# STATISTICS FOR THE SOCIAL SCIENCES

*Bill Pelz*
Herkimer College (SUNY)

LibreTexts™

# Book: Statistics for the Social Sciences (Lumen)

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 04/19/2026

# TABLE OF CONTENTS


Licensing

## 1: Types of Statistical Studies and Producing Data

- 1.1: Why It Matters- Types of Statistical Studies and Producing Data
- 1.2: Introduction- Types of Statistical Studies
- 1.3: Types of Statistical Studies (1 of 4)
- 1.4: Types of Statistical Studies (2 of 4)
- 1.5: Types of Statistical Studies (4 of 4)
- 1.6: Types of Statistical Studies (3 of 4)
- 1.7: Introduction- Sampling
- 1.8: Sampling (1 of 2)
- 1.9: Sampling (2 of 2)
- 1.10: Introduction- Conducting Experiments
- 1.11: Conducting Experiments (1 of 2)
- 1.12: Conducting Experiments (2 of 2)
- 1.13: Putting It Together- Types of Statistical Studies and Producing Data

## 2: Summarizing Data Graphically and Numerically

- 2.1: Why It Matters- Summarizing Data Graphically and Numerically
- 2.2: Introduction- Categorical vs. Quantitative Data
- 2.3: Categorical vs. Quantitative Data
- 2.4: Introduction- Dotplots
- 2.5: Dotplots (1 of 2)
- 2.6: Dotplots (2 of 2)
- 2.7: Introduction- Histograms
- 2.8: Histograms (1 of 4)
- 2.9: Histograms (2 of 4)
- 2.10: Histograms (3 of 4)
- 2.11: Histograms (4 of 4)
- 2.12: Introduction- Measures of Center
- 2.13: Mean and Median (1 of 2)
- 2.14: Mean and Median (2 of 2)
- 2.15: Introduction- Measures of Spread
- 2.16: Interquartile Range and Boxplots (1 of 3)
- 2.17: Interquartile Range and Boxplots (2 of 3)
- 2.18: Interquartile Range and Boxplots (3 of 3)
- 2.19: Introduction- Describing a Distribution
- 2.20: Standard Deviation (1 of 4)
- 2.21: Standard Deviation (2 of 4)
- 2.22: Standard Deviation (3 of 4)
- 2.23: Standard Deviation (4 of 4)
- 2.24: Putting It Together- Summarizing Data Graphically and Numerically
- 2.25: StatTutor- Drinking Habits of College Students
- 2.26: Assignment- Histogram
- 2.27: Assignment- Five-Number Summary

- 2.28: Assignment- Boxplot
- 2.29: Assignment- Standard Deviation


## 3: Examining Relationships- Quantitative Data


- 3.1: Why It Matters- Examining Relationships- Quantitative Data
- 3.2: Introduction- Scatterplots
- 3.3: Scatterplots (1 of 5)
- 3.4: Scatterplots (2 of 5)
- 3.5: Scatterplots (3 of 5)
- 3.6: Scatterplots (4 of 5)
- 3.7: Scatterplots (5 of 5)
- 3.8: Introduction- Linear Relationships
- 3.9: Linear Relationships (1 of 4)
- 3.10: Linear Relationships (2 of 4)
- 3.11: Linear Relationships (3 of 4)
- 3.12: Linear Relationships (4 of 4)
- 3.13: Introduction- Association vs Causation
- 3.14: Causation and Lurking Variables (1 of 2)
- 3.15: Causation and Lurking Variables (2 of 2)
- 3.16: Introduction- Linear Regression
- 3.17: Linear Regression (1 of 4)
- 3.18: Linear Regression (2 of 4)
- 3.19: Linear Regression (3 of 4)
- 3.20: Linear Regression (4 of 4)
- 3.21: Introduction- Assessing the Fit of a Line
- 3.22: Assessing the Fit of a Line (1 of 4)
- 3.23: Assessing the Fit of a Line (2 of 4)
- 3.24: Assessing the Fit of a Line (3 of 4)
- 3.25: Assessing the Fit of a Line (4 of 4)
- 3.26: Putting It Together- Examining Relationships- Quantitative Data
- 3.27: StatTutor- Academic Performance
- 3.28: Assignment- Scatterplot
- 3.29: Assignment- Linear Relationships
- 3.30: Assignment- Linear Regression


## 4: Nonlinear Models


- 4.1: Why It Matters- Nonlinear Models
- 4.2: Introduction- Exponential Relationships
- 4.3: Exponential Relationships (1 of 6)
- 4.4: Exponential Relationships (2 of 6)
- 4.5: Exponential Relationships (3 of 6)
- 4.6: Exponential Relationships (4 of 6)
- 4.7: Exponential Relationships (5 of 6)
- 4.8: Exponential Relationships (6 of 6)
- 4.9: Putting It Together- Nonlinear Models

## 5: Relationships in Categorical Data with Intro to Probability


- 5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability
- 5.2: Introduction- Two-Way Tables
- 5.3: Two-Way Tables (1 of 5)
- 5.4: Two-Way Tables (2 of 5)
- 5.5: Two-Way Tables (3 of 5)
- 5.6: Two-Way Tables (4 of 5)
- 5.7: Two-Way Tables (5 of 5)
- 5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability
- 5.9: StatTutor- Treating Depression- A Randomized Clinical Trial


## 6: Probability and Probability Distributions


- 6.1: Why It Matters- Probability and Probability Distributions
- 6.2: Introduction- Another Look at Probability
- 6.3: Another Look at Probability (1 of 2)
- 6.4: Another Look at Probability (2 of 2)
- 6.5: Introduction- Probability Rules
- 6.6: Probability Rules (1 of 3)
- 6.7: Probability Rules (2 of 3)
- 6.8: Probability Rules (3 of 3)
- 6.9: Introduction- Discrete Probability Distribution
- 6.10: Discrete Random Variables (1 of 5)
- 6.11: Discrete Random Variables (2 of 5)
- 6.12: Discrete Random Variables (3 of 5)
- 6.13: Discrete Random Variables (4 of 5)
- 6.14: Discrete Random Variables (5 of 5)
- 6.15: Introduction- Continuous Probability Distribution
- 6.16: Continuous Probability Distribution (1 of 2)
- 6.17: Continuous Probability Distribution (2 of 2)
- 6.18: Introduction- Normal Random Variables
- 6.19: Normal Random Variables (1 of 6)
- 6.20: Normal Random Variables (2 of 6)
- 6.21: Normal Random Variables (3 of 6)
- 6.22: Normal Random Variables (4 of 6)
- 6.23: Normal Random Variables (5 of 6)
- 6.24: Normal Random Variables (6 of 6)
- 6.25: Putting It Together- Probability and Probability Distribution


## 7: Linking Probability to Statistical Inference


- 7.1: Why It Matters- Linking Probability to Statistical Inference
- 7.2: Parameters vs. Statistics
- 7.3: Introduction- Statistical Inference
- 7.4: Distribution of Sample Proportions (1 of 6)
- 7.5: Distribution of Sample Proportions (2 of 6)
- 7.6: Distribution of Sample Proportions (3 of 6)
- 7.7: Distribution of Sample Proportions (4 of 6)
- 7.8: Distribution of Sample Proportions (5 of 6)
- 7.9: Distribution of Sample Proportions (6 of 6)

- 7.10: Introduction- Distribution of Sample Proportions
- 7.11: Statistical Inference (1 of 3)
- 7.12: Statistical Inference (2 of 3)
- 7.13: Statistical Inference (3 of 3)
- 7.14: Putting It Together- Linking Probability to Statistical Inference


## 8: Inference for One Proportion


- 8.1: Why It Matters- Inference for One Proportion
- 8.2: Introduction- Estimating a Population Proportion
- 8.3: Estimating a Population Proportion (1 of 3)
- 8.4: Estimating a Population Proportion (2 of 3)
- 8.5: Estimating a Population Proportion (3 of 3)
- 8.6: Introduction- Hypothesis Testing
- 8.7: Hypothesis Testing (1 of 5)
- 8.8: Hypothesis Testing (2 of 5)
- 8.9: Hypothesis Testing (4 of 5)
- 8.10: Hypothesis Testing (3 of 5)
- 8.11: Hypothesis Testing (5 of 5)
- 8.12: Introduction- Hypothesis Test for a Population Proportion
- 8.13: Hypothesis Test for a Population Proportion (1 of 3)
- 8.14: Hypothesis Test for a Population Proportion (2 of 3)
- 8.15: Hypothesis Test for a Population Proportion (3 of 3)
- 8.16: Putting It Together- Inference for One Proportion
- 8.17: StatTutor- Cell Phones
- 8.18: Assignment- Hypothesis Testing for the Population Proportion p


## 9: Inference for Two Proportions


- 9.1: Why It Matters- Inference for Two Proportions
- 9.2: Introduction- Distribution of Differences in Sample Proportions
- 9.3: Distribution of Differences in Sample Proportions (1 of 5)
- 9.4: Distribution of Differences in Sample Proportions (2 of 5)
- 9.5: Distribution of Differences in Sample Proportions (3 of 5)
- 9.6: Distribution of Differences in Sample Proportions (4 of 5)
- 9.7: Distribution of Differences in Sample Proportions (5 of 5)
- 9.8: Introduction- Estimate the Difference Between Population Proportions
- 9.9: Estimate the Difference between Population Proportions (1 of 3)
- 9.10: Estimate the Difference between Population Proportions (2 of 3)
- 9.11: Estimate the Difference between Population Proportions (3 of 3)
- 9.12: Introduction- Hypothesis Test for Difference in Two Population Proportions
- 9.13: Hypothesis Test for Difference in Two Population Proportions (1 of 6)
- 9.14: Hypothesis Test for Difference in Two Population Proportions (2 of 6)
- 9.15: Hypothesis Test for Difference in Two Population Proportions (3 of 6)
- 9.16: Hypothesis Test for Difference in Two Population Proportions (4 of 6)
- 9.17: Hypothesis Test for Difference in Two Population Proportions (5 of 6)
- 9.18: Hypothesis Test for Difference in Two Population Proportions (6 of 6)
- 9.19: Putting It Together- Inference for Two Proportions
- 9.20: Assignment- A Statistical Investigation using Software

## 10: Inference for Means


- 10.1: Why It Matters- Inference for Means
- 10.2: Introduction- Distribution of Sample Means
- 10.3: Distribution of Sample Means (1 of 4)
- 10.4: Distribution of Sample Means (2 of 4)
- 10.5: Distribution of Sample Means (3 of 4)
- 10.6: Distribution of Sample Means (4 of 4)
- 10.7: Introduction- Estimating a Population Mean
- 10.8: Estimating a Population Mean (1 of 3)
- 10.9: Estimating a Population Mean (2 of 3)
- 10.10: Estimating a Population Mean (3 of 3)
- 10.11: Introduction- Hypothesis Test for a Population Mean
- 10.12: Hypothesis Test for a Population Mean (1 of 5)
- 10.13: Hypothesis Test for a Population Mean (2 of 5)
- 10.14: Hypothesis Test for a Population Mean (3 of 5)
- 10.15: Hypothesis Test for a Population Mean (4 of 5)
- 10.16: Hypothesis Test for a Population Mean (5 of 5)
- 10.17: Introduction- Inference for a Difference in Two Population Means
- 10.18: Inference for a Difference in Two Population Means
- 10.19: Hypothesis Test for a Difference in Two Population Means (1 of 2)
- 10.20: Hypothesis Test for a Difference in Two Population Means (2 of 2)
- 10.21: Estimating the Difference in Two Population Means
- 10.22: Putting It Together- Inference for Means
- 10.23: StatTutor- Analyzing Data From a Course's Grade Book
- 10.24: Assignment- Distribution of Sample Means
- 10.25: Assignment- Connection between Confidence Intervals and Sampling Distributions
- 10.26: Assignment- Hypothesis Testing for the Population Mean
- 10.27: Assignment- Matched Pairs
- 10.28: Assignment- Checking Conditions
- 10.29: Assignment- Two Independent Samples


## 11: 11- Chi-Square Tests


- 11.1: Why It Matters- Chi-Square Tests
- 11.2: Introduction- Chi-Square Test for One-Way Tables
- 11.3: Goodness-of-Fit (1 of 2)
- 11.4: Goodness-of-Fit (2 of 2)
- 11.5: Introduction- Chi-Square Tests for Two-Way Tables
- 11.6: Test of Independence (1 of 3)
- 11.7: Test of Independence (2 of 3)
- 11.8: Test of Independence (3 of 3)
- 11.9: Test of Homogeneity
- 11.10: Putting It Together- Chi-Square Tests
- 11.11: StatTutor- Risk Factors for Low Birth Weight
- 11.12: Assignment- Test of Independence Using Technology
- 11.13: Assignment- Using Technology with Data to Run a Hypothesis Test

## 12: Main Body

- 12.1: Chapter 1

## 13: Course Information

- 13.1: Welcome to the class!
- 13.2: Course Learning Activities
- 13.3: How Your Course Grade is Determined
- 13.4: Heutagogical Course Objectives
- 13.5: What is Heutagogy?
- 13.6: My Expectations…

## 14: Module 1

- 14.1: Module 1 Learning Activities
- 14.2: Module 1 Discussion Forum
- 14.3: Reflective Blog 1
- 14.4: Blog 1 Forum
- 14.5: Module 1 Homework
- 14.6: Module 1 Exam
- 14.7: Talk with the Professor

## 15: Modules 2, 3, 4 and 5

- 15.1: Structure of Modules 2 through 5

## Index

## Glossary

## Detailed Licensing

# Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

# CHAPTER OVERVIEW

## 1: Types of Statistical Studies and Producing Data


1.1: Why It Matters- Types of Statistical Studies and Producing Data
1.2: Introduction- Types of Statistical Studies
1.3: Types of Statistical Studies (1 of 4)
1.4: Types of Statistical Studies (2 of 4)
1.5: Types of Statistical Studies (4 of 4)
1.6: Types of Statistical Studies (3 of 4)
1.7: Introduction- Sampling
1.8: Sampling (1 of 2)
1.9: Sampling (2 of 2)
1.10: Introduction- Conducting Experiments
1.11: Conducting Experiments (1 of 2)
1.12: Conducting Experiments (2 of 2)
1.13: Putting It Together- Types of Statistical Studies and Producing Data


---


This page titled 1: Types of Statistical Studies and Producing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.1: Why It Matters- Types of Statistical Studies and Producing Data

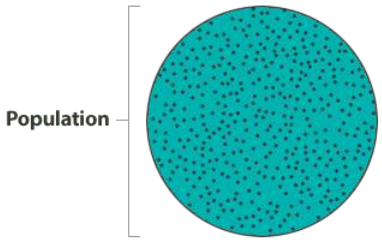
We organized this course around the Big Picture of Statistics. As we learn new material, we will always look at how these new ideas relate to the Big Picture. In this way the Big Picture is a diagram that will help us organize and understand the material we will learn throughout the course.

The Big Picture summarizes the steps in a statistical investigation.

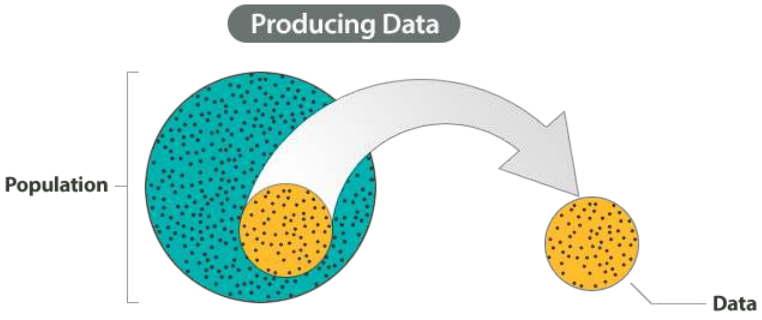
We begin a statistical investigation with a research question. The research question is frequently something we want to know about a **population**. The population can be people or other things, such as animals or objects. For example, we might want to know the answer to questions such as:

- What percentage of U.S. adults supports the death penalty? (Population: U.S. adults)
- Do cell phones affect bees? (Population: bees)
- Do cars get better gas mileage with a new gasoline additive? (Population: cars)

The population is the entire group that we want to know something about:



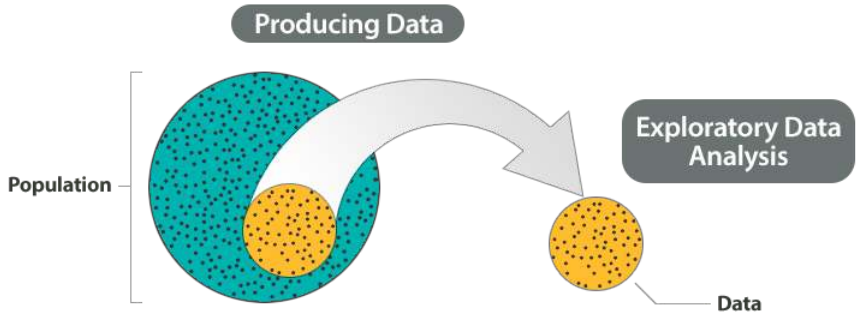

In most cases, the population is a large group. Often, the population is so large that we cannot collect information from every individual in the population. So we select a **sample** from the population. Then we collect data from this sample. This is the first step in the statistical investigation. We call this step **producing data**.



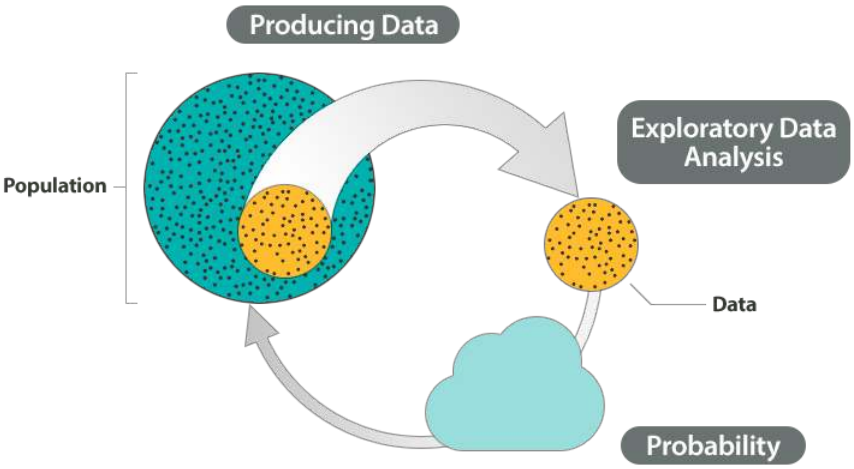

Of course, we need a sample that represents the population well. This involves careful planning but also involves chance. For example, if our goal is to determine the percentage of U.S. adults who favor the death penalty, we do not want our sample to contain only Democrats or only Republicans. So we can give everyone the same opportunity to be in the sample, but we will let chance select the sample.

At this step of the investigation we also carefully define what kind of information we plan to gather. Then we collect the data.

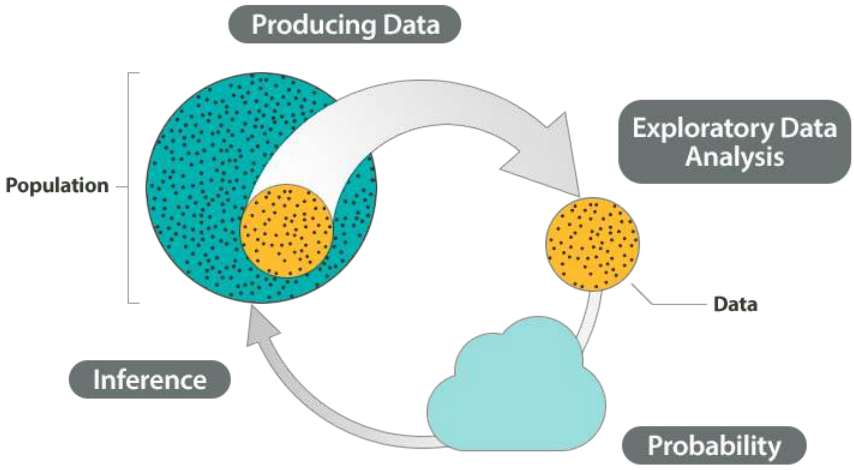
Data is often a long list of information. To make sense of the data, we explore it and summarize it using graphs and different numerical measures, such as percentages or averages. We call this step **exploratory data analysis**.




Remember, our goal is to answer a question about a population based on a sample. Of course, samples will vary due to chance, and we will need to answer our question in spite of this variability. So we need to understand how sample results will vary and how sample results relate to the population as a whole when chance is involved. This is where **probability** comes in.




Probability is the "machinery" behind the last step in the process called **inference**. We infer something about a population based on a sample. This inference is the conclusion we reach from our sample data that answers our original question about the population.



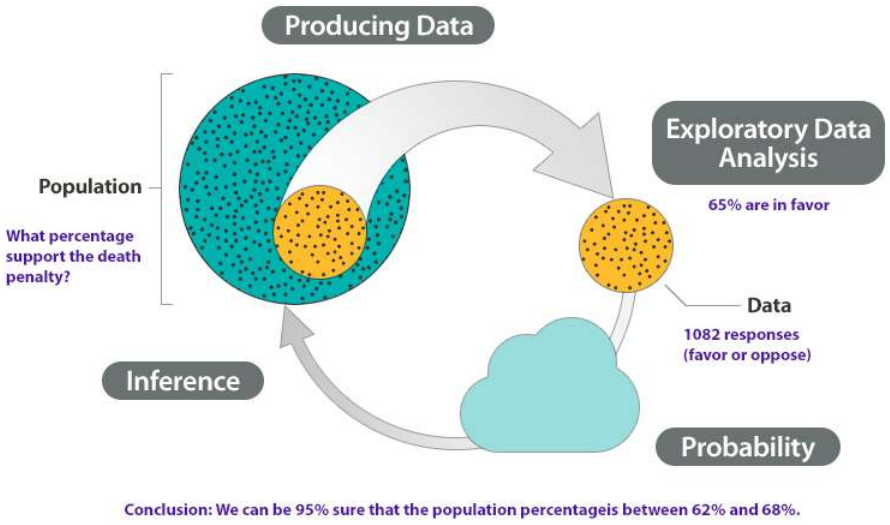

### Example – The big picture of statistics

At the end of April 2005, ABC News and the Washington Post conducted a poll to determine the percentage of U.S. adults who support the death penalty.

**Research question**: What percentage of U.S. adults support the death penalty?

Steps in the statistical investigation:

1. **Produce Data**: *Determine what to measure, then collect the data.*
   The poll selected 1,082 U.S. adults at random. Each adult answered this question: "Do you favor or oppose the death penalty for a person convicted of murder?"
2. **Explore the Data**: *Analyze and summarize the data.*
   In the sample, 65% favored the death penalty.
3. **Draw a Conclusion**: *Use the data, probability, and statistical inference to draw a conclusion about the population.*
   Our goal is to determine the percentage of the U.S. adult population that supports the death penalty. We know that different samples will give different results. What are the chances that our sample reflects the opinions of the population within 3%? Probability describes the likelihood that our sample is this accurate. So we can say with 95% confidence that between 62% and 68% of the population favor the death penalty.



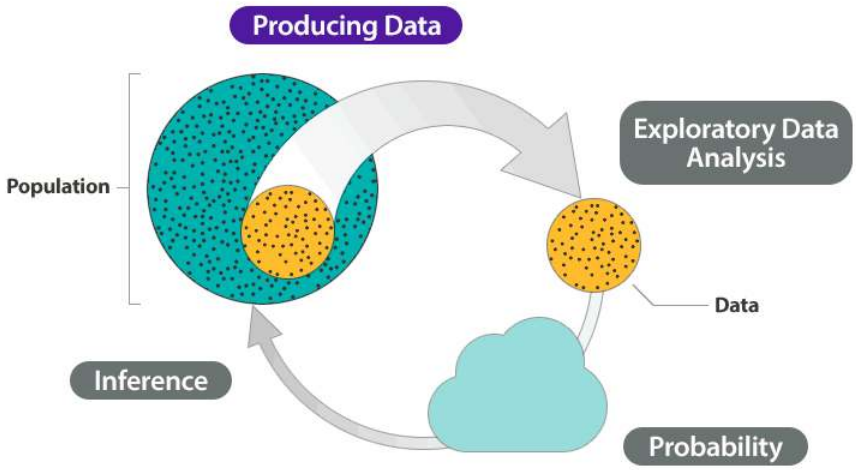

## Let's Summarize

A statistical investigation with a research question. Then the investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data.
- Explore the Data: Analyze and summarize the data (also called *exploratory data analysis*).
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

## Types of Statistical Studies and Producing Data

In this first module, we focus on the *produce data* step in a statistical investigation. We discuss two types of statistical investigations: the observational study and the experiment. Each type of investigation involves a different approach to collecting data. We will also see that our approach to collecting data determines what we can conclude from the data.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 1.1: Why It Matters- Types of Statistical Studies and Producing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.2: Introduction- Types of Statistical Studies

## What you'll learn to do: Describe various types of statistical studies and the types of conclusions that are appropriate.

### LEARNING OBJECTIVES

- From a research question, determine the goal of a statistical study.
- Determine if a study is an experiment or an observational study.
- From a description of a statistical study, determine the goal of the study.
- Based on the study design, determine what types of conclusions are appropriate.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.2: Introduction- Types of Statistical Studies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.3: Types of Statistical Studies (1 of 4)

## Learning Objectives

- From a research question, determine the goal of a statistical study.

Before we begin our discussion of the types of statistical studies, we look closely at the types of research questions used in statistical studies.

## Research Questions about a Population

Recall that a *population* is the entire group of individuals or objects that we want to study. Usually, it is not possible to study the whole population, so we collect data from a part of the population, called a *sample*. We use the sample to draw conclusions about the population.

For example, suppose our research question is "What is the average amount of money spent on textbooks per semester by full-time students at Seattle Central?" We cannot interview every full-time student at Seattle Central because would take too much time and cost too much money. We therefore carefully select a sample of full-time students at Seattle Central to represent the population of all full-time students at that college. Then we collect data from the sample to estimate the average amount spent on textbooks.

This example illustrates how the research question guides the investigation. A well-stated research question contains information about:

- The population (full-time students at Seattle Central).
- The information we will collect from each individual in the sample. We also call this the **variable**. The variable is what we plan to measure (amount of money spent on textbooks per semester).
- A numerical characteristic about the population related to this variable (the *average* amount of money spent on textbooks per semester).

Here are some common types of research questions about a population:

| Type of Research Question | Examples |
|---|---|
| **Make an estimate about the population** (often an estimate about an *average* value or a *proportion* with a given characteristic) | What is the *average* number of hours that community college students work each week? |
| What *proportion* of all U.S. college students are enrolled at a community college? | |
| **Test a claim about the population** (often a claim about an *average* value or a *proportion* with a given characteristic) | Is the *average* course load for a community college student greater than 12 units? |
| Do the *majority* of community college students qualify for federal student loans? | |
| **Compare two populations** (often a comparison of population averages or proportions with a given characteristic) | In community colleges, do female students have a *higher* GPA than male students? |
| Are college athletes *more* likely than nonathletes to receive academic advising? | |
| **Investigate a relationship** between two variables in the population | Is there a *relationship* between the number of hours high school students spend each week on Facebook and their GPA? |
| Is academic counseling *associated* with quicker completion of a college degree? | |

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3820

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3821

## Research Questions about Cause and Effect

A research question that focuses on a **cause-and-effect** relationship is common in disciplines that use experiments, such as medicine or psychology. These types of questions ask how one variable responds as another variable is manipulated. These types of questions involve two variables. Here are some examples:

- Does cell phone usage increase the risk of developing a brain tumor?
- Does drinking red wine lower the risk of a heart attack?
- Does playing violent video games increase aggressive behavior?
- Does sex education lower the incidence of teen pregnancy?

To provide convincing evidence of a cause-and-effect relationship, the researcher designs an experiment. We discuss experiments in "Collecting Data – Conducting an Experiment."

Note: In the previous section, *Research Questions about a Population*, we included examples of questions about the relationship between two variables in a population. But in these types of questions, we used words like *associated*, *correlated*, *linked to*, and *connected*. These words do not imply a cause-and-effect relationship between the variables. We can investigate these types of questions without conducting an experiment – an observational study will do. We study observational studies in "Collecting Data – Sampling."

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3822

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.3: Types of Statistical Studies (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.4: Types of Statistical Studies (2 of 4)

## Learning Objectives

- Determine if a study is an experiment or an observational study.
- From a description of a statistical study, determine the goal of the study.

In general, there are two types of statistical studies: observational studies and experiments.

An **observational study** observes individuals and measures variables of interest. The main purpose of an observational study is to describe a group of individuals or to investigate an association between two variables. We can answer questions about a population with an observational study. We can also investigate a relationship between two variables. But in an observational study, researchers do not attempt to manipulate one variable to cause an effect in another variable. For this reason, an observational study does not provide convincing evidence of a cause-and-effect relationship.

An **experiment** intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables. But the experiment has to be well-designed to provide convincing evidence of a cause-and-effect relationship. We study experiment design in the section "Collecting Data – Conducting an Experiment."

For now, our goal is to distinguish between these two types of studies. We focus on the connection between the research question, the type of study, and the kinds of conclusions we can make.

## Example

### Music and Learning

Many students listen to music while studying. Does listening to music improve learning? Students in a statistics class decide to investigate this question. They write more specific research questions related to the topic of music and learning. Then they design the following three studies:

**Study 1**

Specific research questions: *Do the majority of college students listen to music while they study? Do the majority of college students believe that listening to music improves their learning?*

To investigate these questions, the statistics students conduct a survey in their other classes. They ask these two questions:

- Do you listen to music while you study?
- Do you think listening to music improves your concentration and memory?

This is an observational study designed to answer two questions about a population of college students. Each question contains a claim about the population of college students. We can use data from this study to see if these claims are true. But data from this study cannot provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning.

**Study 2**

Specific research question: *When we compare students who study with music to students who study in a quiet environment, which group gives higher ratings for understanding what they studied?*

To investigate this question, the instructor divides the class into two groups: (1) those who listen to music when they study and (2) those who do not listen to music when they study. The students keep a journal for a week. Each time they study, they record the following information:

- Length of study session (in minutes)
- A rating of how well they understood what they studied, on a scale of 1–10: 1 = no understanding, 10 = excellent understanding.

This investigation is also an observational study. It compares two populations: (1) college students who listen to music when studying and (2) college students who do not listen to music when studying. We can also view this as an observational study of one

population (college students) that investigates the relationship between two variables: *listening to music while studying* and *perceived understanding of material studied*. From this study, we might learn something interesting about the connection between college students' study habits and their perception of their learning. But since this is an observational study, data from this study cannot provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning.

**Study 3**

Specific research question: *Does listening to music improve students' ability to quickly identify information?*

To investigate this question, the instructor uses word-search puzzles. She divides the class into two groups. Students on one side of the room do a word puzzle for 3 minutes while listening to music on an iPod. Students on the other side of the room do a word puzzle for 3 minutes without music. The instructor calculates the average number of words found by each group.

This study is an experiment. The instructor manipulates music to investigate the impact on puzzle completion. Data from this study can provide evidence of a cause-and-effect relationship between listening to music while studying and improvements in learning. But the improvement in learning is more narrowly defined as the ability to quickly identify information, such as words in a puzzle.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3401

https://assessments.lumenlearning.co...sessments/3402

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.4: Types of Statistical Studies (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

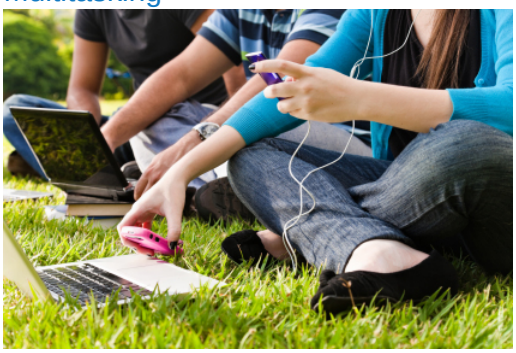
# 1.5: Types of Statistical Studies (4 of 4)

## Learning Objectives

- Based on the study design, determine what types of conclusions are appropriate.

## Example

## Multitasking



Do you constantly text-message while in class? Do you jump from one website to another while doing homework? If so, then you are a high-tech multitasker. In a study of high-tech multitasking at Stanford University, researchers put 100 students into two groups: those who regularly do a lot of media multitasking and those who don't. The two groups performed a series of three tasks:

(1) A task to measure the ability to pay attention:

> Students view two images of red and blue rectangles flashed one after the other on a computer screen. They try to tell if the red rectangles are in a different position in the second frame.

(2) A task to measure control of memory:

> Students view a sequence of letters flashed onto a computer screen, then recall which letters occurred more than once.

(3) A task to measure the ability to switch from one job to another:

> Students view numbers and letters together with the instructions to pay attention to the numbers, then recall if the numbers were even or odd. Then the instructions switch. Students are to pay attention to the letters and recall if the letters were vowels or consonants.

On every task, the multitaskers did worse than the non-multitaskers.

The researchers concluded that "people who are regularly bombarded with several streams of electronic information do not pay attention, control their memory, or switch from one job to another as well as those who prefer to complete one task at a time" (as reported in *Stanford News* in 2009).

"When they're [high-tech multitaskers] in situations where there are multiple sources of information coming from the external world or emerging out of memory, they're not able to filter out what's not relevant to their current goal," said Wagner, an associate professor of psychology at Stanford. "That failure to filter means they're slowed down by that irrelevant information."

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3403

In general, we should not make cause-and-effect statements from observational studies, but in reality, researchers do it all the time. This does not mean that researchers are drawing incorrect conclusions from observational studies. Instead, they have developed techniques that go a long way toward decreasing the impact of confounding variables. These techniques are beyond the scope of this course, but we briefly discuss a simplified example to illustrate the idea.

## Smoking and Cancer

Consider this excerpt from the National Cancer Institute website:

> *Smoking is a leading cause of cancer and of death from cancer. Millions of Americans have health problems caused by smoking. Cigarette smoking and exposure to tobacco smoke cause an estimated average of 438,000 premature deaths each year in the United States.*

Notice that the National Cancer Institute clearly states a cause-and-effect relationship between smoking and cancer. Now let's think about the evidence that is required to establish this causal link. Researchers would need to conduct experiments similar to the hormone replacement therapy experiments done by the Women's Health Initiative. Such experiments would be very difficult to do. The researchers cannot manipulate the smoking variable. Doing so would require them to randomly assign people to smoke or to abstain from smoking their whole life. Obviously, this is impossible. So how can we say that smoking *causes* cancer?

In practice, researchers approach this challenge in a variety of ways. They may use advanced techniques for making *statistical adjustments* within an observational study to control the effects of confounding variables that could influence the results. A simple example is the cell phone and brain cancer study.

> *In this observational study, researchers identified a group of 469 people with brain cancer. They paired each person who had brain cancer with a person of the same sex, of similar age, and of the same race who did not have brain cancer. Then they compared the cell phone use for each pair of people. This matching attempts to control the confounding effects of sex, age, and race on the response variable, cancer. With these adjustments, the study will provide stronger evidence for (or against) a casual link.*

However, even with such adjustments, we should be cautious about using evidence from an observational study to establish a cause-and-effect relationship. Researchers used these types of adjustments in the observational studies with hormone replacement therapy. We saw in that research that the results were still misleading when compared to those of an experiment.

So how can the National Cancer Institute state as a fact that smoking causes cancer?

They used other nonstatistical guidelines to build evidence for a cause-and-effect relationship from observational studies. In this approach, researchers review a large number of observational studies with criteria that, if met, provide stronger evidence of a possible cause-and-effect relationship. Here are some simplified examples of the criteria they use:

(1) There is a reasonable explanation for how one variable might cause the other.

- For example, experiments with rats show that chemicals found in cigarettes cause cancer in rats. It is therefor reasonable to infer that these same chemicals may cause cancer in humans.
- Consider these experiments together with the observational studies showing the association between smoking and cancer in humans. We now have more convincing evidence of a possible cause-and-effect relationship between smoking and cancer in humans.

(2) The observational studies vary in design so that factors that confound one study are not present in another.

- For example, one observational study shows an association between smoking and lung cancer, but the people in the study all live in a large city. Air pollution in a large city may contribute to the lung cancer, so we cannot be sure that smoking is the cause of cancer in this study.
- Another observational study looks only at nonsmokers. This study shows no difference in lung cancer rates for nonsmokers living in rural areas compared to nonsmokers living in cities.
- Consider these two studies together. The second study suggests that air pollution does not contribute to lung cancer, so we now have more convincing evidence that smoking (not air pollution) is the cause of higher cancer rates in the first study.

### Let's Summarize

- There are four steps in a statistical investigation:
  - Ask a question that can be answered by collecting data.
  - Decide what to measure, and then collect data.
  - Summarize and analyze.
  - Draw a conclusion, and communicate the results.
- There are two types of statistical research questions:
  - Questions about a population
  - Questions about cause-and-effect
- To answer a question about a population, we select a sample and conduct an observational study. To answer a question about cause-and-effect we conduct an experiment.
- There are two types of statistical studies:
  - Observational studies: An *observational study* observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an association between two variables. An observational study alone does not provide convincing evidence of a cause-and-effect relationship.
  - Experiments: An *experiment* intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- In statistics, a *variable* is information we gather about individuals or objects.
- When we investigate a relationship between two variables, we identify an *explanatory* variable and a *response* variable. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only thing that impacts the response variable. Other factors, however, may also influence the response. These other factors are called *confounding* variables.
- The influence of confounding variables on the response variable is one of the reasons that an observational study gives weak, and potentially misleading, evidence of a cause-and-effect relationship. A well-designed experiment takes steps to eliminate the effects of confounding variables, such as random assignment of people to treatment groups, use of a placebo, and blind conditions. For this reason, a well-designed experiment provides convincing evidence of cause-and-effect.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.5: Types of Statistical Studies (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.6: Types of Statistical Studies (3 of 4)

## Learning Objectives

- Based on the study design, determine what types of conclusions are appropriate.

We now focus more closely on studies that investigate a relationship between two variables. In these types of studies, one variable is the **explanatory variable**, and the other is the **response variable**. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only thing that impacts the response variable. We therefore get rid of all other factors that might affect the response. Then we manipulate the explanatory variable. Our goal is to see if it really does affect the response.

In an observational study, researchers may take steps to reduce the influence of these other factors on the response. But it is difficult in an observational study to get rid of all the factors that may have an influence. In addition, the researchers do not manipulate the explanatory variable to see if it affects the response. They just collect data and look for an association between the two variables. For these reasons, observational studies do not give convincing evidence of a cause-and-effect relationship.

In an experiment, researchers use a variety of techniques to eliminate the influence of these other factors. Then they manipulate the explanatory variable to see if it affects the response. For this reason, experiments give the strongest evidence for a cause-and-effect relationship.

## Example

### Hormone Replacement Therapy

When women go through menopause, the production of hormones in their bodies changes. The hormonal changes can cause a variety of symptoms that may be reduced by hormone replacement therapy. In the 1980s, hormone replacement therapy was common in the United States.

In the early 1990s, observational studies suggested that hormone replacement therapy had additional benefits, including a reduction in the risk of heart disease. In these observational studies, researchers compared women who took hormones to those who did not take hormones. Health records showed that women taking hormones after menopause had a lower incidence of heart disease. But women who take hormones are different from other women. They tend to be richer and more educated, to have better nutrition, and to visit the doctor more frequently. These women have many habits and advantages that contribute to good health, so it is not surprising that they have fewer heart attacks. But can we conclude from these studies that the hormones caused the reduction in heart attacks? No. The results are *confounded* by the influence of these other factors.

In 2002, the Women's Health Initiative sponsored a large-scale, well-designed experiment to study the health implications of hormone replacement therapy. In this experiment, researchers randomly assigned over 16,000 women to one of two treatments. One group took hormones. The other group took a **placebo**. A placebo is a pill with no active ingredients that looks like the hormone pill. The experiment was **double-blind**. *Blind* means that women did not know if they were receiving hormones or the placebo. *Double-blind* means that the information was coded, so researchers administering the pills did not know which treatment the women received. After 5 years, the group taking hormones had a *higher* incidence of heart disease and breast cancer. This is exactly the opposite result from the result found in the observational studies! In fact, the differences were so significant that the researchers ended the experiment early. The National Institutes of Health declared that the observational studies were wrong. Hormone replacement therapy to treat menopausal symptoms is now rarely used.

## What's the Main Point?

An observational study can provide evidence of a link or an association between two variables. But other factors may also influence the results. These other factors are called *confounding variables*. The influence of confounding variables on the response variable is one of the reasons that an observational study gives weak, and potentially misleading, evidence of a cause-and-effect relationship. A well-designed experiment takes steps to eliminate the effects of confounding variables, including random assignment of people to treatment groups, use of a placebo, or blind conditions. Using these precautions, a well-designed experiment provides convincing evidence of cause-and-effect.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3404

https://assessments.lumenlearning.co...sessments/3405

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 1.6: Types of Statistical Studies (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.7: Introduction- Sampling

## What you'll learn to do: For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

### LEARNING OBJECTIVES

- For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.7: Introduction- Sampling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.8: Sampling (1 of 2)

## Learning Objectives

- For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

We now focus on observational studies and how to collect reliable and accurate data for an observational study.

We know that an observational study can answer questions about a population. But populations are generally large groups, so we cannot gather data from every individual in the population. Instead, we select a sample and gather data from the sample. We use the data from the sample to make statements about the population.

Here are two examples:

- A political scientist wants to know what percentage of college students consider themselves conservatives. The population is college students. It would be too time consuming and expensive to poll every college student, so the political scientist selects a sample of college students. Of course, the sample must be carefully selected to represent the political perspectives that are present in the population.
- A government agency plans to test airbags from Honda to determine if the airbags work properly. Testing an airbag means it has to be inflated and punctured, which ruins the airbag, so the researchers certainly cannot test every airbag. Instead, they test a sample of airbags and draw a conclusion about the quality of airbags from Honda.

## Important Point

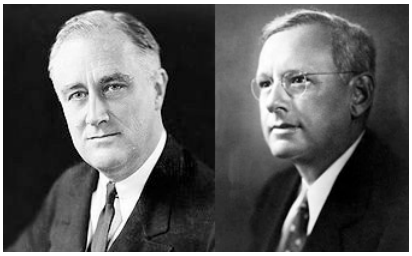
Our goal is to use a sample to make valid conclusions about a population. Therefore, the *sample must be representative of the population*. A representative sample is a subset of the population that reflects the characteristics of the population.

A **sampling plan** describes exactly how we will choose the sample. A sampling plan is **biased** if it systematically favors certain outcomes.

In our discussion of sampling plans, we focus on surveys. The next example is a famous one that illustrates how biased sampling in a survey leads to misleading conclusions about the population.

## Example

### The 1936 Presidential Election



In 1936, Democrat Franklin Roosevelt and Republican Alf Landon were running for president. Before the election, the magazine *Literary Digest* sent a survey to 10 million Americans to determine how they would vote. More than 2 million people responded to the poll; 60% supported Landon. The magazine published the findings and predicted that Landon would win the election. However, Roosevelt defeated Landon in one of the largest landslide presidential elections ever.

What happened?

The magazine used a biased sampling plan. They selected the sample using magazine subscriptions, lists of registered car owners, and telephone directories. The sample was not representative of the American public. In the 1930s, Democrats were much less likely to own a car or have a telephone. The sample therefore systematically *underrepresented* Democrats. The poll results did not represent the way people in the general population voted.

Before we discuss a method for avoiding bias, let's look at some examples of common survey plans that produce unreliable and potentially biased results.

Example

## How to Sample Badly

**Online polls:** The American Family Association (AFA) is a conservative Christian group that opposes same-sex marriage. In 2004, the AFA began a campaign in support of a constitutional amendment to define marriage as strictly between a man and a woman. The group posted a poll on its website asking AFA members to voice their opinion about same-sex marriage. The AFA planned to forward the results to Congress as evidence of America's opposition to same-sex marriage. Almost 850,000 people responded to the poll. In the poll, 60% *favored* legalizing same-sex marriage.

What happened? Against the wishes of the AFA, the link to the poll appeared in blogs, social-networking sites, and a variety of email lists connected to gay/lesbian/bisexual groups. The AFA claimed that gay rights groups had *skewed* its poll. Of course, the results of the poll would have been skewed in the other direction had only AFA members been allowed to participate.

This is an example of a **voluntary response sample**. The people in a voluntary response sample are self-selected, not chosen. For this reason, a voluntary response sample is biased because only people with strong opinions make the effort to participate.

**Mall surveys:** Have you ever noticed someone surveying people at a mall? People shopping at a mall are more likely to be teenagers, retired people, or people who have more money than the typical American. In addition, unless interviewers are carefully trained, they tend to interview people with whom they are comfortable talking. For these reasons, mall surveys frequently *overrepresent* the opinions of white middle-class or retired people. Mall surveys are an example of a **convenience sample**.

Example

## How to Eliminate Bias in Sampling

In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer chooses who will be part of the sample. In both cases, personal choice produces a biased sample. **Random sampling** is the best way to eliminate bias. Collecting a random sample is like pulling names from a hat (assuming every individual in the population has a name in the hat!). In a **simple random sample** everyone in the population has an equal chance of being chosen.

Reputable polling firms use techniques that are more complicated than pulling names out of a hat. But the goal is the same: eliminate bias by using random chance to decide who is in the sample.

Random samples will eliminate bias, even bias that may be hidden or unknown. The next three activities will reveal a bias that most of us have but don't know that we have! We will see how random sampling avoids this bias.

### Random Samples

**Instructions:** Use the simulation below for this activity. You will see 60 circles. This is the "population." *Our goal is to estimate the average diameter of these 60 circles by choosing a sample.*

1. Choose a sample of five circles that look representative of the population of all 60 circles. Mark your five circles by clicking on each of them. They will turn orange. Record the average diameter for the five circles. (Make sure you have five orange circles before you record the average diameter.)
2. Reset the simulation.
3. Choose another five circles and record the average diameter for this sample of circles. You can reuse a circle, but the sample should not have all the same circles. You now have the averages for two samples.
4. Reset and repeat for a total of 10 samples. Record the average diameter for each sample.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=36

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3823

https://assessments.lumenlearning.co...sessments/3824

Now we estimate the average diameter of the 60 circles using *random* samples.

**Instructions:** Use the simulation below for this activity. You will again see the same 60 circles. As before, this is the "population." *Our goal is to estimate the average diameter of these 25 circles by choosing a random sample.*

1. Click on the "Generate sample" button to get a random sample of five circles by clicking on the random sample button. The simulation randomly chooses five circles. Record the average diameter for the random sample.
2. Reset the simulation using the reset button.
3. Click on the "Generate sample" button to get another random sample. Record the average diameter for this random sample. You now have the averages for two samples.
4. Reset and repeat for a total of 10 samples. Record the average diameter for each sample.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=36

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3825

https://assessments.lumenlearning.co...sessments/3826

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3827

### Comment

Random selection also guarantees that the sample results do not change haphazardly from sample to sample. When we use random selection, the variability we see in sample results is due to chance. The results obey the mathematical laws of probability. We looked at this idea briefly in the Big Picture of Statistics. Probability is the machinery for drawing conclusions about a population on the basis of samples. To use this machinery, the sample must be chosen by random chance.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3406

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3407

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.8: Sampling (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.9: Sampling (2 of 2)

## Learning Objectives

- For an observational study, critique the sampling plan. Recognize implications and limitations of the plan.

Let's briefly summarize the main points about sampling:

- We draw a conclusion about the population on the basis of the sample.
- To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population that reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome.
- Random selection eliminates bias.

We have not mentioned the size of the sample. Are larger samples more accurate? Well, the answer is yes and no.

Recall the 1936 presidential election. A sample of over 2 million people did not correctly identify the winner of the election. Two million people is a huge sample, yet the results were completely wrong. So a large sample does not guarantee reliable results.

However, if the samples are randomly selected, then size does matter. We see this in the next example.

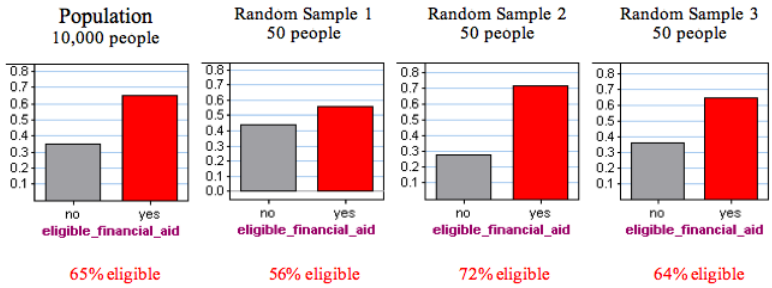
## Example

## For Random Samples, Size Matters

Let's compare the accuracy of random samples of different sizes.

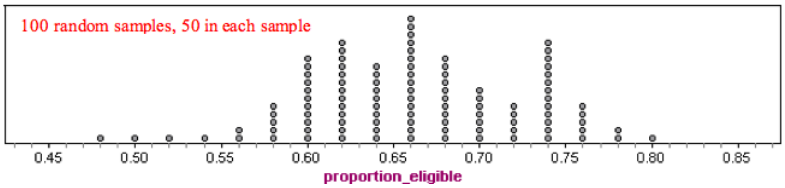
Suppose there are 10,000 students at your college. Also suppose that 65% of these students are eligible for financial aid. How accurate are random samples at predicting this population value?

To answer this question, we randomly select 50 students and determine the proportion who are eligible for financial aid. We repeat this several times. Here are the results for three random samples:




Notice that each random sample has a different result. Some results are larger than the true population value of 65%; some results are smaller than the true population value. Because there is no bias in random samples, we expect results above and below the true value to occur with similar frequency.

Now we use a simulation to take many more random samples. Again, each sample is composed of 50 randomly selected people. Here is a dotplot of the proportion who are eligible for financial aid in 100 samples. Each dot is a random sample.



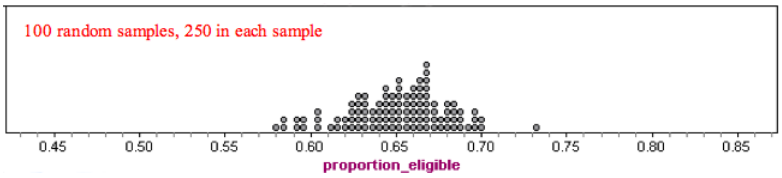

We see that the results from random samples vary from 0.48 to 0.80. Typical values range from about 0.58 to 0.74.

Note: Many samples have results below the true population value of 0.65, and many have results above 0.65. This shows that random samples are not biased. For the question *Are you eligible for financial aid?*, there is no systematic favoring of one response

over another. The samples are representative of the population.

**What happens when we increase the number of people in the random sample?**

We increased the number of people in each sample to 250. Here is dotplot of the results from 100 of these larger random samples.



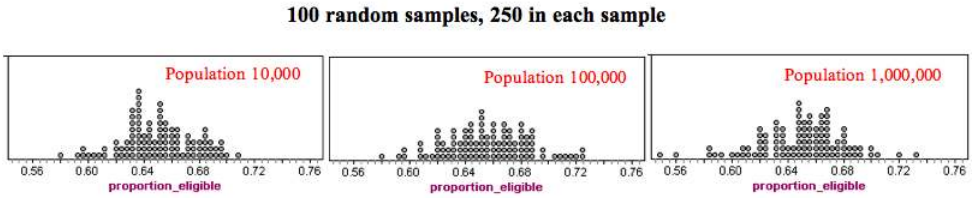

Notice there is less variability in these larger samples. Results range from about 0.58 to 0.73. Typical values range from about 0.62 to 0.68. These samples give results that are closer to the true population value of 0.65.

So what's the point? *Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly.*

## Comment

The precision of the sample results depends on the size of the sample, not the size of the population. The following dotplots illustrate this point. Here we selected samples with 250 people in each sample, but we *varied the size of the population*. Each dotplot contains 100 samples.

Notice that the sample results look very similar. For each population, the sample results fall between about 0.58 and 0.73. In each graph, it is common for sample results to fall between about 0.62 and 0.68.




What's the main point? *The size of the population does not affect the accuracy of a random sample as long as the population is large.*

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3408

https://assessments.lumenlearning.co...sessments/3828

## Comment

If an attempt is made to include every individual from a population in a sample, then the investigation is called a **census**. Every 10 years, the U.S. Census Bureau conducts a population census. It attempts to collect information about every person living in the United States. However, the population census misses between 1% and 3% of the U.S. population and accidentally counts some people more than once. A full census is possible only for small populations.

## Let's Summarize

- We draw a conclusion about the population on the basis of the sample.
- To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population. It also reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome.
- Random selection eliminates bias.
- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly.
- The size of the population does not affect the accuracy of a random sample as long as the population is large.
- If an attempt is made to include every individual from a population in a sample, then the investigation is called a *census*.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 1.9: Sampling (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.10: Introduction- Conducting Experiments

## What you'll learn to do: Identify features of experiment design that control the effects of confounding.

### LEARNING OBJECTIVES

- Identify features of experiment design that control the effects of confounding.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.10: Introduction- Conducting Experiments is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.11: Conducting Experiments (1 of 2)

## Learning Objectives

- Identify features of experiment design that control the effects of confounding.

We now focus on experiments.

The primary goal of an **experiment** is to provide evidence for a cause-and-effect relationship between two variables. An experiment intentionally manipulates the explanatory variable in an attempt to cause an effect on the response variable. To establish a cause-and-effect relationship, we want to make sure that the explanatory variable is the only factor that impacts the response variable. We therefore attempt to get rid of all other factors that might affect the response. These other factors are called **confounding variables**.

To confound means to mix up or to confuse. Confounding variables mix up our ability to determine if the explanatory variable causes a change in the response variable. If we do not control the effects of confounding variables, the experiment does not provide evidence of a cause-and-effect relationship between the explanatory and response variables.

Researchers use two common strategies to control the effects of confounding variables:

- Direct control
- Random assignment

## Example

## Direct Control

Researchers compare bacteria reduction for three different hand-drying methods. In this experiment, participants handled uncooked chicken for 45 seconds, then washed their hands with one squirt of soap for 60 seconds, and then used one of three hand-drying methods. After participants completely dried their hands, researchers measured the bacteria count on their hands. The *Infectious Disease News* published the results in 2010.

In this experiment, the explanatory variable is *hand-drying method*.The response variable is *bacteria count*. Notice that the explanatory variable determines the three treatments in the experiment. Each treatment is a different hand-drying method. For this reason, the explanatory variable is also called the **treatment variable**.

In this experiment, researchers attempt to directly control the influence of three variables that could affect the bacteria count:

(1) Length of time participants handle the raw chicken.

- Direct control: All participants handle the raw chicken for 45 seconds.

(2) Amount of soap participants use.

- Direct control: All participants use one squirt of soap.

(3) Amount of time participants wash hands.

- Direct control: All participants wash their hands for 60 seconds.

Notice that the control works by stabilizing the impact of the confounding variable across the treatments. For example, the amount of soap will still influence the bacteria count. We cannot avoid this. But if all participants use the same amount of soap, then *differences* in bacteria count among the three treatments cannot be due to the amount of soap used.

Similarly, the amount of time that participants wash their hands will influence the bacteria count. But if all participants wash their hands for the same amount of time, then *differences* in bacteria count among the three treatments cannot be due to the amount of time participants washed their hands. This is what we mean when we say that the control works by stabilizing the impact of the confounding variable across the treatments.

Now we examine random assignment. Random assignment controls the effects of confounding variables that a researcher cannot control directly or that are difficult to identify in advance.

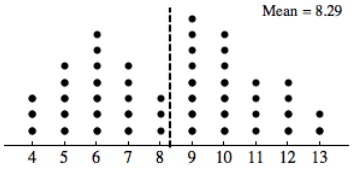
## Random Assignment

Medical researchers conducted an experiment to compare two different types of surgery for children with hernias. They compared the recovery times for each type of surgery. The two surgery types are laparoscopic repair (a surgery that involves three small incisions) and open repair (a surgery that involves one large incision). Researchers identified a variety of variables that might influence recovery time, such as child's age, weight, and physical fitness level.

Let's consider one of these variables: age. How could the researchers control the impact of age on recovery time?

Direct control involves use of children of the same age. For example, researchers might use only 10-year-old children in the experiment. But it may be difficult to find enough 10-year-old children with hernias. So how do researchers create treatment groups that are similar with respect to age? One way is to assign children at random to treatment groups.

The goal of random assignment is to create similar groups with respect to age, weight, and other characteristics that might influence recovery time. To illustrate how random assignment creates similar groups, we focus on age. Here is a dotplot of the ages of the 48 children with hernias who participated in this experiment. Each dot represents a child. The average age of the 48 children is 8.29 years.




If random assignment is working, the average age for each treatment group should be about equal. We see how random assignment works in the next activity.

Click Random Assignment to randomly assign the 48 children to the two treatments. Repeat this process several times to investigate whether random assignment creates groups with similar ages. The average age is labeled as the mean and marked with a vertical line. Compare the average ages for the treatment groups.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=42

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3829

### What Is the Main Point?

The goal of random assignment is to create similar treatment groups. If the groups are similar, then any differences we see in the response variable are due to the differences in treatments. In this way, random assignment controls the impact of confounding variables. Random assignment in an experiment eliminates confounding, just as random selection in a survey eliminates bias.

### Comment

How do we make random assignments? We use any method that allows random chance to choose the treatment for each participant. Random assignment means that each participant has an equal chance of receiving any one of the treatment options. For example, in the hernia experiment, you could put every child's name in a hat. The first 24 names drawn get the first treatment. The rest of the children get the second treatment. You could also flip a coin. Heads means the child is assigned to the first treatment. This method could create groups with slightly different sizes, which is fine.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3830

### Learn By Doing

The following paragraph is from a 1999 USA Today article titled “Heart care reflects race and sex, not symptoms.”

“Previous research suggested that blacks and women were less likely than whites and men to get cardiac catheterization or coronary bypass surgery for chest pain or a heart attack. Scientists blamed differences in illness severity, insurance coverage, patient preference, and health care access. The researchers eliminated those differences by videotaping actors – two black men, two white men, two black women, two white women – describing chest pain from identical scripts. They wore identical gowns, used identical hand gestures, and were taped from the same position. Researchers asked 720 primary care doctors at meetings of the American College of Physicians or the American Academy of Family Physicians to watch a tape and recommend care. The doctors thought the study focused on clinical decision making.”

Researchers rolled a four-sided die to determine which video each doctor watched.

https://assessments.lumenlearning.co...sessments/3409

https://assessments.lumenlearning.co...sessments/3410

https://assessments.lumenlearning.co...sessments/3831

https://assessments.lumenlearning.co...sessments/3832

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 1.11: Conducting Experiments (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.12: Conducting Experiments (2 of 2)

## Learning Objectives

- Avoid overgeneralization of experiment results.

Let's summarize what we know about experiments:

- The goal of the experiment is to provide evidence for a cause-and-effect relationship between two variables.
- A well-designed experiment controls the effects of confounding variables to isolate the effect of the explanatory variable on the response.
- Two commonly used methods for controlling the effects of confounding variables are *direct control* and *random assignment*.
- Random assignment uses random chance to assign participants to treatments. This creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.

Now we discuss a few more strategies that are commonly used to control the effects of confounding variables.

In an experiment, we manipulate the explanatory variable to determine if it has an effect on the response variable. Could the change we observe in the response variable happen without manipulating the explanatory variable? Maybe what we observe would have happened anyway.

For this reason, it is important to include a **control group**. A control group is a group that receives no treatment. The control group provides a baseline for comparison.

### Example

## Control Groups

**Music and rats:** In David Merrell's experiment with rats, he wanted to examine the relationship between music and the ability of rats to run a maze. He had three treatment groups: exposure to music by the heavy metal band Anthrax, exposure to music by Mozart, and no exposure to music. The group of rats that did not listen to music is the control group. Merrell's experiment lasted 1 month. With a month of practice, the rats in all the groups would probably get faster at running the maze. The control group provides a baseline for comparison. At the end of 1 month, the rats in the Mozart group were much faster at running the maze than were the rats in the other two groups. Comparison to the control group shows that the improvement in the Mozart group is not due to the rats being more experienced with the maze.

**Hernia treatments for children:** In this experiment, researchers compared two different surgeries. The response variable was recovery time, so it would not have made sense to have a no-treatment group. However, one type of surgery was the standard treatment for hernias, and children who received this surgery represented the control group. This group provides a baseline for comparing recovery times.

In experiments that use human participants, use of a control group may not be enough to establish whether a treatment really has an effect. A substantial amount of research shows that people respond in positive ways to treatments that have no active ingredients, a response called the **placebo effect**. A placebo is a treatment with no active ingredients, sometimes called a "sugar pill."

### Example

## The Placebo Effect

An article published in the *Washington Post* in 2002 illustrates the placebo effect in medical experiments.

> *After thousands of studies, hundreds of millions of prescriptions and tens of billions of dollars in sales, two things are certain about pills that treat depression: Antidepressants like Prozac, Paxil and Zoloft work. And so do sugar pills. A new analysis has found that in the majority of trials conducted by drug companies in recent decades, sugar pills have done as well as – or better than – antidepressants.…The new research may shed light on findings such as those from a trial last month that compared the herbal remedy St. John's wort against Zoloft. St. John's wort fully cured 24 percent of the depressed people who received it, and Zoloft cured 25 percent – but the placebo fully cured 32 percent.*

The placebo effect can confound the results of medical experiments. It is uncertain what is behind the placebo effect, but because people in medical experiments improve when taking a placebo, a placebo group provides a baseline for comparing treatments. We cannot eliminate the placebo effect on a treatment group. Both the placebo group and the drug group experience the placebo effect. If a treatment produces better results than a placebo, then we have evidence that the treatment (and not the placebo effect) is responsible for the improvement.

In experiments that use a placebo, participants do not know whether they are receiving the drug or a placebo. The participants are *blind* to the treatment to prevent their own beliefs about the drug (or placebo) from confounding the results.

### Example

## Blinding

Recall our discussion of the experiment conducted by the Women's Health Initiative to study the health implications of hormone replacement therapy. In this experiment, researchers randomly assigned over 16,000 women to one of two treatments. One group took hormones. The other group took a placebo. The experiment was also double-blind, meaning that neither the women nor the researchers knew who had which treatment.

In a *single-blind*, experiment only one of the two (either the researcher or the participants) do not know which treatment the participants receive.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3833

https://assessments.lumenlearning.co...sessments/3834

To end our discussion of experiments, we consider one last question: *If an experiment is well-designed, can we generalize the results?*

Recall the hormone replacement experiment. This experiment has all of the features of a well-designed experiment:

- A large number of participants (over 16,000 women)
- Use of a placebo group
- Random assignment of women to hormone treatment or placebo
- Double-blind design

After 5 years, the group taking hormones had a higher incidence of breast cancer and heart disease. Researchers were so alarmed by the results that the experiment was ended early to prevent further harm to the health of the women participating in the hormone group.

As a result of this experiment, the use of hormone replacement therapy fell by 66%.

Yet the British Menopause Society and the International Menopause Society questioned this reaction. The Women's Health Concern, a British nonprofit group that provides independent and unbiased information about women's health, wrote:

> *It must be remembered that the WHI data on which the concerns were raised related to overweight North American women in their mid-sixties and not to the women that are treated with HRT for their menopausal symptoms in the United Kingdom, who are usually around the age of menopause, namely 45–55 years.*

The concerns expressed here do not challenge the validity of the results of the WHI experiment. Instead, they question whether the results apply to a larger group of women: women who are younger and not overweight when they go through menopause.

This is an important consideration. If our goal is to generalize the results of an experiment to a more general population, we must consider issues of sampling design as well as random assignment.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3835

**An Important Point about the Role of Random Chance**

We now know that in an experiment we intentionally manipulate the explanatory variable to observe changes in the response variable. We use the explanatory variable to create different treatments. If we see different responses in the different treatments, we want to be able to say that the differences are the result of the explanatory variable. We must rule out other possible explanations for the differences we observed, so we use direct control and random assignment, as well as a control group, a placebo group, or blinding as appropriate.

But none of these strategies will rule out the influence of **chance variation**. When we randomly assign participants to treatments, we produce similar groups most of the time. But there is a small chance that we will end up with treatment groups that are not similar.

For example, in the hernia experiment with children, we saw that random assignment creates two groups with average ages that are close. But there is a very small chance that we could get two groups that significantly differ in ages. This will not happen very often, but it could. And if it does happen, the results of our experiment are confounded by age.

Similarly, when we investigated how well a random sample estimates the proportion of students receiving financial aid in the population, we saw that the proportions from random samples gave good estimates – most of the time. Occasionally, a random sample did not give a good estimate. Larger random samples varied less, but they still varied.

### What's the Main Point?

Good study design is important. Random selection in sampling can control bias. Random assignment in experiments can control the effects of confounding variables. But there is always a small chance, even when we randomly sample, that the results we observe in a poll do not represent the population well. Similarly, there is always a small chance, even when we use random assignment, that the differences we observe in an experiment are due to random variation and not the explanatory variable. For this reason, we have to understand how random chance behaves. This is the role of probability. We study probability later in the course, before we learn more formal statistical methods for determining if what we observe could be a result explained by chance.

### Let's Summarize

- The goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- A well-designed experiment controls the effects of confounding variables to isolate the effect of the explanatory variable on the response.
- Two commonly used methods for controlling the effects of confounding variables are *direct control* and *random assignment*.
- Random assignment uses random chance to assign participants to treatments, which creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.
- Other strategies for controlling confounding variables include use of a control group, use of a placebo group, and blinding.
- A well-designed experiment provides evidence for a cause-and-effect relationship. But even in a well-designed experiment, differences in the response might be due to chance. We learn to describe chance behavior when we study probability later in the course.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.12: Conducting Experiments (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 1.13: Putting It Together- Types of Statistical Studies and Producing Data

## Let's Summarize

- There are four steps in a statistical investigation:
  - Ask a question that can be answered by collecting data.
  - Decide what to measure, and then collect data.
  - Summarize and analyze.
  - Draw a conclusion, and communicate the results.
- There are two types of statistical studies:
  - Observational studies: An *observational study* observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an association between two variables. An observational study alone does not provide convincing evidence of a cause-and-effect relationship.
  - Experiments: An *experiment* intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- In statistics, a *variable* is information we gather about individuals or objects.

### Observational Studies

- In an observational study, we draw a conclusion about the population on the basis of a sample. To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population. It also reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome. Voluntary response samples (such as Internet polls) and convenience samples (such as surveys at a mall) are biased.
- Random selection eliminates bias. In a simple random sample, everyone in the population has an equal chance of being chosen. In this way, random selection helps ensure that the sample is representative of the population.
- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly. The size of the population does not affect the accuracy of a random sample as long as the population is large.
- If an attempt is made to include every individual from a population in a sample, then the investigation is called a *census.*

### Experiments

The goal of the experiment is to provide evidence for a cause-and-effect relationship between two variables. When we investigate a relationship between two variables, we identify an explanatory variable and a response variable. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only factor that impacts the response variable. But other factors, called *confounding variables*, may also influence the response.

- A well-designed experiment takes steps to eliminate the effects of confounding variables. These steps include direct control, random assignment of people to treatment groups, use of a control or placebo, and blind conditions. Incorporating such precautions, a well-designed experiment provides convincing evidence of cause-and-effect.
- Random assignment uses random chance to assign participants to treatments, which creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.
- A well-designed experiment provides evidence for a cause-and-effect relationship. But even in a well-designed experiment, differences in the response might be due to chance. We learn to describe chance behavior when we study probability later in the course.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.13: Putting It Together- Types of Statistical Studies and Producing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 2: Summarizing Data Graphically and Numerically


2.1: Why It Matters- Summarizing Data Graphically and Numerically
2.2: Introduction- Categorical vs. Quantitative Data
2.3: Categorical vs. Quantitative Data
2.4: Introduction- Dotplots
2.5: Dotplots (1 of 2)
2.6: Dotplots (2 of 2)
2.7: Introduction- Histograms
2.8: Histograms (1 of 4)
2.9: Histograms (2 of 4)
2.10: Histograms (3 of 4)
2.11: Histograms (4 of 4)
2.12: Introduction- Measures of Center
2.13: Mean and Median (1 of 2)
2.14: Mean and Median (2 of 2)
2.15: Introduction- Measures of Spread
2.16: Interquartile Range and Boxplots (1 of 3)
2.17: Interquartile Range and Boxplots (2 of 3)
2.18: Interquartile Range and Boxplots (3 of 3)
2.19: Introduction- Describing a Distribution
2.20: Standard Deviation (1 of 4)
2.21: Standard Deviation (2 of 4)
2.22: Standard Deviation (3 of 4)
2.23: Standard Deviation (4 of 4)
2.24: Putting It Together- Summarizing Data Graphically and Numerically
2.25: StatTutor- Drinking Habits of College Students
2.26: Assignment- Histogram
2.27: Assignment- Five-Number Summary
2.28: Assignment- Boxplot
2.29: Assignment- Standard Deviation



This page titled 2: Summarizing Data Graphically and Numerically is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.1: Why It Matters- Summarizing Data Graphically and Numerically

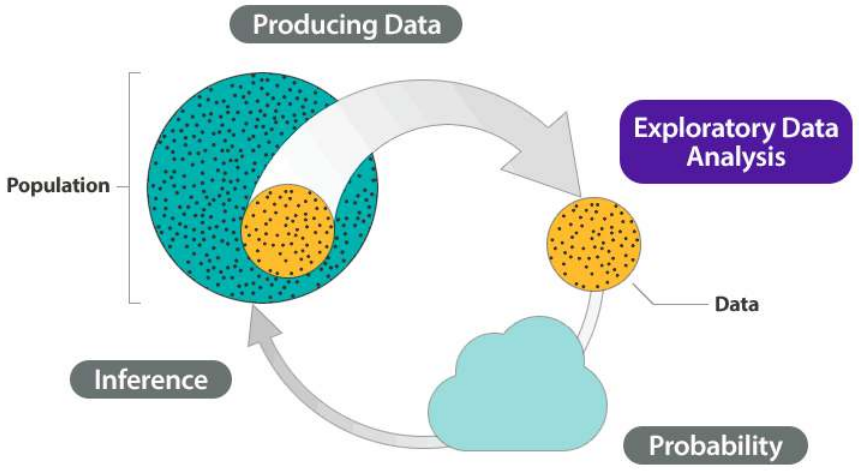
Before we begin *Summarizing Data Graphically and Numerically*, let's see how the new ideas in this module relate to what we learned in the previous module, *Types of Statistical Studies and Producing Data*.

Recall the Big Picture:

We begin a statistical investigation with a research question. The investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data. ← **Types of Statistical Studies and Producing Data**
- Explore the Data: Analyze and summarize the data (also called exploratory data analysis). ← **Summarizing Data Graphically and Numerically**
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

The previous module focused on methods for collecting reliable data. In this module, we focus on summarizing and analyzing data. In the Big Picture of Statistics, we call this **exploratory data analysis**.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.1: Why It Matters- Summarizing Data Graphically and Numerically is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.2: Introduction- Categorical vs. Quantitative Data

## What you'll learn to do: Distinguish between quantitative and categorical variables in context.

### LEARNING OBJECTIVES

- Distinguish between quantitative and categorical variables in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.2: Introduction- Categorical vs. Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.3: Categorical vs. Quantitative Data

## Learning Objectives

- Distinguish between quantitative and categorical variables in context.

**Data** consist of **individuals** and **variables** that give us information about those individuals. An individual can be an object or a person. A variable is an attribute, such as a measurement or a label.

## Example

## Medical Records

This dataset is from a medical study. In this study, researchers wanted to identify variables connected to low birth weights.

| | Age at delivery | Weight prior to pregnancy (pounds) | Smoker | Doctor visits during 1st trimester | Race | Birth Weight (grams) |
|---|---|---|---|---|---|---|
| Patient 1 | 29 | 140 | Yes | 2 | Caucasian | 2977 |
| Patient 2 | 32 | 132 | No | 4 | Caucasian | 3080 |
| Patient 3 | 36 | 175 | No | 0 | African-Am | 3600 |
| * | * | * | * | * | * | * |
| * | * | * | * | * | * | * |
| Patient 189 | 30 | 95 | Yes | 2 | Asian | 3147 |

In this example, the individuals are the patients (the mothers). There are six variables in this dataset:

- Mother's age at delivery (years)
- Mother's weight prior to pregnancy (pounds)
- Whether mother smoked during pregnancy (yes, no)
- Number of doctor visits during first trimester of pregnancy
- Mother's race (Caucasian, African American, Asian, etc.)
- Baby's birth weight (grams)

There are two types of variables: quantitative and categorical.

- **Categorical variables** take category or label values and place an individual into one of several groups. Each observation can be placed in only one category, and the categories are mutually exclusive. In our example of medical records, smoking is a categorical variable, with two groups, since each participant can be categorized only as either a nonsmoker or a smoker. Gender and race are the two other categorical variables in our medical records example.
- **Quantitative variables** take numerical values and represent some kind of measurement. In our medical example, age is an example of a quantitative variable because it can take on multiple numerical values. It also makes sense to think about it in numerical form; that is, a person can be 18 years old or 80 years old. Weight and height are also examples of quantitative variables.

## Learn By Doing

We took a random sample from the 2000 US Census. Here is part of the dataset.

US Census 2000

| | State | zipcode | Family_Size | Annual_income |
|---|---|---|---|---|
| 1 | Florida | 32716 | 8 | 200 |
| 2 | Alabama | 35236 | 5 | 800 |
| 3 | Florida | 32116 | 6 | 13500 |
| 4 | Florida | 33679 | 5 | 21000 |
| 5 | Alabama | 36374 | 4 | 21000 |
| 6 | California | 94565 | 1 | 23000 |

https://assessments.lumenlearning.co...sessments/3411

https://assessments.lumenlearning.co...sessments/3412

## Learn By Doing

*Consumer Reports* analyzed a dataset of 77 breakfast cereals. Here is a part of the dataset.

(Note: Consumer Reports is an non-profit organization that rates products in an effort to help consumers make informed decisions.)

Cereals

| | Name | Manufacturer | Target | Shelf | Calories | Sodium | Fat |
|---|---|---|---|---|---|---|---|
| 1 | 100% Bran | Nabisco | adult | top | 70 | 130 | 1 |
| 2 | 100% Natural Bran | Quaker Oats | adult | top | 120 | 15 | 5 |
| 3 | All-Bran | Kelloggs | adult | top | 70 | 260 | 1 |
| 4 | All-Bran Extra Fiber | Kelloggs | adult | top | 50 | 140 | 0 |
| 5 | Almond Delight | Ralston Purina | adult | top | 110 | 200 | 2 |
| 6 | Apple Cinnamon Cheerios | General Mills | child | bottom | 110 | 180 | 2 |
| 7 | Apple Jacks | Kelloggs | child | middle | 110 | 125 | 0 |

https://assessments.lumenlearning.co...sessments/3415

https://assessments.lumenlearning.co...sessments/3416

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.3: Categorical vs. Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.4: Introduction- Dotplots

## What you'll learn to do: Describe the distribution of quantitative data using a dotplot.

### LEARNING OBJECTIVES

- Describe the distribution of quantitative data using a dotplot.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.4: Introduction- Dotplots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.5: Dotplots (1 of 2)

## Learning Objectives

- Describe the distribution of quantitative data using a dot plot.

## Introduction

When we work with data, the data is usually in a table. In this form, we can easily see the variable value for each individual. But when we analyze data, we are not focused on information about an individual. We want to describe a group of individuals. In data analysis, our goal is to describe patterns in the data and create a useful summary about a group. A table is not a useful way to view data because patterns are hard to see in a table. For this reason, our first step in data analysis is to create a graph of the **distribution** of the variable.

In a graph that summarizes the distribution of a variable, we can see

- the possible values of the variable.
- the number of individuals with each variable value or interval of values.

In this module, *Summarizing Data Graphically and Numerically*, we focus on summarizing the distribution of a quantitative variable. We discuss the distribution of a categorical variable in depth in the module *Relationships in Categorical Data with Intro to Probability*.

## Example

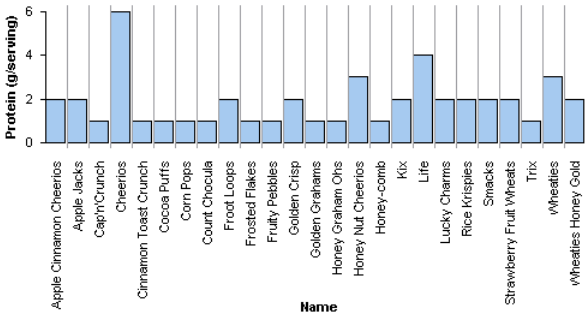
## Breakfast Cereals

Here are two graphs of the variable *protein* for a group of breakfast cereals targeted at children.

In both graphs, the individuals and the variable are the same:

- Individuals: Children's cereals
- Variable: Grams of protein in a serving of cereal

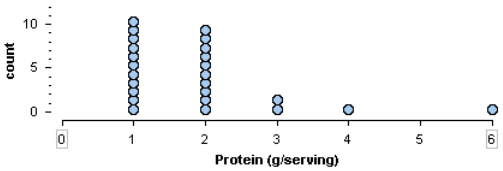
Let's compare the graphs to determine which graph is a better summary of the distribution of protein.




This graph is called a **case-value graph**. You can see the names of the individual cereals (the cases) and the amount of protein in a serving of each cereal (the variable values). For example, Apple Jacks has 2 grams of protein in a serving. This graph is NOT a good way to summarize the distribution of protein values because we cannot easily determine the number of cereals with each protein amount.

For example, how many cereals have 2 grams of protein in a serving? This graph does not make it easy to answer this question. We have to move across the graph and count the cereals with 2 grams of protein. In this way, a case-value graph is like a table. We cannot easily see patterns in the data or determine the number of individuals with a given variable value.

Here is a second graph of the same data. This graph is called a **dotplot**. A dotplot gives a better summary of the distribution of protein.




In a dotplot, each dot represents one individual. Here, each dot is a children's cereal. The numbers on the horizontal axis are the variable values. The vertical axis gives the count of cereals. We can easily see that 10 children's cereals have 2 grams of protein in a serving.

From the dotplot, we can easily describe the distribution of protein. Here are some observations about this distribution:

- The amount of protein in a serving varies from 1 to 6 grams.
- Most of the cereals have 1 or 2 grams of protein in a serving.
- Larger amounts of protein are less typical.
- One cereal has 6 grams of protein. This much protein is unusual for this group of children's cereals.

These observations are a good summary of the data.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3421

https://assessments.lumenlearning.co...sessments/3422

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.5: Dotplots (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

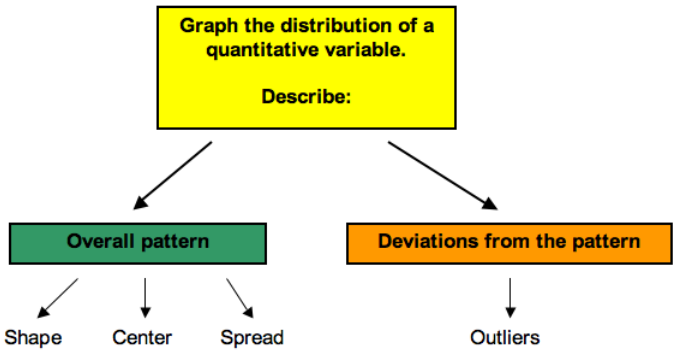
# 2.6: Dotplots (2 of 2)

## Learning Objectives

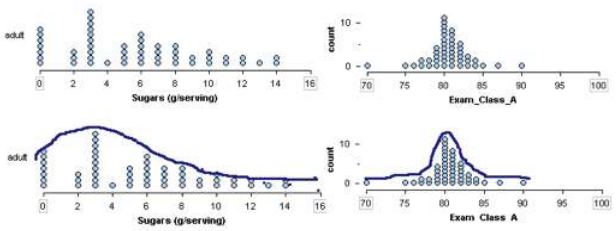
- Describe the distribution of quantitative data using a dot plot.

Now we will give more specific advice on how to describe the distribution of a quantitative variable.

When we describe patterns in data, we use descriptions of **shape**, **center**, and **spread**. We also describe exceptions to the pattern. We call these exceptions **outliers**.




**Shape:** To describe the shape of a distribution, imagine sketching the outline of the data to emphasize the general trend.

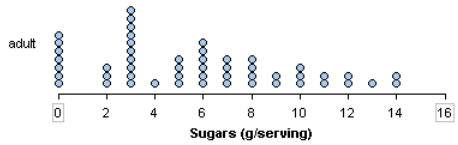



## Example

## Some Common Descriptions of Shape Used to Categorize Distributions

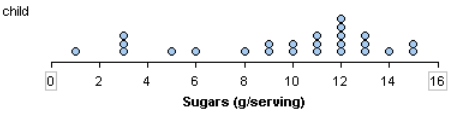
**Right skewed:** A cluster of data on the left with a *tail* of data tapering off to the right. A right-skewed distribution has a lot of data at lower variable values with smaller amounts of data at higher variable values.

- The distribution of sugar in adult cereals is *skewed to the right.*



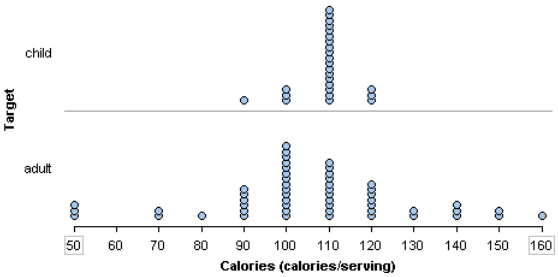

**Left skewed:** A cluster of data on the right with a tail of data tapering off to the left. A left skewed distribution has a lot of data at higher variable values with smaller amounts of data at lower variable values.

- The distribution of sugar in children's cereals is *skewed to the left.*



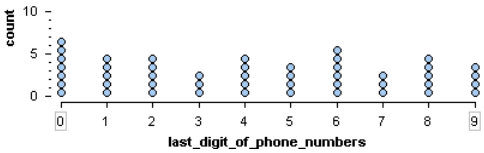
**Symmetric with a central peak (also called *bell-shaped*):** A central peak with a tail in both directions. A bell-shaped distribution has a lot of data in the center with smaller amounts of data tapering off in each direction.

- The distribution of calories in children's cereals is *symmetric with a central peak*. It is **bell-shaped**. The distribution of calories in adult cereals is also roughly bell-shaped.




**Uniform:** A rectangular shape, the same amount of data for each variable value.

- Here is the last digit from 47 students' telephone numbers. The distribution of the digits is roughly **uniform**.




### Learn By Doing

https://assessments.lumenlearning.co...sessments/3423

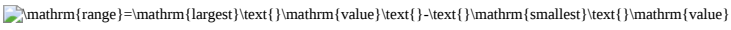
### Center and Spread

To describe the pattern in a distribution of a quantitative variable, we describe more than the shape. We also describe the center and spread. Later in this module, we develop more precise ways to identify the center of a distribution and to measure the spread. For now, we discuss these concepts informally.

When we describe a distribution of a quantitative variable, it is helpful to identify a typical value. We choose a single value of the variable to represent the entire group. This is one way to think about the center of the distribution.

We also want to describe how much the data varies among individuals in the group. **Variability** is another word for spread. We describe the spread in two ways:

- We look at the smallest value and the largest value to describe an **overall range** in the data.



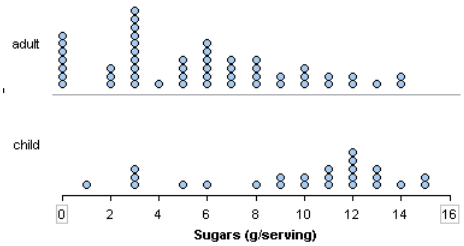

- We also describe a **range of typical values** to represent common variable values for the group.

### Example

## Cereals

Here we use shape, center, and spread to compare the distribution of sugar content in adult cereals and children's cereals.

**Compare the shapes:**



The sugar content in adult cereals is skewed to the right. Many adult cereals have less than 8 grams of sugar in a serving. A smaller number of adult cereals contain high amounts of sugar. The sugar content for children's cereals is skewed to the left. Many children's cereals have more than 8 grams of sugar in a serving, with a smaller number of children's cereals with low amounts of sugar.

*Comment:* There is nothing special about the number 8. We chose 8 as a convenient reference point to describe the opposite trends in these two distributions.

**Compare the centers:**

A typical adult cereal has 3 grams of sugar in a serving. A typical children's cereal has 12 grams of sugar in a serving.

*Comment:* Here we looked at the most common value in each distribution. We develop more precise ways to describe the center of a distribution in the next section. For now, just choose a reasonable typical value to represent the group.

**Compare the spreads:**

**Overall range:** Adult cereals have 0 to 14 grams of sugar in a serving. Children's cereals vary from 1 to 15 grams. So both types of cereal vary over a range of 14 grams.

(Note: Overall range = highest value – lowest value. For adult cereals: 14 – 0 = 14. For children's cereals: 15 – 1 = 14)

**Typical range:** Typical adult cereals have between 0 and 6 grams of sugar in a serving, compared to 9 to 13 grams in typical children's cereals.

*Comment:* Here we looked at clumps in the data to identify a range of typical values. We develop more precise ways to describe the spread a distribution in the last two sections of this module.

When comparing two distributions, we usually tie all of these ideas into one paragraph:

In this sample, children's cereals have more sugar per serving than adult cereals. A typical children's cereal has 12 grams of sugar in a serving. It is not uncommon for children's cereals to have 9 to 13 grams of sugar per serving, but it is unusual for a children's cereal to have less than 8 grams of sugar. A typical adult cereal has 3 grams of sugar in a serving. It is not uncommon for adult cereals to have 0 to 6 grams of sugar in a serving. Larger amounts of sugar are less common.

Here is a paragraph that uses more formal vocabulary to summarize the comparison:

In this sample, children's cereals have more sugar per serving than adult cereals. The distribution of sugar in children's cereals is skewed left with an overall range of 14 grams. Typical children's cereals have 9 to 13 grams of sugar per serving with 12 grams as the most common amount. The distribution of sugar in adult cereals is skewed right with the same overall range of 14 grams. Typical adult cereals have 0 to 6 grams of sugar per serving with 3 grams as the most common amount.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3836

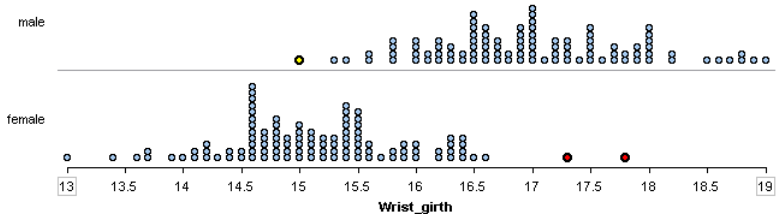
**Outliers:** Outliers are observations that fall outside the overall pattern. We develop a more precise method for identifying outliers later in this module. For now, use your judgment to identify values that appear to be exceptions to the general trend in the data.

### Example

## Wrist Measurements

In the distribution of wrist measurements, there are two women with unusually large wrists. These women might be outliers. They are marked in red.

The man with the smallest wrist measurement is shown in yellow. This man is probably not an outlier.




## Learn By Doing

https://assessments.lumenlearning.co...sessments/3424

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.6: Dotplots (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.7: Introduction- Histograms

## What you'll learn to do: Describe the distribution of quantitative data using a histogram.

### LEARNING OBJECTIVES

- Describe the distribution of quantitative data using a histogram.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.7: Introduction- Histograms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.8: Histograms (1 of 4)

## Learning Objectives

- Describe the distribution of quantitative data using a histogram.

Here we continue our discussion of graphs that describe the distribution of a quantitative variable.

Recall that our goal in data analysis is to describe patterns in data and create a useful summary about a group. When a graph summarizes the distribution of a variable, we can see

- the possible values of the variable.
- the number of individuals with each variable value or interval of values.

As we have seen, a dotplot is a useful graphical summary of a distribution.

A **histogram** is an alternative way to display the distribution of a quantitative variable. Histograms are particularly useful for large data sets. A histogram divides the variable values into equal-sized intervals. We can see the number of individuals in each interval.

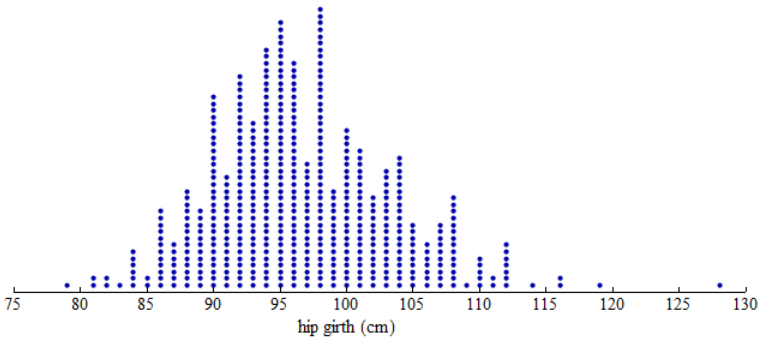
## Example

## A Histogram of Hip Measurements

Here we have three graphs of the same set of hip girth measurements for 507 adults who exercise regularly. (*Hip girth* is the measurement around the hips.)

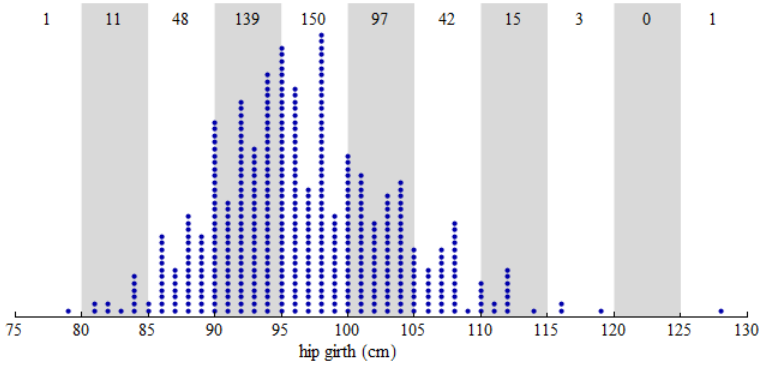
**Dotplot:**

From the dotplot, we can see that the distribution of hip measurements has an overall range of 79 to 128 cm. For convenience, we started the axis at 75 and ended the axis at 130.



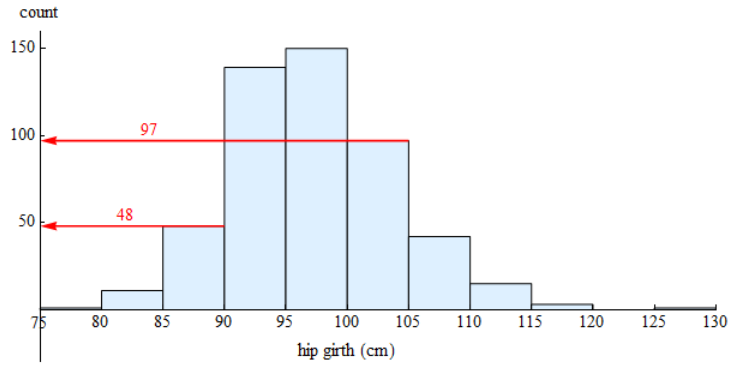

**Dotplot with Bins:**

To create a histogram, divide the variable values into equal-sized intervals called **bins**. In this graph, we chose bins with a width of 5 cm. Each bin contains a different number of individuals. For example, 48 adults have hip measurements between 85 and 90 cm, and 97 adults have hip measurements between 100 and 105 cm.



**Histogram:**

Here is a histogram. Each bin is now a bar. The height of the bar indicates the number of individuals with hip measurements in the interval for that bin. As before, we can see that 48 adults have hip measurements between 85 and 90 cm, and 97 adults have hip measurements between 100 and 105 cm.




**Comment:** In the histogram, the count is the number of individuals in each bin. The count is also called the **frequency**. From these counts, we can determine a percentage of individuals with a given interval of variable values. This percentage is called a **relative frequency**.

The following questions require us to calculate relative frequencies:

- Approximately what percentage of the sample has hip measurements between 85 and 90 cm?

**Answer:** Of the 507 adults in the data set, 48 have hip measurements between 85 and 90 cm.

48 out of 507 is 48 ÷ 507 ≈ 0.095 = 9.5%

So approximately 9.5% of the adults in this sample have hip girths between 85 and 90 cm.

(This calculation might include adults with as 85-cm hip measurement but not adults with a 90-cm hip measurement. See note below.)

- A pants manufacturer plans to produce three sizes of sweatpants. Size Large will fit hip girths of 100 cm or more. What percentage of the sample will wear size Large sweatpants?

**Answer:** Of the 507 adults in the data set, 158 adults (97 + 42 + 15 + 3 + 1) = 158 have hip measurements of 100 cm or more.

158 out of 507 is 158 ÷ 507 ≈ 0.312 = 31.2%

So 31.2% of the adults in this sample will wear size Large sweatpants.

Note: In these calculations, we assume that the value of the left-hand endpoint of each bin is included in the count for that bin. The value of the right-hand endpoint is not included in the count for that bin. For example, the bin corresponding to the interval 85 to 90 includes individuals with values of 85 but not 90. In histograms pictured in this course, bins will always include values for the left-hand endpoint but not the right-hand endpoint.

**Spotlight on percentages**

Percent means “per hundred.” A percentage describes a number as a fraction out of 100.

EXAMPLE

## What percentage of adults in this sample wear a large size sweatpants?

1. Identify the appropriate ratio: 158 out of 507 adults will wear large size sweatpants.
2. Calculate a percentage:
   - Divide to convert the ratio into a decimal form: 158÷507 ≈ 0.312
   - Multiply by 100 to convert the decimal form to a percentage: 0.312 x 100 = 31.2%
   - 31.2% is 31.2 out of 100

3. Interpret the percentage:
    - For every 100 adults in the sample, 29.6 will wear a large.
    - 29.6% of the adults in this sample wear large sweatpants.

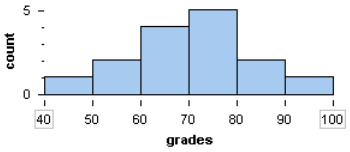
**General steps:**

1. Identify the appropriate ratio: You can think of the ratio as a fill-in-the-blank: ***(a part) out of (the group)***
    - The "part" is often a subset of the group with a special characteristic.
2. Calculate the percentage:
    - Divide: (part) ÷ (group size)
    - Multiply by 100
3. Interpret the percentage in context:

For every 100 individuals in the group, (the percentage) will have the special characteristic. You can interpret the percentage as: **Percentage of (group) has (special characteristic).**

## Learn By Doing

Here is a histogram of the distribution of grades on a quiz.




https://assessments.lumenlearning.co...sessments/3425

https://assessments.lumenlearning.co...sessments/3426

This next exercise will remind us when to use a histogram.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3431

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.8: Histograms (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.9: Histograms (2 of 4)

## Learning Objectives

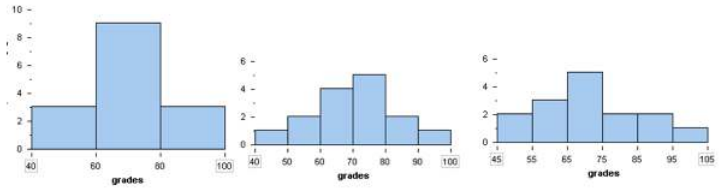
- Describe the distribution of quantitative data using a histogram.

We have discussed two types of graphs that summarize a distribution of a quantitative variable: dotplots and histograms.

From a dotplot, we also described the pattern in the data with statements about shape, center, and spread. We have to be more cautious making similar statements using a histogram because our perception of shape, center, and spread can be affected by how the bins are defined. We investigate this important point in the next example.

## Example

We used the *same set of data* to construct these three histograms of student scores. Are you surprised by how different the distribution looks in each histogram?




The histogram on the left has a bin width of 20. The first bin starts at 40. To create the middle histogram, we changed the bin width to 10 but kept the first bin starting at 40. To create the last histogram, we kept the bin width at 10 but started the first bin at 45.

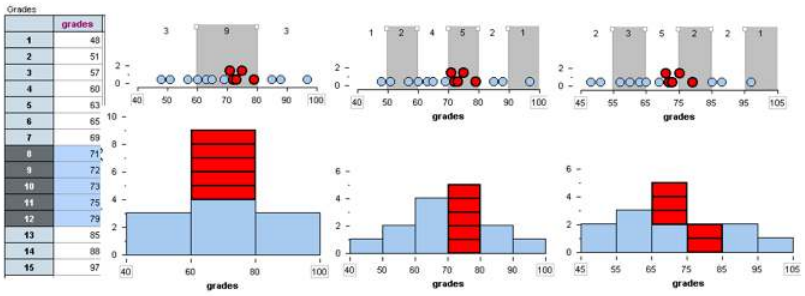
These changes affect our description of the shape, center, and spread of this set of data. For example, in the histogram on the left, the distribution looks symmetric with a central peak. In the histogram on the right, the distribution looks slightly skewed to the right. Based on the middle histogram, we might estimate that most students scored between 70 and 80. But the histogram on the right suggests that typical students scored between 65 and 75.

**Why does changing the bin size and the starting point of the first bin change the histogram so drastically?**

When we change the bins, the data gets grouped differently. The different grouping affects the appearance of the histogram.

To illustrate this point, we highlighted the five students who scored in the 70s in each histogram.

- In the histogram on the left, these five students are grouped in the middle bin with other students who scored between 60 and 80.
- In the histogram in the middle, these five students form a bin of their own, since no other students scored between 70 and 80.
- In the histogram on the right, these five students are in separate bins.




**Which histogram gives the most helpful summary of the distribution?**

For this situation, the middle histogram is probably the most useful summary because the intervals correspond to letter grades.

Our general advice is as follows:

- Avoid histograms with large bin widths that group data into only a few bins. A histogram constructed with large bin widths will show the distribution as a “skyscraper.” This does not give good information about variability in the distribution.

- Avoid histograms with small bin widths that group data into lots of bins. A histogram constructed with small bin widths will show the distribution as a "pancake." This does not help us see the pattern in the data.

Use the simulation below to answer the questions in the next Learn By Doing.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=66

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3432

https://assessments.lumenlearning.co...sessments/3433

https://assessments.lumenlearning.co...sessments/3837

These next exercises focus on recognizing the shape of a distribution using a histogram. We know that changes in the bin width can change the appearance of the distribution. But a histogram with an appropriate bin width can give good information about the shape of the distribution.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3434

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3435

https://assessments.lumenlearning.co...sessments/3436

https://assessments.lumenlearning.co...sessments/3437

https://assessments.lumenlearning.co...sessments/3438

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.9: Histograms (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.10: Histograms (3 of 4)

## Learning Objectives

- Describe the distribution of quantitative data using a histogram.

In the next example, we use a histogram to describe the shape, center, and spread of the distribution of a quantitative variable.

## Example

## Oscar for Best Actress

Here we have the ages of the actresses who won an Oscar for Best Actress from 1970 to 2001.

Click here to see the entire data set.

**Shape:** The distribution of ages appears skewed to the right. Most of the Oscar winners for Best Actress are young. More precisely, we see that 91% (29 of 32) of the winners under 50 years of age, and 56% (18 of 32) of the winners are under the age of 35.

**Center:** The distribution of ages appears to be centered between 30 and 35 years; 28% (9 of 32) of the winners are in this age range.

**Spread:** The data range from about 20 to about 80, so the overall range is approximately 60. There is a lot of variability in the ages of actresses who have won the Oscar for Best Actress.

**Outliers:** Winners older than 60 years are unusual. There are three outliers: one in each of the following intervals: 60–65, 70–75, 75–80.

Now we summarize all of these observations in a paragraph:

Between the years of 1970 and 2001, the Oscar for Best Actress was awarded most often to young actresses: 56% (18 of 32) of the winners were under the age of 35, with 28% (9 of 32) of the winners between 30 and 35 years of age. Winners ranged in age from about 20 to about 80, but only 3 of the 32 were over 60.

Here is a paragraph that uses more formal vocabulary to summarize the distribution of ages:

Between the years of 1970 and 2001, the Oscar for Best Actress was awarded most often to young actresses. The distribution of ages is skewed to the right: 56% (18 of 32) of the winners were under the age of 35, with the center of the distribution between 30 and 35 years of age. With winners ranging in age from about 20 to about 80, the overall range of the distribution is about 60. But much of this variability is due to three outliers who were older than 60 when they won the award.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.10: Histograms (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.11: Histograms (4 of 4)

## Learning Objectives

- Describe the distribution of quantitative data using a histogram.

We now use histograms to compare the distributions of a quantitative variable for two groups of individuals. Previously, we did a similar comparison using dotplots. As before, our descriptions focus on the overall pattern (shape, center, and spread) as well as deviations from the pattern (outliers). We also use percentages to describe and compare different intervals of variable values, since histograms make it easy to do so.

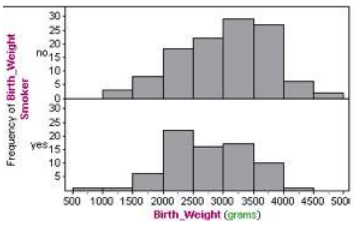
## Example

## Smoking and Birth Weight

Does smoking during pregnancy have an impact on birth weight? To investigate this question, doctors collected data on 189 new mothers who gave birth at a hospital in Massachusetts during the 1980s.

Here we use histograms to compare the distribution of birth weights for mothers who smoked during pregnancy with mothers who did not smoke. The table shows the numbers of mothers with babies in each interval of birth weights. (Left endpoints are included in the bin, so a 1,000-gram baby is in the interval 1,000–1,500 grams.)

Note: For easy and more accurate visual comparisons, both histograms have the same horizontal scale and bin width. Also, the scale on the vertical axis is the same. So we can directly compare the heights of the bars to compare the number of mothers with babies in each interval of birth weights.




| | Smoker No | Smoker Yes |
|---|---|---|
| 500-1000 | 0 | 1 |
| 1000-1500 | 3 | 1 |
| 1500-2000 | 8 | 6 |
| 2000-2500 | 18 | 22 |
| 2500-3000 | 22 | 16 |
| 3000-3500 | 29 | 17 |
| 3500-4000 | 27 | 10 |
| 4000-4500 | 6 | 1 |
| 4500-5000 | 2 | 0 |
| | total = 115 | total =74 |

Following are some observations about the shape, center, and spread:

**Nonsmokers:** The distribution of birth weights for the nonsmokers appears skewed slightly to the left. We estimate that birth weights for this group fall between approximately 1,000 and 5,000 grams for an overall range of approximately 4,000 grams. For nonsmokers, nearly half of the babies have a birth weight between 3,000 and 4,000 grams (29 + 27 = 56, 56/115 = 48.7%) with fewer babies in the lower weight ranges.

**Smokers:** The distribution of birth weights for the smokers appears slightly skewed to the right. We estimate the birth weights for this group fall between approximately 500 and 4,500 grams for an overall range of approximately 4,000 grams. For smokers, nearly half of the babies have a birth weight between 2,000 and 3,000 grams (16 + 22 = 38, 38 / 74 = 51%) with fewer babies in heavier weight ranges.

Comment: As we have seen, the choice of bin width can affect the shape of a histogram. We also cannot make precise statements about center and spread because our sense of “typical” range is also affected by the choice of bin width.

Another strategy for comparing distributions is to use a **benchmark**. Here are some examples:

1. Doctors define *low birth weight* as a birth weight below 2,500 grams. Calculate and compare the percentage of smokers and nonsmokers with low-birth-weight babies by this definition.Nonsmokers: Of babies born to mothers who did not smoke, 3 + 8 + 18 = 29 weighed less than 2,500 grams, so 25.2% (29 of 115) of the babies born to nonsmokers fit the definition of low birth weight.Smokers: Of babies born to mothers who smoked, 1 + 1 + 6 + 22 = 30 weighed less than 2,500 grams, so 40.5% (30 of 74) of the babies born to smokers fit the definition of low birth weight.
2. A condition called *macrosomia* (also known as big baby syndrome) is defined as a birth weight of 4,000 grams or more. Calculate and compare the percentage of smokers and nonsmokers with babies that fit the definition of macrosomia.Nonsmokers: Of babies born to mothers who did not smoke, 6 + 2 = 8 weighed 4,000 grams or more, so 7.0% (8 of

115) of the babies born to nonsmokers fit the definition of macrosomia.Smokers: Of babies born to mothers who smoked, only 1 weighed 4,000 grams or more, so 1.4% (1 of 74) of the babies born to smokers fit the definition of macrosomia.

**Now we synthesize these observations into a paragraph.**

Tip: Be sure to emphasize the comparison of the groups. Develop a thesis statement if appropriate.

In this observational study, we compared mothers who smoked during pregnancy to mothers who did not smoke during pregnancy. The variable is the birth weights of their babies. Both groups had a lot of variability in birth weights, with identical overall range estimates of 4,000 grams.

There was also a lot of overlap in the distributions. Nonsmokers had babies that weighed between approximately 1,000 and 5,000 grams. Smokers had babies that weighed between approximately 500 and 4500 grams.

However, we also observe some important differences in the typical ranges of birth weights for the two groups. For nonsmokers, nearly half of the babies have a birth weight between 3,000 and 4,000 grams (56 out of 115, 48.7%) with fewer babies in the lower weight ranges. For smokers, nearly half of the babies have a birth weight between 2,000 and 3,000 grams (40 of 74, 54%) with fewer babies in heavier weight ranges.

If we use the medical definition of low birth weight (under 2,500 grams), we see that smokers in this study have a much higher incidence of low birth weights: 25.2% (29 of 115) of the babies born to nonsmokers fit the definition of low birth weight, compared to 40.5% (30 of 74) of the babies born to smokers. In this study, smoking is associated with lower birth weights, though the variability in the data suggests that other variables also contribute to birth weight.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3838

## Let's Summarize

In "Distributions for Quantitative Data," we focused on describing the *distribution of a quantitative variable*.

- In a graph that summarizes the distribution of a quantitative variable, we can see
  - the possible values of the variable.
  - the number of individuals with each variable value or interval of values.
- To analyze the distribution of a quantitative variable, we described the *overall pattern of the data* (shape, center, spread), and any *deviations from the pattern* (outliers).
  - We described the *shape* of a distribution as left-skewed, right-skewed, symmetric with a central peak (bell-shaped), or uniform. Not all distributions have a simple shape that fits into one of these categories.
  - The *center* of a distribution is a typical value that represents the group. We discuss ways to identify the center of a distribution in "Measures of Center."
  - The *spread* of a distribution is a description of how the data varies. One measurement of spread is the overall range of the data (largest value – smallest value). We also looked at a typical range of values. We discuss ways to identify a typical range in "Quantifying Variability Relative to the Median" and "Quantifying Variability Relative to the Mean."
  - *Outliers* are data points that fall outside the overall pattern of the distribution.
- We used two types of graphs to analyze the distribution of a quantitative variable:
  - Dotplots
  - Histograms
- Following are some observations about *dotplots:*
  - Individual variable values are visible, particularly when the data set is small.
  - Descriptions of shape, center, and spread are not affected by how the dotplot is constructed.
  - We can accurately calculate the overall range (largest value – smallest value).
- Following are some observations about *histograms:*
  - Individual variable values are not visible.
  - Grouping individuals into bins of equal-sized intervals is particularly useful when analyzing large data sets.

- We can easily use percentages, also called relative frequencies, to describe the distribution.
- Descriptions of shape, center, and spread are affected by how the bins are defined.

- How do we decide when to use a dotplot and when to use a histogram? There are no rules here. Each type of graph can highlight different aspects of the data.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.11: Histograms (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.12: Introduction- Measures of Center

## What you'll learn to do: Use mean and median to describe the center of a distribution.

### LEARNING OBJECTIVES

- Use mean and median to describe the center of a distribution.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

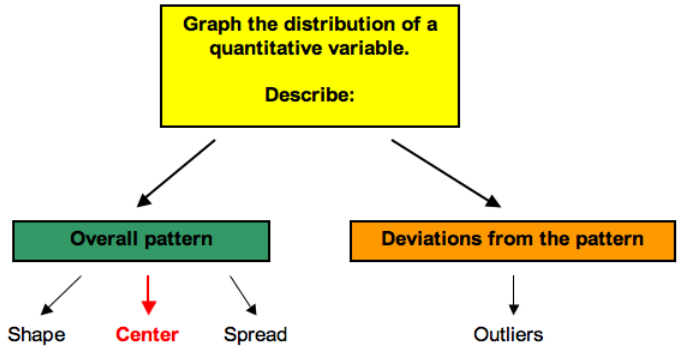
This page titled 2.12: Introduction- Measures of Center is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.13: Mean and Median (1 of 2)

## Learning Objectives

- Use mean and median to describe the center of a distribution.

Recall that when we describe the distribution of a quantitative variable, we describe the overall pattern (shape, center, and spread) in the data and deviations from the pattern (outliers). In our previous discussion of patterns in quantitative data, we identified a typical value in the distribution. We used this single value of the variable to represent the entire group. This is an informal way to think about the center of the distribution. In “Measures of Center,” we focus on describing the center of a distribution more precisely.




We develop two different measurements for identifying the center of a distribution: the mean and the median. Each measure has special properties.

## Mean

The **mean** is the average. It is written as $\stackrel{¯}{x}$ and pronounced “x-bar.” To calculate the mean, we add the data values and divide by the number of data points.

We can write this as a formula.

$$\stackrel{¯}{x}=\frac{\mathrm{\Sigma}x}{n}$$

In this formula, the symbol $\mathrm{\Sigma}$ means sum (add up the values). The x represents the data values. The letter “n” represents the number of data values.

## Example

## Calculating the Mean

Let’s find the mean of a set of three quiz scores: 70, 85, 82. In this situation, n is 3 because there are 3 quiz scores. We add the “x” values, 70 + 85 + 82 to get 237, then divide by 3 to get a mean of 79.

We could write this calculation using the formula:

$$\stackrel{¯}{x}=\frac{\mathrm{\Sigma}x}{n}=\frac{70+85+82}{3}=\frac{237}{3}=79$$

## Example

## Average Homework Score

Suppose Beth’s homework scores are70, 80, 80, 80, 85, 86, 90, 90, 95. There is variability in her homework scores, but the mean represents her typical performance on homework.

The mean of her scores is

$$\stackrel{¯}{x}=\frac{70+80+80+80+85+86+90+90+95\text{}}{9}\text{}=\text{}\frac{756}{9}=84$$

So Beth's performance on homework varies, but on average, she makes an 84 on each assignment. In other words, we can understand the mean as the score Beth would have on every assignment if she always made the same grade – that is, if she made an 84 on all nine homework assignments.

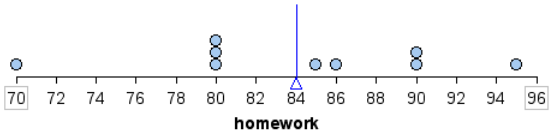
Her mean score is 84, since

$$\stackrel{\frown}{x}=\frac{84+84+84+84+84+84+84+84+84}{9}=\frac{9(84)}{9}=\frac{756}{9}=84$$

From this viewpoint, the mean is the **fair share** measure of center.

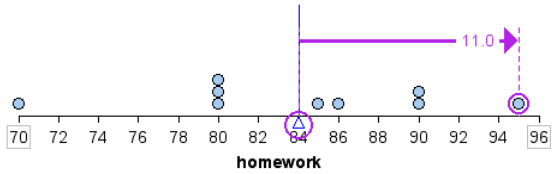
Notice, however, that Beth did not actually make an 84 on any assignment. The mean does not give us information about any individual homework score or about how the homework scores vary. It only gives us a sense of her performance by averaging the values across all the assignments.

Here is the mean marked on a dotplot of the distribution of homework scores. For this set of scores, the mean appears to be a pretty good measure of how Beth performed overall.




The mean is also referred to as the **balancing point** of a distribution. If we measure the distance between each data point and the mean, the distances are balanced on each side of the mean.

For example, a homework score of 95 is 11 points above the mean, as shown.




A homework score of 80 is 4 points below the mean. In the table, we calculate the sum of the distances above and below the mean. Notice that the sum of the distances above and below the mean are equal. In this way, the mean is a balancing point for the distribution.

| Homework scores | 70 | 80 | 80 | 80 | 85 | 86 | 90 | 90 | 95 |
|---|---|---|---|---|---|---|---|---|---|
| Distance from the mean of 84 | 14 | 4 | 4 | 4 | 1 | 2 | 6 | 6 | 11 |
| | Distances below the mean 14+4+4+4 = 26 | | | | Distances above the mean 1+2+6+6+11 = 26 | | | | |

We can also view the distances below the mean as negative and the distances above the mean as positive. When we add these "signed" distances together, we get 0

$$\mathbf{(-14)} + \mathbf{(-4)} + \mathbf{(-4)} + \mathbf{(-4)} + 1 + 2 + 6 + 6 + 11$$

$$\mathbf{(-26)} + 26$$

The mean is the only measure of center with this special property.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3840

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3439

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3440

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3841

### Median

The **median** is another way to identify a typical value. The median is the middle of the data when all the values are listed in order. The median divides the data into two equal-sized groups. There is as much data below the median as above it.

### Example

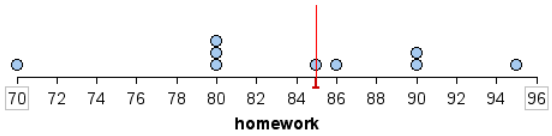
## Median Homework Score

Let's return to Beth's homework scores: 70, 80, 80, 80, **85**, 86, 90, 90, 95.

The median score is 85. This is the center score. There are four homework scores below 85 and four homework scores above 85.

For this data set, the median was one of the homework scores. This will not always be the case. So, like the mean, the median does not give us information about any individual homework score or about how the homework scores vary. It only gives us a sense of Beth's performance by locating a value that is the middle of the actual scores.

Here is the median marked on a dotplot of the distribution of homework scores. For this set of scores, the median is also a pretty good measure of how Beth performed overall.




### Learn By Doing

https://assessments.lumenlearning.co...sessments/3441

https://assessments.lumenlearning.co...sessments/3442

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.13: Mean and Median (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.14: Mean and Median (2 of 2)

## Learning Objectives

- Use mean and median to describe the center of a distribution.

## Choosing between Median and Mean

We now have a choice between two measurements of center. We can use the median, or we can use the mean. How do we decide which measurement to use?

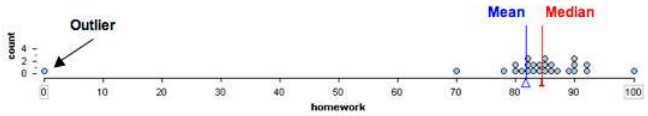
In these next examples, we learn that the shape of the distribution and the presence of outliers helps us answer this question.

### Example

## Homework Scores with an Outlier

Here is a dotplot of the 26 homework scores earned by a student. Notice that the distribution of scores has an outlier. This student typically scores between 80 and 90 on homework, but there is one score of 0. Which measurement of center gives a better summary of this distribution?

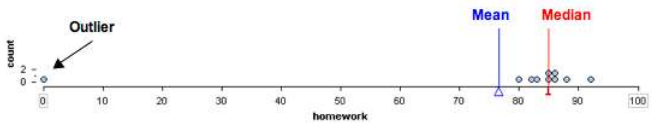
- Median = 84.5
- Mean = 81.8




Both measures of center are in the B grade range, but the median is a better summary of this student's homework scores. The outlier does not affect the median. This makes sense because the median depends primarily on the order of the data. Changing the lowest score does not affect the order of the scores, so the median is not affected by the value of this point.

The mean is not a good summary of this student's homework scores. The outlier decreases the mean so that the mean is a bit too low to be a representative measure of this student's typical performance. This makes sense because when we calculate the mean, we first add the scores together, then divide by the number of scores. Every score therefore affects the mean.

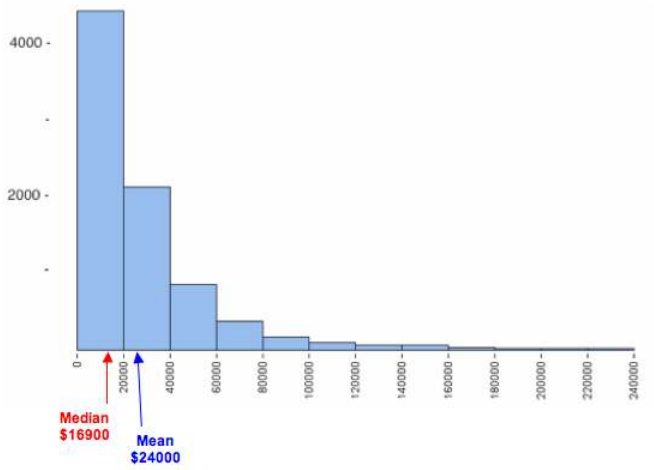
Note: In the distribution above, there are 26 homework scores for this student. If the teacher made fewer homework assignments, a zero would have a greater impact on the mean. We can see this in the distribution below. This distribution has only 10 scores. The one grade of 0 moves the mean into the C grade range.




### Example

## Skewed Incomes

In this example, we look at how skewness in a data set affects the mean and median. The following histogram shows the personal income of a large sample of individuals drawn from U.S. census data for the year 2000. Notice that it is strongly skewed to the right. This type of skewness is often present in data sets of variables such as income.




The mean and median for this data set are

- Mean = $24,000
- Median = $16,900

Here again we see that the mean income does not represent the typical income for this sample very well. The small number of people with higher incomes increase the mean. The mean is too high to represent the large number of people making less than $20,000 a year. A small number of high incomes gives the misleading impression that the typical income in the sample is $24,000. The small number of people with higher incomes does not impact the median, so the median income of $16,900 better represents the typical income in this sample.

### What's the Main Point?

These examples illustrate some general guidelines for choosing a measure of center:

- Use the mean as a measure of center *only* for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean is not a good choice.
- Use the median as a measure of center for all other cases.

Both of these examples also highlight another important principle: *Always plot the data*.

We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measures of center best describe the data.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3443

https://assessments.lumenlearning.co...sessments/3444

Instructions for using the simulation:

- To add a point, move the slider to the value you want, then click **Add**.
- To remove a point, move the slider to the value you want, then click **Minus**.
- To reset the simulation, click the button in the upper left corner that says **Reset**.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=76

### Let's Summarize

- We have two different measurements for determining the center of a distribution: mean and median. When we use the term *center*, we mean a typical value that can represent the distribution of data.

- The *mean* is the average. We calculate the mean by adding the data values and dividing by the number of individual data points.
- The mean has the following properties:
  - It is the *fair-share* measure. For example, imagine that you have 10 homework scores. Say that your scores vary, but the mean is 84. Then you have 84(10) = 840 points, which is like having an 84 on each of the 10 assignments.
  - The mean is also referred to as the *balancing point* of a distribution. If we measure the distance between each data point and the mean, the distances are balanced on each side of the mean.
- The *median* is the physical center of the data when we make an ordered list. It has the same number of values above it as below it.
- **General Guidelines for Choosing a Measure of Center**
  - Use the mean as a measure of center *only* for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean is not a good choice.
  - Use the median as a measure of center for all other cases.
- *Always plot the data.* We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measures of center best describe the data.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.14: Mean and Median (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.15: Introduction- Measures of Spread

## What you'll learn to do: Use a five-number summary and a boxplot to describe a distribution.

### LEARNING OBJECTIVES

- Use a five-number summary and a boxplot to describe a distribution.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.15: Introduction- Measures of Spread is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

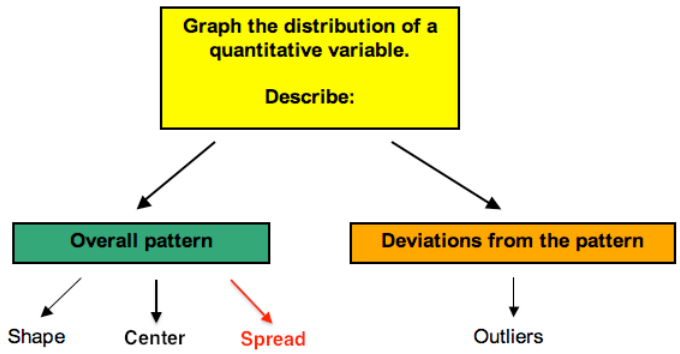
# 2.16: Interquartile Range and Boxplots (1 of 3)

## Learning Objectives

- Use a five-number summary and a boxplot to describe a distribution.

## Introduction

Recall that when we describe the distribution of a quantitative variable, we describe the overall pattern (shape, center, and spread) in the data and deviations from the pattern (outliers). In “Distributions for Quantitative Data” and “Measures of Center,” we focused on describing the shape and center of a distribution. We also investigated how the shape influences our choice of measurements of center. In “Quantifying Variability Relative to the Median” and Quantifying Variability Relative to the Mean,” we focus on describing the spread of a distribution more precisely.

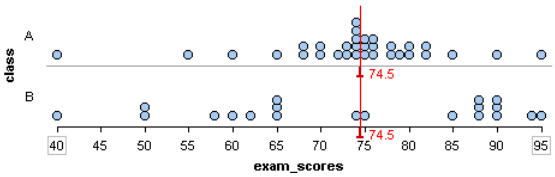



We begin with describing spread about the median.

## Example

### Two Sets of Exam Scores

Consider the following two distributions of exam scores:




Both distributions have a median of 74.5. Which distribution has more variability?

The answer to this question depends on how we measure variability. Both distributions have the same range. The range is the distance spanned by the data. We calculate the range by subtracting the minimum value from the maximum value.

- Range = Maximum value – minimum value

For both of these data sets, the range is 55 (here is how we calculated the range: 95 – 40 = 55). If we use the range to measure variability, we say the distributions have the same amount of variability.

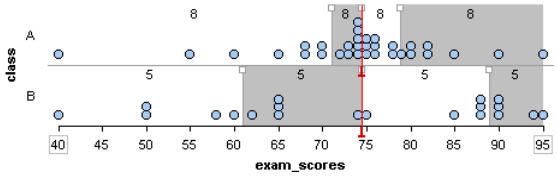
But the variability in the distributions differ when we look at how the data is distributed about the median. Set A has a large portion of its data close to the median. This is not true for Set B. From this viewpoint, Set A has less variability about the median.

Now we develop a way to measure the variability about the median. To do so, we use *quartiles*. Quartile marks divide the data set into four groups with equal counts.

## Example

## Quartiles and the Interquartile Range

We added dividers to show the quartile marks for the two sets of exam scores. Quartile marks divide the data set into four subgroups with the same number of individuals in each subgroup.




Notice: For a data set, there is an equal amount of data in each quartile.

- Class A has 32 scores, so each quartile contains eight scores (32 ÷ 4 = 8).
- Class B has 20 scores, so each quartile contains five scores (20 ÷ 4 = 5).

The quartiles together with the minimum and maximum scores give the five-number summary:

- Class A: Min: 40 Q1: 71 Q2: 74.5 Q3: 78.5 Max: 95
- Class B: Min: 40 Q1: 61 Q2: 74.5 Q3: 89 Max: 95

Notice: The second quartile mark (Q2) is the median.

Notice: Some quartiles exhibit more variability in the data even though each quartile contains the same amount of data.

- For example, 25% of the scores in Class A are between 40 and 71. There is a lot of variability in this first quartile (Q1). The eight scores in Q1 vary by 30 points.
- Compare this to the third quartile (Q3) for Class A: 25% of the scores in Class A are between 74.5 and 78.5. There is not much variability in Q3. The 8 scores in Q3 vary by only 4 points.

How are quartiles used to measure variability about the median? The *interquartile range (IQR)* is the distance between the first and third quartile marks. The IQR is a measurement of the variability about the median. More specifically, the IQR tells us the range of the middle half of the data.

Here is the IQR for these two distributions:

- Class A: IQR = Q3 – Q1 = 78.5 – 71 = 7.5
- Class B: IQR = Q3 – Q1 = 89 – 61 = 28

As we observed earlier, Class A has less variability about its median. Its IQR is much smaller. The middle 50% of exam scores for Class A vary by only 7.5 points. The middle 50% of exam scores for Class B vary by 28 points.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3445

### Using the IQR to Identify Outliers

We consider a point an outlier when it is substantially above Q3 or substantially below Q1. To make this statement more precise, we mark off a distance of 1.5 IQRs above Q3 and below Q1. Any point outside of this range is considered an outlier.

We can write this idea as a formula:

A value is an outlier when

- the value is greater than Q3 + 1.5 * IQR or
- the value is less than Q1 – 1.5 * IQR

To make more sense of this rule, let's look at a visual example.

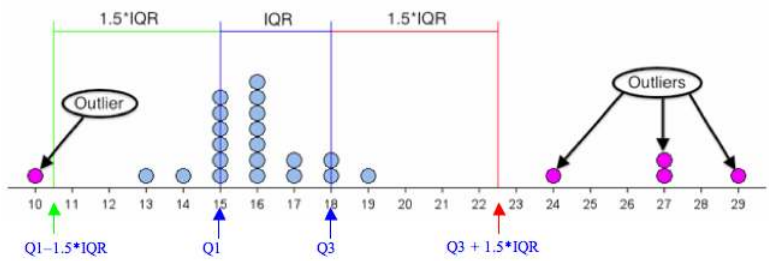
### Example

## Using IQR to Identify Outliers

For the data set in the dotplot, Q1 = 15 and Q3 = 18, so the IQR = 18 – 15 = 3.

- Q1 – 1.5 * IQR = 15 – 1.5 * 3 = 15 – 4.5 = 10.5
  - This cutoff is shown in green on the dotplot.
  - The data point at 10 is considered an outlier because it is below 10.5.
- Q3 + 1.5 * IQR = 18 + 1.5 * 3 = 18 + 4.5 = 22.5
  - This cutoff is shown in red on the dotplot.
  - The data points at 24, 27, and 29 are considered outliers because they are above 22.5.

The points in purple are outliers by the IQR definition.




## Learn By Doing

https://assessments.lumenlearning.co...sessments/3446

## Let's Summarize

- The range measures the variability of a distribution by looking at the interval covered by *all* the data. The IQR measures the variability of a distribution by giving us the interval covered by the *middle* 50% of the data.
- The five-number summary of a distribution consists of the minimum, quartile 1, median, quartile 3, and maximum.
- The IQR is the measure of spread we should use when using the median to measure center.
- When using the median and IQR to measure center and spread, a data point is considered an outlier if it satisfies one of the following conditions.
  - The data value is more than 1.5 IQRs greater than Q3 (i.e., the value is greater than Q3 + 1.5 * IQR)
  - The data value is more than 1.5 IQRs less than Q1 (i.e., the value is less than Q1 − 1.5 * IQR)

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.16: Interquartile Range and Boxplots (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.17: Interquartile Range and Boxplots (2 of 3)

## Learning Objectives

- Use a five-number summary and a boxplot to describe a distribution.

## Introduction

On the previous page, we learned about the five-number summary. At this point, you should know the following:

- The five-number summary uses quartiles to identify the center and spread of a distribution.
- The median (which is Q2) is a measure of center. We also view the median as a typical value that represents the distribution.
- The values between Q1 and Q3 give a typical range of values.
- The IQR is a way to measure the variability about the median.

Now we use the five-number summary to make a new type of graph, the **boxplot**. Boxplots are commonly used to summarize a distribution of a quantitative variable.

## Example

## Boxplots for Exam Scores

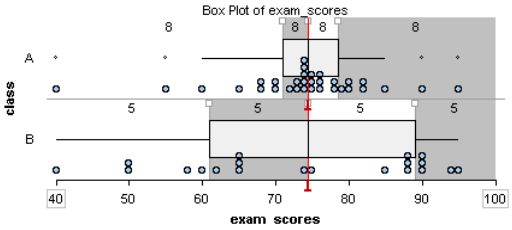
Here are the two sets of exam scores from the previous example. Recall that we divided the data into quartiles. In a data set, each quartile contains the same number of scores. In other words, each quartile contains 25% of the data.

Here is the five-number summary for these two distributions:

- Class A: Min: 40 Q1: 71 Q2: 74.5 Q3: 78.5 Max: 95
- Class B: Min: 40 Q1: 61 Q2: 74.5 Q3: 89 Max: 95

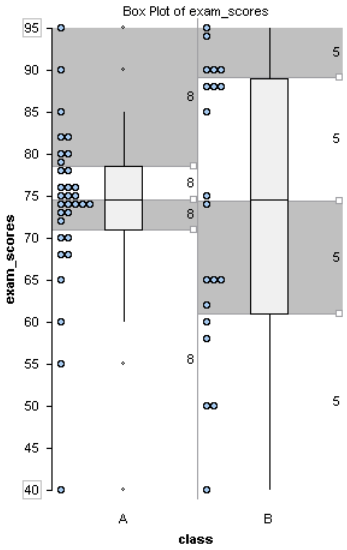
To create the boxplot for each distribution,

- Draw a box from Q1 to Q3.
- Draw a vertical line in the box at the median.
- Extend a tail from Q1 to the smallest value that is not an outlier and from Q3 to the largest value that is not an outlier.
- Indicate outliers with asterisks (*).




Notice: A long box in the boxplot indicates a large IQR, so the middle half of the data has a lot of variability. A short box in the boxplot indicates a small IQR. In this case, the middle half of the data has little variability.

Frequently, side-by-side boxplots are drawn vertically. Here we drew vertical dotplots with their boxplots for the exam scores from the two classes.



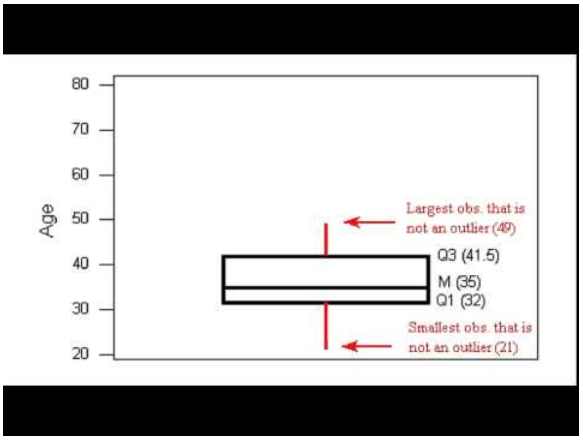

Note: Some statistical packages offer two options: a boxplot and a modified boxplot. We drew modified boxplots in this example. In a modified boxplot, outliers are marked with an asterisk (*). For a boxplot that is not modified, the tails extend to the minimum and maximum values. In this type of boxplot, we cannot see outliers.

## Making a Boxplot:

Now we walk through the steps for making a modified boxplot using the distribution of ages for winners of the Oscar Award for Best Actress from 1970 to 2001. The five-number summary for this distribution is

- Min: 21 Q1: 32 Median: 35 Q3: 41.5 Max: 80

Using the IQR definition of an outlier, there are three outliers: 61, 74, and 80.




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=82

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3447

https://assessments.lumenlearning.co...sessments/3448

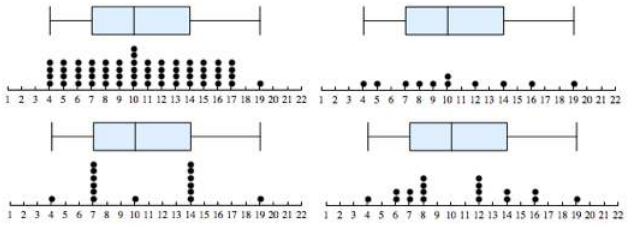
At this point, you should know how to

- Create a boxplot from a five-number summary.
- Use a boxplot to identify and interpret quartiles.
- Identify the median and the IQR of a distribution from a boxplot.

Now we want to focus on what a boxplot does *not* tell us. A boxplot does not give us information about the following:

- The number of data points in the data set.
- The number of data points within each quartile (though each quartile contains the same number of data points).
- The pattern of the data within each quartile.

Here are four data sets that illustrate these ideas.



How are these data sets *similar*? Notice that the four data sets have the same boxplot. This is because the five-number summary is the same for each data set. The data sets have identical minimum value, maximum value, and quartile marks, so we could say that these data sets have the same center and spread.

- Center: Each data set has a median of 10.
- Spread: In each data set, the middle half of the data varies from 7 to 14, so the IQR is 7. In each data set, the data varies from 4 to 19, so the overall range is 15.

How are these data sets *different*? The data sets do not have the same number of data points. Also, the shape of each distribution is different.

The goal of the next Learn By Doing activity is to develop a deeper understanding of how the interquartile range (IQR) measures variability about the median. Use the simulation below for the next activity. You have used a similar simulation before. Recall the instructions for adding or removing data points:

- To add a point, move the slider to the value you want, then click **Add**.
- To remove a point, move the slider to the value you want, then click **Minus**.
- To reset the simulation, click the button in the upper left corner that says **Reset**.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=82

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3842

https://assessments.lumenlearning.co...sessments/3843

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.17: Interquartile Range and Boxplots (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.18: Interquartile Range and Boxplots (3 of 3)

## Learning Objectives

- Use a five-number summary and a boxplot to describe a distribution.

## Comparing Distributions with Side-by-Side Boxplots

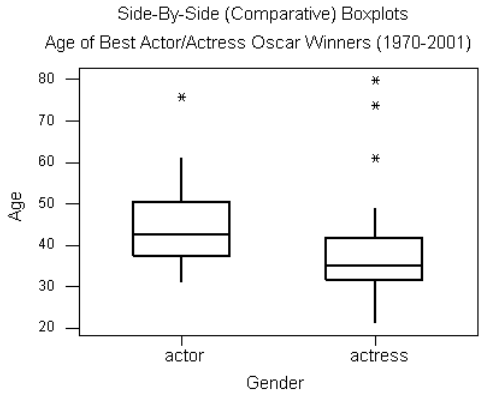
In the next two examples, we again use boxplots to compare two distributions. This time we focus on writing a description of the two distributions. We practiced writing descriptions in the earlier section, "Distributions for Quantitative Data," using dotplots and histograms. Now we use boxplots. As before, we describe shape, center, spread, and outliers. But now we use the five-number summary to make our descriptions more precise.

## Example

## Best Actor/Actress Oscar Winners

So far we have examined the age distributions of Oscar winners for males and females separately.

It will be interesting to *compare* the age distributions of actors and actresses who won best acting Oscars. To do that, we look at side-by-side boxplots of the age distributions by gender.




- Actors: Min = 31, Q1 = 37.75, M = 42.5, Q3 = 48.75, Max = 76
- Actresses: Min = 21, Q1 = 32, M = 35, Q3 = 41.5, Max = 80

Based on the graph and numerical measures, we can make the following comparison between the two distributions:

Note: A good summary compares the two distributions using shape, center, spread, and outliers. Let's begin with observations about these characteristics of the distributions.

**Shape**: The shape of a distribution can be hard to determine from the boxplot, but we can compare the variability in the upper half of the data (Max − Median) to the variability in the lower half of the data (Median − Min) to get a sense of shape. For the men, the distribution appears skewed to the right because the lower half of the data has less variability than the upper half. The lower half of the data has a range of 11.5 years (42.5 − 31), compared to the upper half of the data with a range of 33.5 years (76 − 42.5). The distribution for women also appears right-skewed. The lower half of the data has a range of 14 years (35 − 21), compared to a range of 45 years for the upper half of the data (80 − 35). In both cases, the shape suggests that the Oscar is awarded to younger actors and actresses.

**Center**: Actresses tend to win the Oscar at a younger age than do actors. The median age for females (35) is lower than for the males (42.5). Note also that the third quartile of the females' distribution (41.5) is lower than the median age for males. It tells us that only 25% of the actresses were 41.5 years old or older when they won the Oscar, compared to 50% of the males who were 42.5 years old or older.

**Spread**: Not only do actresses win at a younger age, but the Oscar is awarded more consistently to younger actresses, as we can see by comparing the interquartile ranges. There is less variability in the middle half of the actresses' ages (IQR = 9.5) than in the

actors' ages (IQR = 11). On the other hand, the actresses have more variability in their overall ages (range = 59) compared to the actors (range = 45).

**Outliers**: We see that we have outliers in both distributions. There is only one high outlier in the actors' distribution (76, Henry Fonda, On Golden Pond), compared with three high outliers in the actresses' distribution.

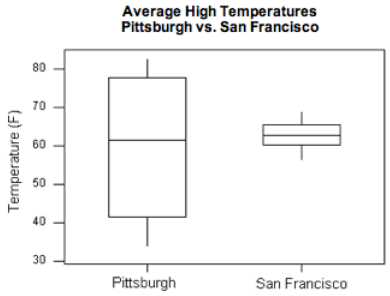
*Now let's pull these observations together into a paragraph. A good paragraph compares the two distributions and uses observations about the distributions to support a central thesis.*

In general, actresses win the Best Actress Oscar at a younger age than do actors. The median age for actresses is 35, compared to 42.5 for actors. Not only do actresses win at a younger age, the Oscar is awarded more consistently to younger actresses, as seen when we compare the interquartile ranges. There is less variability in the middle half of the actresses' ages (IQR = 9.5) than in the actors' ages (IQR = 11). Both distributions have older winners that are outliers. These older winners are unusual and skew the distribution of ages to the right.

### Example

## Temperature of Pittsburgh vs. San Francisco

To compare the average high temperatures of Pittsburgh to those of San Francisco, we look at the following side-by-side boxplots and supplement the graph with the descriptive statistics of each of the two distributions.




| Average High Temperatures | | |
|---|---|---|
| | Pittsburgh | San Francisco |
| Min | 33.7 | 56.3 |
| Q1 | 41.2 | 60.2 |
| Q2 (Median) | 61.4 | 62.7 |
| Q3 | 77.8 | 65.4 |
| Max | 82.6 | 68.7 |

When looking at the graph, the similarities and differences between the two distributions are striking. Both distributions have roughly the same center (medians are 61.4 for Pittsburgh and 62.7 for San Francisco). However, the temperatures in Pittsburgh have a much larger variability than the temperatures in San Francisco (Range: 49 vs. 12; IQR: 36.5 vs. 5).

The practical interpretation of the results we obtained is that the weather in San Francisco is much more consistent than the weather in Pittsburgh, which varies a lot during the year. Also, because the temperatures in San Francisco vary so little during the year, knowing that the median temperature is around 63 is actually very informative. On the other hand, knowing that the median temperature in Pittsburgh is around 61 is practically useless, since temperatures vary so much during the year and can get much warmer or much colder than in San Francisco.

Note that this example provides more intuition about variability by interpreting small variability as consistency and large variability as lack of consistency. Also, through this example, we learned that the center of the distribution is more meaningful as a typical value for the distribution when there is little variability (or, as statisticians say, little "noise") around it. When there is large variability, the center loses its practical meaning as a typical value.

### Let's Summarize

- The range measures the variability of a distribution by looking at the interval covered by *all* the data. The IQR measures the variability of a distribution by giving us the interval covered by the *middle* 50% of the data.
- The five-number summary of a distribution consists of the minimum, quartile 1, median, quartile 3, and maximum.
- The IQR is the measure of spread we should use when using the median to measure center.

- When using the median and IQR to measure center and spread, a data point is considered an outlier if it satisfies one of the following conditions.
  - More than 1.5 IQRs greater than Q3 (i.e., the value is greater than Q3 + 1.5 * IQR).
  - More than 1.5 IQRs less than Q1 (i.e., the value is less than Q1 – 1.5 * IQR).
- The boxplot is a graphical representation of a data set. It displays the five-number summary and highlights any points that are considered outliers (using the 1.5 * IQR rule described in the previous bullet).
- Side-by-side boxplots are commonly used to compare two data sets.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.18: Interquartile Range and Boxplots (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.19: Introduction- Describing a Distribution

## What you'll learn to do: Describe a distribution using mean and standard deviation

### LEARNING OBJECTIVES

- Use mean and standard deviation to describe a distribution.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.19: Introduction- Describing a Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.20: Standard Deviation (1 of 4)

## Learning Objectives

- Use mean and standard deviation to describe a distribution.

## Introduction

In the section “Distributions for Quantitative Data,” we discussed the spread of a distribution in terms of a *typical range* of values. In “Quantifying Variability Relative to the Median,” we made this idea more precise with the interquartile range, IQR. The IQR gives us a measure of spread about the median. We defined a typical range of values about the median as the values between the first and third quartiles.

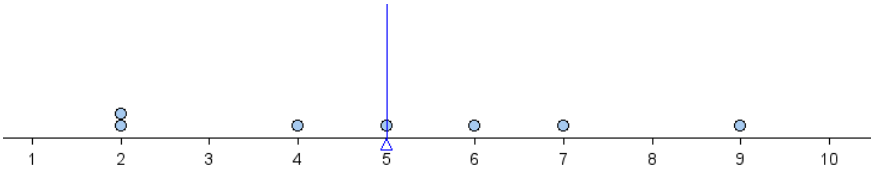
Now we want to develop a numerical measure of spread that we can use with the mean. In constructing a measure of spread about the mean, we want to compute how far a “typical” number is away from the mean.

## Measuring Spread about the Mean

Let’s consider the sample data set 2, 2, 4, 5, 6, 7, 9. The mean of this data set is

$$\stackrel{\frown}{x}=\frac{2+2+4+5+6+7+9}{7}=\frac{35}{7}=5$$

Here is a dotplot of this data set with the mean marked by the vertical blue line.

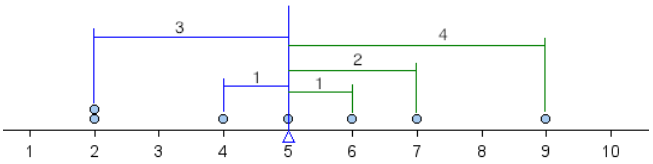



We can see that some data is close to the mean and some data is further from the mean.

Since we want to see how the data points deviate from the mean, we determine how far each point is from the mean. We compute the difference between each of these values and the mean. These differences are called the *deviations from the mean* for each point.

| |
|---|
| $2-5=-3$ |
| $2-5=-3$ |
| $4-5=-1$ |
| $5-5=0$ |
| $6-5=1$ |
| $7-5=2$ |
| $9-5=4$ |

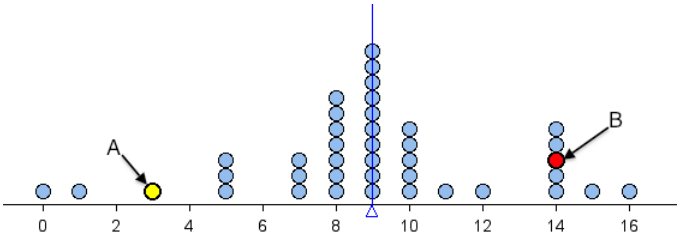
When visualized on a dotplot, these differences are viewed as distances between each point and the mean. A negative difference indicates that the data point is to the *left* of the mean (shown in blue on the graph below). A positive difference indicates that the data point is to the *right* of the mean (shown in green on the graph below).




Our goal is to develop a single measurement that summarizes a typical distance from the mean. Before we continue, let’s practice determining the distance of a single data point from the mean.

## Learn By Doing

The two questions below refer to the following dotplot. The mean is 9 and it is marked by the vertical blue line.




https://assessments.lumenlearning.co...sessments/3449

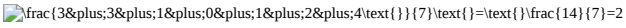
https://assessments.lumenlearning.co...sessments/3450

Since we want to determine how far a typical number is away from the mean, we might try to average these numbers. However, if we add them all up, we will get 0 (try it). Getting 0 with this procedure (finding differences from mean and adding them all together) is no accident – it *always* produces 0. We have to overcome this problem.

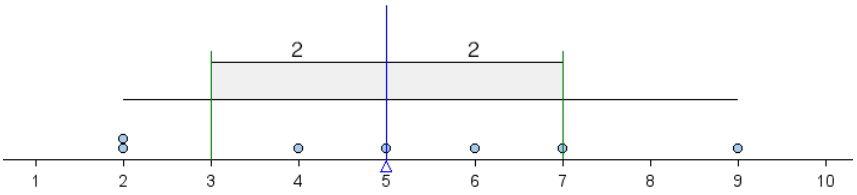
Recall that we are trying to find the typical *distance* between data points and the mean. It therefore makes sense to take the absolute value of each of these differences.

| |
|---|
| $\lvert 2-5 \rvert = \lvert -3 \rvert = 3$ |
| $\lvert 2-5 \rvert = \lvert -3 \rvert = 3$ |
| $\lvert 4-5 \rvert = \lvert -1 \rvert = 1$ |
| $\lvert 5-5 \rvert = \lvert 0 \rvert = 0$ |
| $\lvert 6-5 \rvert = \lvert 1 \rvert = 1$ |
| $\lvert 7-5 \rvert = \lvert 2 \rvert = 2$ |
| $\lvert 9-5 \rvert = \lvert 4 \rvert = 4$ |

Now we can compute the average of these deviations. There are seven data points, so we add these seven distances and divide by 7. The result is a measure of spread about the mean called the **average deviation from the mean (ADM)**.




We can indicate this average deviation on a dotplot with a graphic similar to a boxplot as follows.




The shaded box in the middle is centered at the mean. It extends left and right a distance of 1 average deviation from the mean. Because the average deviation about the mean for this data set is 2, the box starts at 3 (because 5 − 2 = 3) and ends at 7 (because 5 + 2 = 7). In this way, we can use the ADM to define a typical range of values about the mean. Notice that this typical range of values (within 1 ADM of the mean) contains more than half of the values in the data set.

The goal of the next Learn By Doing exercise is to improve our intuition of what the ADM measures. We use the following simulation to investigate how the ADM responds to changes in a data set.

Instructions for adding or removing data points:

- To add a point, move the slider to the value you want, then click the + sign.

- To remove a point, move the slider to the value you want, then click the – sign.
- To reset the simulation to a blank screen, click the button in the upper left corner that says **Reset**.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=88

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3451

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3452

Before we continue, let's summarize our main points:

- The ADM (average distance from the mean) is a measurement of spread about the mean. More precisely, ADM measures the average distance of the data from the mean.
- We can use the ADM to define a typical range of values about the mean. We mark the mean, then we mark 1 ADM below the mean and 1 ADM above the mean. This interval is centered at the mean and captures typical values about the mean.

Using these two ideas, we can estimate the ADM by looking at a graph of the distribution of data. We practice this important skill in the next Learn By Doing.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3453

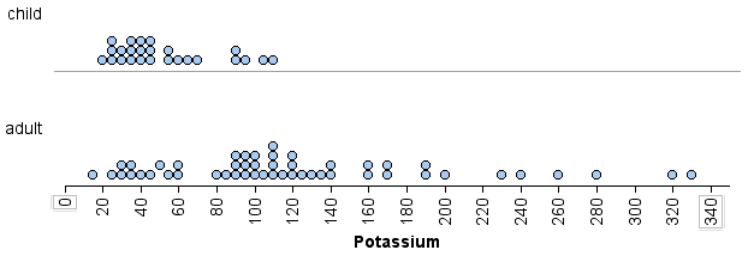
https://assessments.lumenlearning.co...sessments/3454

In the next example, we compare the ADM as a measure of spread to the other ways we have measured spread.

### Example

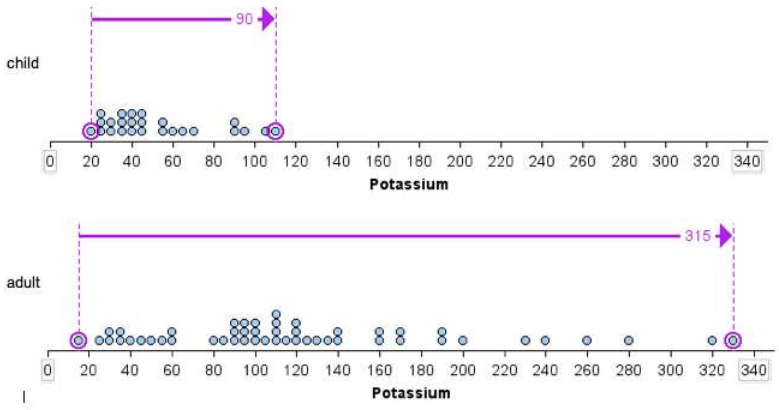
## Measuring Variability in Different Ways

The following dotplots show the potassium content in 76 cereals. Compare children's cereals to adult cereals. Which type of cereal has more variability in potassium content?

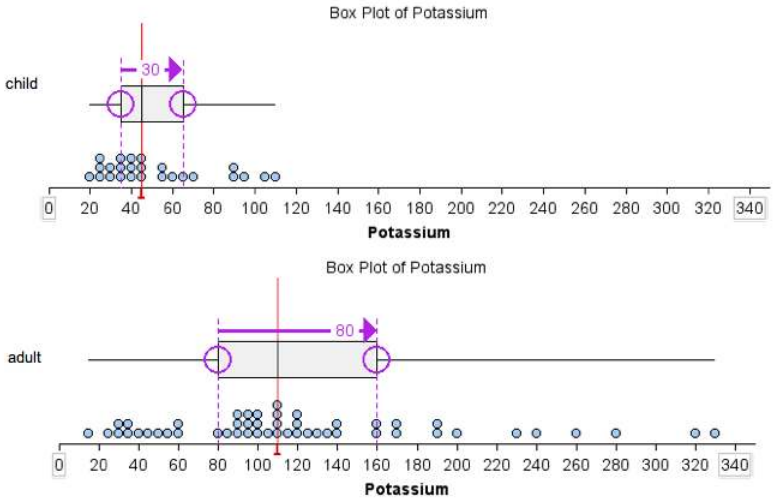



We can visually see that there is more variability in the potassium content of the adult cereals than in the children's cereals. We can measure this spread in three ways:

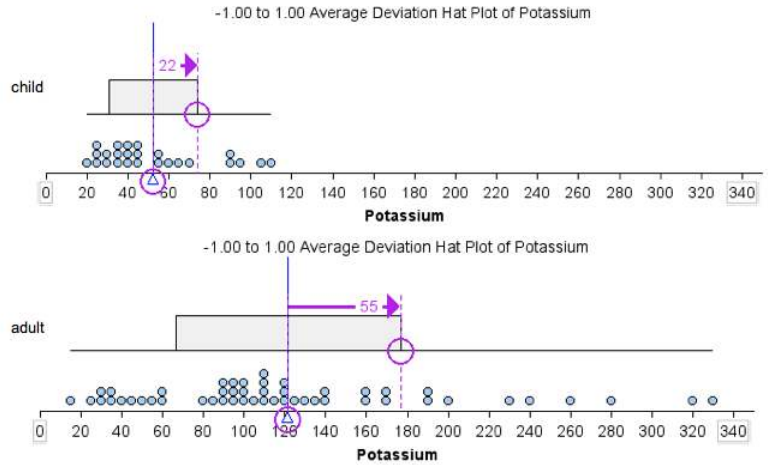
- Using overall range: The range of potassium content is larger for the adult cereals than for the children's cereals. The children's cereal set has a range of 90 (because 110 − 20 = 90), whereas the adult cereal set has a range of 315 (because 330 − 15 = 315).




- Using IQR: The IQR of the adult cereal set is larger than the IQR of the children's cereal set. The adult cereal set has an IQR of 80, since for that set Q1 = 80 and Q3 = 160. The children's cereal set has an IQR of 30, since for that set Q1 = 35 and Q3 = 65. Notice here we use the median as a measure of center. The median is marked with a red line. IQR measures spread about the median.




- Using ADM: The children's cereal data set has an ADM of 22. The adult cereal data set has an ADM of 55. Notice here we use the mean as a measure of center. The mean is marked with a blue line. ADM measures spread about the mean.

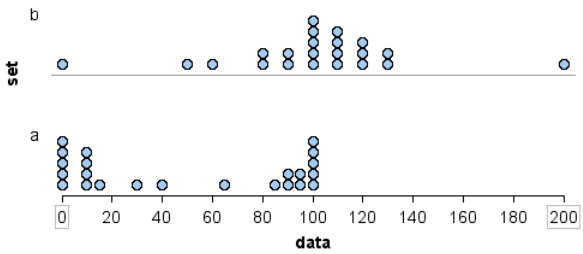


Based on the preceding example, we might expect the data set with the larger range to also have the larger ADM. This is not true, as we illustrate in the next example.

### Example

## Comparing Range and ADM

Which data set has more variability? Our answer to this question depends on how we measure variability.

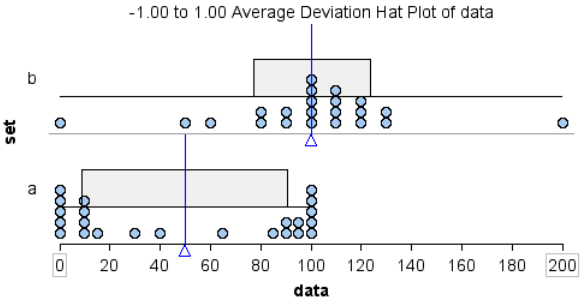



We can see that the overall range is larger for data set b.

- Data set a: range = 100 − 0 = 100
- Data set b: range = 200 − 0 = 200

If we use overall range to measure spread, we will say that data set b has more variability.

Does our answer change if we use ADM to measure spread? Yes!




- Data set a: ADM = 41
- Data set b: ADM = 23

Most of the data in data set a is located away from the mean, so the ADM is large: 41. Compare this to data set b. Most of the data in data set b is located close to the mean, so the ADM is small: 23.

If we use ADM as a measure of spread, we will say that data set a has more variability.

The ADM is a reasonable measure of spread about the mean, but there is another measure that is used much more often: the standard deviation (SD). The standard deviation behaves very much like the average deviation. So all of the work we have done on this page is useful in understanding standard deviation. We discuss standard deviation next.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.20: Standard Deviation (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.21: Standard Deviation (2 of 4)

## Learning Objectives

- Use mean and standard deviation to describe a distribution.

## A More Common Measure of Spread about the Mean: The Standard Deviation

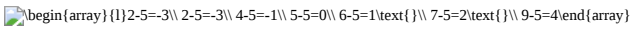
The **standard deviation (SD)** is a measurement of spread about the mean that is similar to the average deviation. We think of standard deviation as roughly the average distance of data from the mean. In other words, the standard deviation is approximately equal to the average deviation. We develop the formula for standard deviation in the following example.

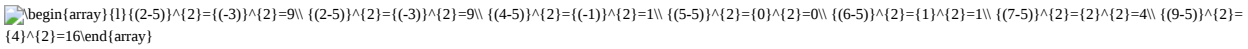
## Example

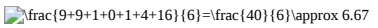
## Calculating the Standard Deviation

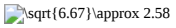
Let's consider the same data set we used on the previous page: 2, 2, 4, 5, 6, 7, 9. We already know that the mean is 5. We compute the standard deviation similarly to the way we compute the average deviation. We begin by computing the deviation of each point from the mean, but instead of taking the absolute value of the differences, we square them. Here are the steps:

1. We start by finding the differences between each value and the mean (just like before):

   
   
2. We square each of the differences:

   
   
3. As before, we find the average of these squared differences. We add the squared differences and divide by $n - 1$ (the count minus 1). Note that we divide by $n - 1$ instead of $n$. (The reason is subtle. We do not discuss it in this course.)

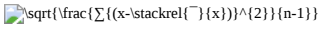
   
   
4. To scale back the value to account for the squaring we did in step 2, we take the square root of the value we found in step 3:

Notice that the standard deviation is a little bit larger than the average deviation (which was 2). We can get a good approximation of the standard deviation by estimating the average distance from the mean. The shaded box on the following dotplot indicates 1 SD to the right and left of the mean.

## Comment

The formula for the standard deviation of a data set can be described by the following expression. However, we will always use technology to perform the actual computation of the standard deviation.




The symbols in the expression are defined as follows:

- $n$ is the number of values in the data set (the count).
- Recall that ∑ means to add up (compute the sum).
- is the mean of the data set.
- The individual values are denoted by $x$.

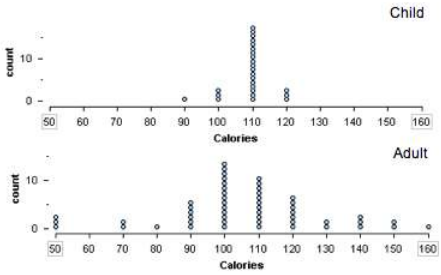
Note: In the formula you can see

- the deviations from the mean .
- the squaring of these deviations.
- the averaging of the squared deviations: add them up (∑) and divide by ($n - 1$).

Before we learn to use technology to compute the standard deviation, we practice estimating it. We can estimate standard deviation in the same ways we estimated ADM. Think of standard deviation as roughly equal to ADM, so standard deviation is roughly the average distance of data from the mean.

## Learn By Doing

Let’s consider the same collection of cereals we worked with previously, except this time we’ll look at the calorie content.




https://assessments.lumenlearning.co...sessments/3455

https://assessments.lumenlearning.co...sessments/3456

https://assessments.lumenlearning.co...sessments/3457

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.21: Standard Deviation (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.22: Standard Deviation (3 of 4)

## Learning Objectives

- Next we summarize our observations with a focus on comparing the two groups:

  From this analysis, we can see that the group of students in the fifth and seventh grades are carrying more weight on average in their backpacks. The mean weight for this group is 14.2 pounds compared to 5.8 pounds for the group of first and third graders. There is also more variability in backpack weights in the fifth- seventh-grade group. The standard deviation for this group is 7.2 pounds, compared to 2.1 pounds for the younger students.

  One plausible explanation is that as children get older, they are assigned more homework, so they carry more in their backpacks. But at this age, we may also see more students making independent decisions about how much homework they will do, so some students will carry more books home and others will carry fewer.

## Learn By Doing

Consider the following two quantitative data sets:

- https://assessments.lumenlearning.co...sessments/3458

Can two data sets have the same mean but different standard deviations? Can two data sets have different means but the same standard deviation? Use the simulation to investigate these questions in the next two activities.

Instructions for adding or removing data points:

- Click here to open this simulation in its own window.

  An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=92

### Activity 1

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3463

### Activity 2

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3464

Remark:

The examples we constructed in the preceding activity should make it clear that the mean and standard deviation measure independent characteristics of a data set. The mean is a measure of center, and the standard deviation is a measure of spread. The size of the mean does not give us information about the size of the standard deviation. Similarly, the size of the standard deviation does not give us information about the size of the mean.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.22: Standard Deviation (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.23: Standard Deviation (4 of 4)

## Learning Objectives

- Use mean and standard deviation to describe a distribution.

## Deciding Which Measurements to Use

We now have a choice between two measurements of center and spread. We can use the median with the interquartile range, or we can use the mean with the standard deviation. How do we decide which measurements to use?

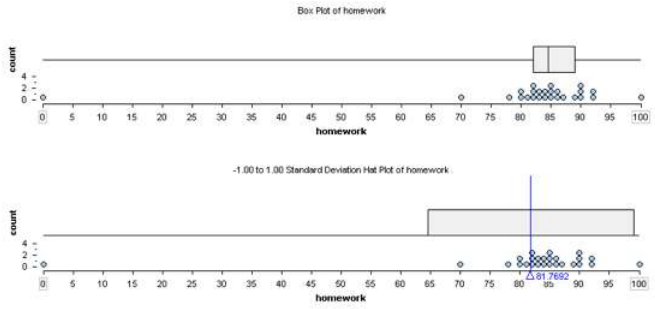
Our next examples show that the shape of the distribution and the presence of outliers helps us answer this question.

## Example

## Homework Scores with an Outlier

Here are two summaries of the same set of homework scores earned by a student: a boxplot and an SD hatplot. Notice that the distribution of scores has an outlier. This student has mostly high homework scores with one score of 0. Here are some observations about the homework data.

- Five-number summary: low: 0 Q1: 82 Q2: 84.5 Q3: 89 high: 100
- Median is 84.5 and IQR is 7
- Mean = 81.8, SD = 17.6




The typical range of scores based on the first and third quartiles is 82 to 89.

The typical range of scores based on Mean ± SD is 64.2 to 99.4 (Here's how we calculated this: 81.8 – 17.6 = 64.2, 81.8 + 17.6 = 99.4.)

Which is the better summary of the student's performance on homework?

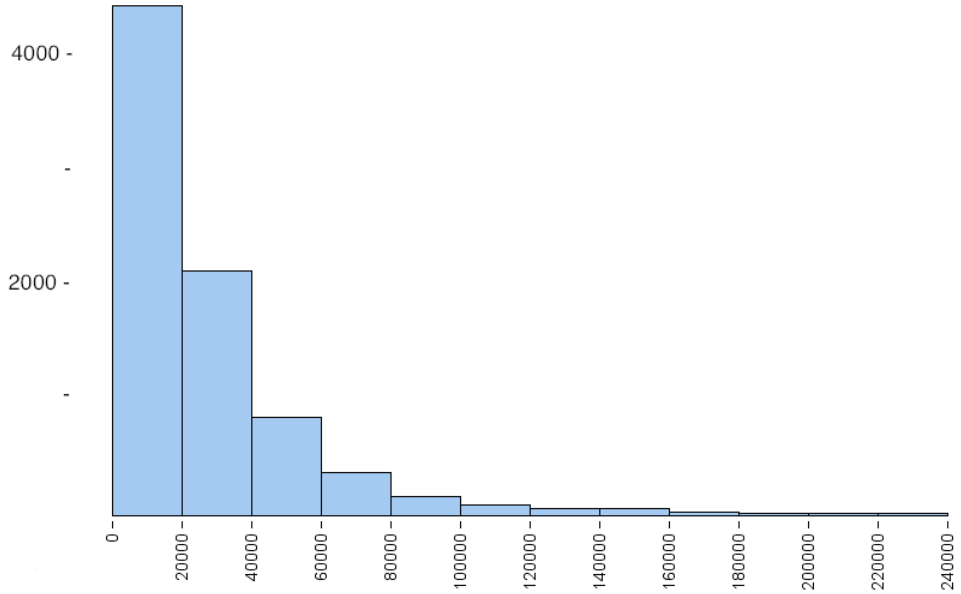
The typical range based on the mean and standard deviation is not a good summary of this student's homework scores. Here we see that the outlier decreases the mean so that the mean is too low to be representative of this student's typical performance. We also see that the outlier increases the standard deviation, which gives the impression of a wide variability in scores. This makes sense because the standard deviation measures the average deviation of the data from the mean. So a point that has a large deviation from the mean will increase the average of the deviations. In this example, a single score is responsible for giving the impression that the student's typical homework scores are lower than they really are.

The typical range based on the first and third quartiles gives a better summary of this student's performance on homework because the outlier does not affect the quartile marks.

## Example

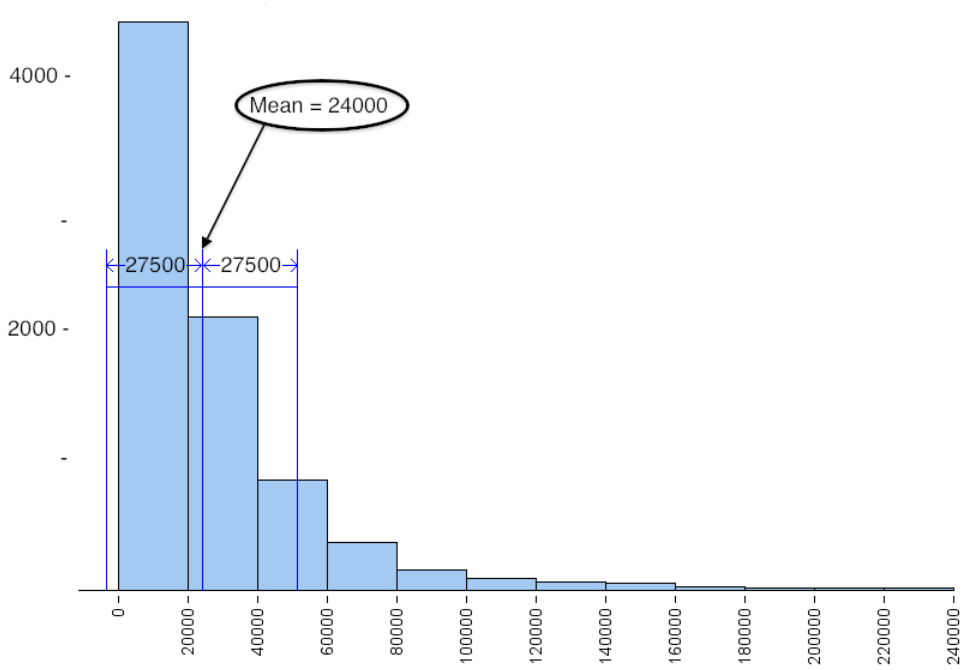
## Skewed Incomes

In this example, we look at how skewness in a data set affects the standard deviation. The following histogram shows the personal income of a large sample of individuals drawn from U.S. census data in the year 2000. Notice that it is strongly skewed to the right. This type of skewness is often present in data sets of variables such as income.




Following are some summary statistics for this data:

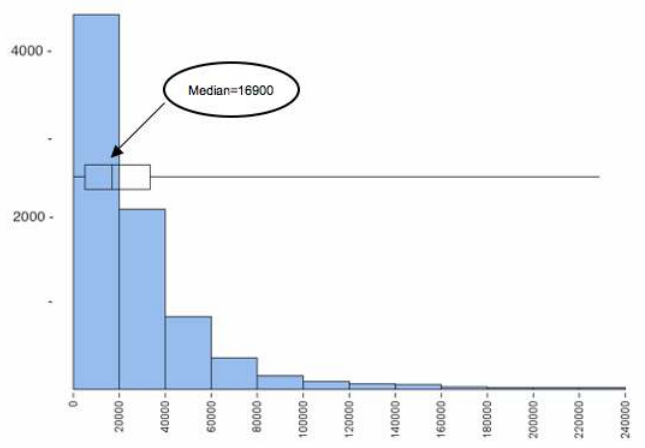
- Mean = $24,000, SD = $27,500
- Median = $16,900, IQR = $28,000




The typical range based on the mean and standard deviation is not a good summary of the distribution of incomes. The small number of people with higher incomes increases the mean. The mean is too high to represent the large number of people making less than $20,000 a year. The small number of people with higher incomes also increase the standard deviation, so a small number of high incomes gives the misleading impression that typical incomes in the sample are higher than they really are.

Notice also that Mean ± SD gives an awkward range of typical values. The left endpoint is at −3,500 (Mean − SD = 24,000 − 27,500 = −3,500), but there are no negative values in this data set. This is another reason why it is better to use the IQR when measuring the spread of a skewed data set.

Let's take a look at the same histogram, except this time we overlay a boxplot.




We see that the median represents the typical income of people in this sample better than the mean. The small number of people with higher incomes does not impact the median or the other quartile marks, so the first and third quartile marks give a range of incomes that more accurately represent typical incomes in the sample. Notice also that this range is always within the overall range of the data, so we will never have the problem that we encountered earlier with the standard deviation.

In a skewed distribution, the upper half and the lower half of the data have a different amount of spread, so no single number such as the standard deviation could describe the spread very well. We get a better understanding of how the values are distributed if we use the quartiles and the two extreme values in the five-number summary.

These examples illustrate some general guidelines for choosing numerical summaries:

- Use the mean and the standard deviation as measures of center and spread *only* for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean and standard deviation are not a good choice.
- Use the five-number summary (which gives the median, IQR, and range) for all other cases.

Both of these examples also highlight another important principle: *Always plot the data*.

We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measures of center and spread best describe the data.

### Let's Summarize

- The average deviation from the mean (ADM) is a measurement of spread about the mean. More precisely, ADM measures the average distance of the data from the mean. In practice, ADM is not commonly used, but it helps us understand the standard deviation (SD).
- The standard deviation is a measure of spread. We use it as a measure of spread when we use the mean as a measure of center.
- The standard deviation is approximately the average distance of the data from the mean, so it is approximately equal to ADM.
- We can use the standard deviation to define a typical range of values about the mean. We mark the mean, then we mark 1 SD below the mean and 1 SD above the mean. This interval is centered at the mean and defines typical values about the mean. We often write this interval as Mean ± SD.
- We use technology to calculate the standard deviation.
- Like the mean, the standard deviation is strongly affected by outliers and skew in the data.

When choosing numerical summaries,

- Use the mean and the standard deviation as measures of center and spread *only* for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean and standard deviation are not a good choice.
- Use the five-number summary (which gives the median, IQR, and range) for all other cases.
- *Always plot the data.* We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measures of center and spread best describe the data.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.23: Standard Deviation (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.24: Putting It Together- Summarizing Data Graphically and Numerically

## Let's Summarize

In *Summarizing Data Graphically and Numerically*, we focused on describing the *distribution of a quantitative variable*.

- To analyze the distribution of a quantitative variable, we describe the *overall pattern of the data* (shape, center, spread) and any *deviations from the pattern* (outliers). We use three types of graphs to analyze the distribution of a quantitative variable: dotplots, histograms, and boxplots.
- We described the *shape* of a distribution as left-skewed, right-skewed, symmetric with a central peak (bell-shaped), or uniform. Not all distributions have a simple shape that fits into one of these categories.
- The *center* of a distribution is a typical value that represents the group. We have two different measurements for determining the center of a distribution: mean and median.
  - The *mean* is the average. We calculate the mean by adding the data values and dividing by the number of individual data points. The *mean* is the *fair share* measure. The mean is also called the *balancing point* of a distribution. If we measure the distance between each data point and the mean, the distances are balanced on each side of the mean.
  - The *median* is the physical center of the data when we make an ordered list. It has the same number of values above it as below it.
  - **General Guidelines for Choosing a Measure of Center**
    - *Always plot the data.* We need to use a graph to determine the shape of the distribution. By looking at the shape, we can determine which measure of center best describes the data.
    - Use the mean as a measure of center *only* for distributions that are reasonably symmetric with a central peak. When outliers are present, the mean is not a good choice.
    - Use the median as a measure of center for all other cases.
- The *spread* of a distribution is a description of how the data varies. We studied three ways to measure spread: *range* (max – min), the *interquartile range* (Q3 – Q1), and the *standard deviation*. When we use the median, Q1 to Q3 gives a typical range of values associated with the middle 50% of the data. When we use the mean, Mean ± SD gives a typical range of values.
  - The interquartile range (IQR) measures the variability in the middle half of the data.
  - Standard deviation measures roughly the average distance of data from the mean.
- *Outliers* are data points that fall outside the overall pattern of the distribution. When using the median and IQR to measure center and spread, we use the 1.5 * IQR interval to identify outliers. Specifically, points outside the interval Q1 – 1.5 * IQR to Q3 + 1.5 * IQR are labeled as outliers.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.24: Putting It Together- Summarizing Data Graphically and Numerically is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.25: StatTutor- Drinking Habits of College Students

You are now ready to practice what you learned in this module by doing a StatTutor exercise. We design StatTutor exercises to help you apply what you have learned to a real life data analysis question.

**Instructions**: One of the first few screens in StatTutor will have a link to download the dataset for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click **Save**, which will allow you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you're using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you're using the TI Calculator.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=98

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.25: StatTutor- Drinking Habits of College Students is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.26: Assignment- Histogram

We will use the Best Actor Oscar winners (1970–2001) to learn how to create a histogram using a statistics package, and practice what we've learned about describing the histogram.

Click here to see the entire dataset.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 1:

Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.26: Assignment- Histogram is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.27: Assignment- Five-Number Summary

In this activity, we will use the Best Actor Oscar winners (1970-2001) to:

- Learn how to use a statistics package to produce the numerical measures, or “descriptive statistics” of a distribution.
- Get some information about the distribution from its five-number summary.

Click here to see the entire dataset.

Choose your statistical package and follow the instructions to compute numerical measures. Note that “n” represents the sample size, which is the number of individuals in the data set.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 1:

Getting information from the output:

a. How many observations are in this data set?

b. What is the mean age of the actors who won the Oscar?

c. What is the five-number summary of the distribution?

## Question 2:

Get information from the five-number summary:

a. Half of the actors won the Oscar before what age?

b. What is the range covered by all the actors’ ages?

c. What is the range covered by the middle 50% of the ages?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.27: Assignment- Five-Number Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.28: Assignment- Boxplot

The objectives of this activity are:

- To teach you how to use

  to produce side-by-side boxplots and the relevant descriptive statistics,
- To let you practice comparing and contrasting distributions, and
- To help you gain more intuition about variability through the interpretation of your results in context.

The percentage of each entering Freshman class that graduated on time was recorded for each of six colleges at a major university over a period of several years. (Source: This data is distributed with the software package, Data Desk. (1993). Ithaca, NY: Data Description, Inc., and appears in http://lib.stat.cmu.edu/DASL/)

In order to compare the graduation rates among the different colleges, we will create side-by-side boxplots (graduation rate by college), and supplement the graph with numerical measures. Follow the instructions, and then answer the questions based on the output you got.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

Answer the following questions:

## Question 1:

Compare and contrast the distributions of the graduation rates at the different colleges. Be sure to address center, spread and outliers.

## Question 2:

If you had to choose one college among the six colleges based on this data, which college would it be? Explain your reasoning.

## Question 3:

If you were debating between colleges B and F only, which one would you choose based on this data? Explain your reasoning.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 2.28: Assignment- Boxplot is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 2.29: Assignment- Standard Deviation

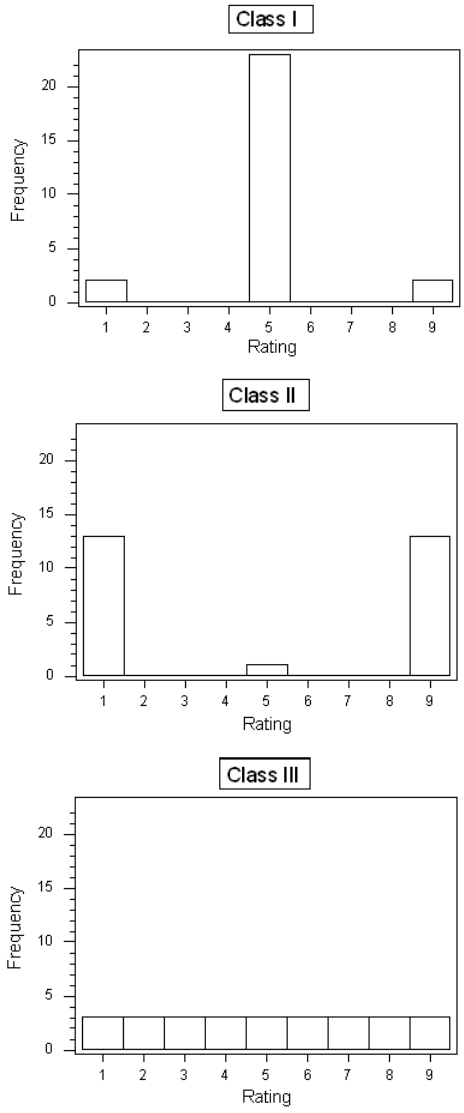
The concept of standard deviation is less intuitive as a measure of spread than the range or the IQR. The following activity is designed to help you develop a better intuition for the standard deviation.

## Background

At the end of a statistics course, students in three different classes rated their instructor on a number scale of 1 to 9 (1 being "very poor," and 9 being "best instructor I've ever had"). The following table provides three hypothetical rating data:

| Rating | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Class I | 2 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 2 |
| Class II | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 13 |
| Class III | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

And here are the histograms of the data:




## Question 1:

Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a "typical" or "average" distance between the data points and their mean. Judging from

the table and the histograms, which class would have the largest standard deviation, and which one would have the smallest standard deviation? Explain your reasoning.

Now check your intuition by finding the actual standard deviations of the three rating distributions.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 2:

What are the standard deviations of the three rating distributions? Was your intuition correct?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 2.29: Assignment- Standard Deviation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 3: Examining Relationships- Quantitative Data


3.1: Why It Matters- Examining Relationships- Quantitative Data
3.2: Introduction- Scatterplots
3.3: Scatterplots (1 of 5)
3.4: Scatterplots (2 of 5)
3.5: Scatterplots (3 of 5)
3.6: Scatterplots (4 of 5)
3.7: Scatterplots (5 of 5)
3.8: Introduction- Linear Relationships
3.9: Linear Relationships (1 of 4)
3.10: Linear Relationships (2 of 4)
3.11: Linear Relationships (3 of 4)
3.12: Linear Relationships (4 of 4)
3.13: Introduction- Association vs Causation
3.14: Causation and Lurking Variables (1 of 2)
3.15: Causation and Lurking Variables (2 of 2)
3.16: Introduction- Linear Regression
3.17: Linear Regression (1 of 4)
3.18: Linear Regression (2 of 4)
3.19: Linear Regression (3 of 4)
3.20: Linear Regression (4 of 4)
3.21: Introduction- Assessing the Fit of a Line
3.22: Assessing the Fit of a Line (1 of 4)
3.23: Assessing the Fit of a Line (2 of 4)
3.24: Assessing the Fit of a Line (3 of 4)
3.25: Assessing the Fit of a Line (4 of 4)
3.26: Putting It Together- Examining Relationships- Quantitative Data
3.27: StatTutor- Academic Performance
3.28: Assignment- Scatterplot
3.29: Assignment- Linear Relationships
3.30: Assignment- Linear Regression



This page titled 3: Examining Relationships- Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.1: Why It Matters- Examining Relationships- Quantitative Data

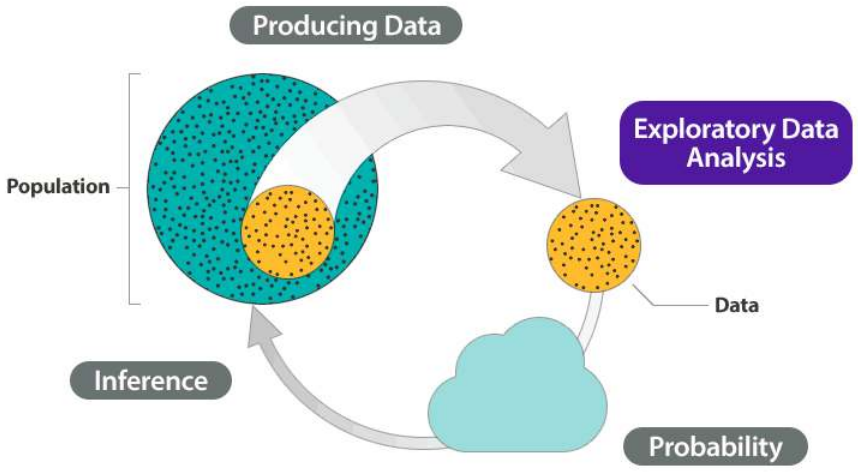
Before we begin *Examining Relationships: Quantitative Data*, let's see how the new ideas in this module relate to what we learned in the previous modules, *Types of Statistical Studies and Producing Data* and *Summarizing Data Graphically and Numerically*.

Recall the Big Picture:

We begin a statistical investigation with a research question. The investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data. ← **Types of Statistical Studies and Producing Data**
- Explore the Data: Analyze and summarize the data. ← **Summarizing Data Graphically and Numerically, Examining Relationships: Quantitative Data**
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population.

*Types of Statistical Studies and Producing Data* focused on methods for collecting reliable data. *Summarizing Data Graphically and Numerically* focused on summarizing and analyzing data for a quantitative variable. In this module, we focus on summarizing and analyzing the relationship between two quantitative variables. In the Big Picture of Statistics, the material in *Examining Relationships: Quantitative Data* is still part of exploratory data analysis.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.1: Why It Matters- Examining Relationships- Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.2: Introduction- Scatterplots

## What you'll learn to do: Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

### LEARNING OBJECTIVES

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.2: Introduction- Scatterplots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.3: Scatterplots (1 of 5)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

## Example

## Highway Signs

A research firm conducts a study to explore the relationship between a driver's age and the driver's ability to read highway signs. The subjects are a random sample of 30 drivers between the ages of 18 and 82. *(Source: Jessica M. Utts and Robert F. Heckard, Mind on Statistics [Brooks/Cole, 2002]. Original source: Data collected by The Last Resource, Inc., Bellfonte, PA.)*

Because the purpose of this study is to explore the effect of age on the driver's ability to read highway signs,

- the *explanatory* variable is *age*, and
- the *response* variable is the maximum distance at which the driver can read a highway sign, or *maximum reading distance*.

Both variables are quantitative.

Here is what the raw data look like:

**Explanatory** **Response**

| | Age | Distance |
|---|---|---|
| | | |
| **Driver 1** | **18** | **510** |
| Driver 2 | 32 | 410 |
| Driver 3 | 55 | 420 |
| Driver 4 | 23 | 510 |
| . | . | . |
| . | . | . |
| . | . | . |
| Driver 30 | 82 | 360 |

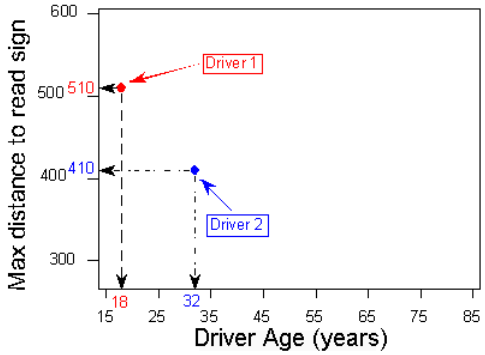
In this data set, the individuals are the 30 drivers. For each driver, we have two values: age and maximum reading distance.

To explore the relationship between age and distance, we create a graph called a **scatterplot**. To create a scatterplot, we use an ordered pair $(x, y)$ to represent the data for each driver. The ***x*-coordinate** is the explanatory variable: driver's age. The ***y*-coordinate** is the response variable: maximum reading distance.

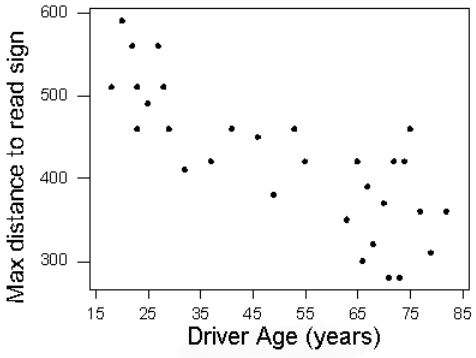
For this example, the ordered pair (18, 510) represents an 18-year-old driver who can read a highway sign at a maximum distance of 510 feet. We plot a point for each ordered pair. In the scatterplot, each driver appears as a single point.

Generally, each point in a scatterplot represents *one individual*. The $x$-coordinate is the value of the explanatory variable for that individual. The $y$-coordinate is the value of the response variable for that individual.

| | Age (X) | Distance (Y) |
|---|---|---|
| | | |
| **Driver 1** | **18** | **510** |
| **Driver 2** | **32** | **410** |
| Driver 3 | 55 | 420 |
| Driver 4 | 23 | 510 |
| . | . | . |
| . | . | . |
| . | . | . |
| Driver 30 | 82 | 360 |

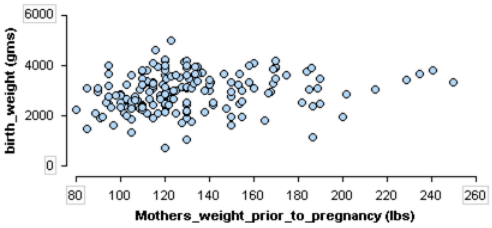



Here is the completed scatterplot:




## Learn By Doing

Recall this dataset from a medical study. In this study researchers collected data on new mothers to identify variables connected to low birth weights. This scatterplot investigates the relationship between two quantitative variables in the study: mother's weight prior to pregnancy and baby's birth weight.




https://assessments.lumenlearning.co...sessments/3856

https://assessments.lumenlearning.co...sessments/3466

https://assessments.lumenlearning.co...sessments/3467

https://assessments.lumenlearning.co...sessments/3468

https://assessments.lumenlearning.co...sessments/3469

## Comment

Remember: The explanatory variable is on the horizontal x-axis. The response variable is on the vertical y-axis. Sometimes the variables do not have a clear explanatory–response relationship. In this case, there is no rule to follow. Plot the variables on either axis.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.3: Scatterplots (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

https://stats.libretexts.org/@go/page/32013

# 3.4: Scatterplots (2 of 5)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

## Interpreting the Scatterplot

How do we describe the relationship between two quantitative variables using a scatterplot? We describe the overall pattern and deviations from that pattern.

This is the same way we described the distribution of one quantitative variable using a dotplot or a histogram in *Summarizing Data Graphically and Numerically*. To describe the overall pattern of the distribution of one quantitative variable, we describe the shape, center, and spread. We also describe deviations from the pattern (outliers).

Similarly, in a scatterplot, we describe the overall pattern with descriptions of **direction**, **form**, and **strength**. Deviations from the pattern are still called outliers.

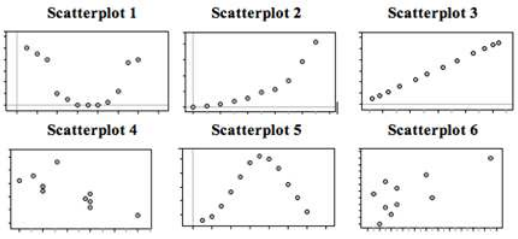
- A *positive (or increasing) relationship* means that an increase in one of the variables is associated with an increase in the other.

  A *negative (or decreasing) relationship* means that an increase in one of the variables is associated with a decrease in the other.
- The form of the relationship is its general shape. To identify the form, describe the shape of the data in the scatterplot. In practice, forms that we commonly use have mathematical equations. We look at a few of these equations in this course. For now, we simply describe the shape of the pattern in the scatterplot. Here are a couple of forms that are quite common:**Linear** form: The data points appear scattered about a line. We use a line to summarize the pattern in the data. We study the equation for a line in this module.
- In the top scatterplot, the data points closely follow the linear pattern. This is an example of a *strong linear* relationship. In the bottom scatterplot, the data points also follow a linear pattern, but the points are not as close to the line. The data is more scattered about the line. This is an example of a *weaker linear* relationship.

  Labeling a relationship as strong or weak is not very precise. We develop a more precise way to measure the strength of a relationship shortly.

*Outliers* are points that deviate from the pattern of the relationship. In the scatterplot below, there is one outlier.

## Learn By Doing




Fill in the letter of the description that matches each scatterplot.

Descriptions:

**A:** X = month (January = 1), Y = rainfall (inches) in Napa, CA in 2010 (Note: Napa has rain in the winter months and months with little to no rainfall in summer.)

**B:** X = month (January = 1), Y = average temperature in Boston MA in 2010 (Note: Boston has cold winters and hot summers.)

**C:** X = year (in five-year increments from 1970), Y = Medicare costs (in $) (Note: the yearly increase in Medicare costs has gotten bigger and bigger over time.)

**D:** X = average temperature in Boston MA (°F), Y = average temperature in Boston MA (°C) each month in 2010

**E:** X = chest girth (cm), Y = shoulder girth (cm) for a sample of men

**F:** X = engine displacement (liters), Y = city miles per gallon for a sample of cars (Note: engine displacement is roughly a measure of engine size. Large engines use more gas.)

https://assessments.lumenlearning.co...sessments/3470

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.4: Scatterplots (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.5: Scatterplots (3 of 5)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

Now we return to our previous example. We apply the ideas of direction, form, and strength to describe the relationship between the age of the driver and the maximum distance to read a highway sign. Here is the scatterplot:

**Direction:** The direction of the relationship is negative. An increase in age is associated with a decrease in reading distance, which makes sense because older drivers tend to have diminished eyesight. So most older drivers can read the sign only when they are close to it. In other words, they have a shorter maximum reading distance.

**Form:** The form of the relationship is linear.

**Strength:** The data points are fairly close to the line, so the relationship is moderately strong. Do not worry if you feel uncertain about describing the strength of a relationship. We mentioned earlier that descriptions of strength are not very precise. We develop a more precise measure of the strength shortly.

**Outliers:** There are no outliers. All the data points tend to follow the linear pattern.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.5: Scatterplots (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.6: Scatterplots (4 of 5)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

We now look at two more examples:

## Example

### Average Length of Pregnancy

What is the relationship between an animal's lifespan and the length of its pregnancy? To investigate this question, we have data from 40 different species of animals living in captivity. We use average lifespan as the explanatory variable, $x$. The average length of pregnancy is the response variable, $y$. (Source: Allen J. Rossman and Beth L. Chance, *Workshop Statistics: Discovery with Data and Minitab* [Key College Publishing, 2001]. Original source: *World Almanac and Book of Facts, 1993* [World Almanac, 1993].)

What can we learn about the relationship from the scatterplot?

The *direction* of the relationship is positive. An increase in lifespan is associated with an increase in pregnancy length. In other words, animals that live longer tend to have longer pregnancies. The *form* of the relationship is linear. The relationship is moderately *strong*.

Is there an outlier? There is a data point that deviates from the rest of the data because it has large $x$– and $y$-values. This is the elephant. Elephants live a long time (large $x$-value) and have a long pregnancy (large $y$-value). So the elephant is an outlier in the distribution of both the lifespan and the pregnancy variables. But this data point follows the positive direction of the data and fits the linear pattern. With respect to the form and direction of the relationship between the variables, this point is not an outlier.

Notice that the variation in pregnancy length is larger for animals that live longer. For example, animals that live 5 years have pregnancies that range from about 30 days to 120 days. The short, red vertical line on the left illustrates this range. Animals that live 12 years have pregnancies that vary more, ranging from about 60 days to over 400 days. The longer red vertical line on the right illustrates this range. So the relationship is stronger for animals with shorter lifespans.

## Example

### Fuel Usage

When you drive a car, what is the relationship between the speed you drive and the amount of gas the car uses? In this study, engineers measured the amount of fuel (in liters) used to drive 100 kilometers. They made these fuel measurements for a car driving at a fixed speed (in kilometers per hour). They then made fuel measurements for different fixed speeds.

What can we learn about the relationship from the scatterplot?

The data describe a relationship that decreases, then increases, so the direction of the relationship is negative and then becomes positive. In other words, at slow speeds, the car uses a lot of fuel. The amount of fuel decreases rapidly to a low point when the speed is 60 kilometers per hour, so the car uses the least amount of fuel at a speed of 60 km/h. The amount of fuel increases gradually for speeds above 60 km/h. This forms a *curvilinear* relationship that is very *strong*. All of the data fit a smooth curve.

Is there an outlier? The point (10, 21) lies above the rest of the data. With respect to speed ($x$), this point is not an outlier. The $x$-value does not deviate from the pattern for the other $x$-values in the data. In this study, it appears that the engineers varied the speeds by increments of 10 km/h. However, the $y$-value is much higher than the other $y$-values. With respect to fuel usage, this point is an outlier. But the point fits the overall curvilinear pattern in the data, so with respect to direction and form, this point is not an outlier.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3476

## Comment

In *Summarizing Data Graphically and Numerically*, we developed a method for identifying outliers in a distribution of one quantitative variable. The method was the 1.5 * IQR rule. In a scatterplot, you can use this rule to determine if the *x*-value of a point is an outlier with respect to the *x*-values in data. Similarly, you can use this rule to determine if a *y*-value of a point is an outlier with respect to the *y*-values in the data. However, this rule does not help us identify a point that deviates from the overall pattern in the data.

*Is there a method to identify outliers that deviate from the overall pattern in a scatterplot?* The answer is yes, but we do not discuss these techniques in this course. For now, just look at the scatterplot and see if a point deviates from the overall pattern. In other words, see if the point deviates from the direction and form of the data. We will see later that this type of outlier can influence measures of center and spread for two quantitative variables.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.6: Scatterplots (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.7: Scatterplots (5 of 5)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

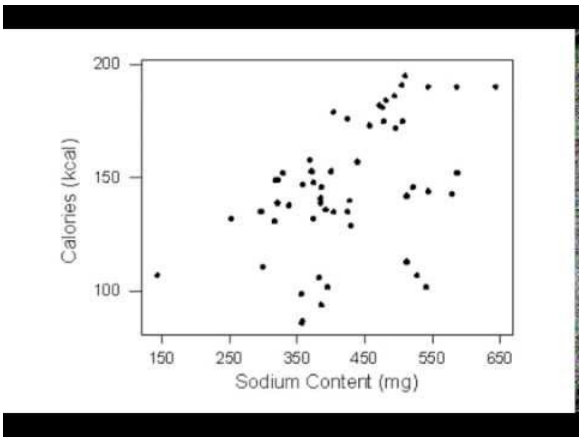
## Labeling Groups in a Scatterplot

If we graph data from two or more groups in a scatterplot, the relationship between the two quantitative variables can be hidden or unclear. We can use a categorical variable to label groups within the scatterplot, then look for patterns within each group. The relationship may be clearer within each group.

## Example

## Hot Dogs

A study was conducted by a concerned health group in which 54 major hot dog brands were examined. Using this data, we explore the relationship between sodium content and calories. We begin by making a scatterplot with data from the three types of hot dogs: beef, poultry, and meat (meat is a combination of pork, beef, and poultry).




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=122

## Let's Summarize

- The relationship between two quantitative variables is visually displayed using the scatterplot, where each point represents an individual. We always plot the explanatory variable on the horizontal x-axis and the response variable on the vertical y-axis.
- When we explore a relationship using the scatterplot, we should describe the *overall pattern* of the relationship and any *deviations* from that pattern. To describe the overall pattern, consider the *direction*, *form*, and *strength* of the relationship. Assessing the strength just by looking at the scatterplot can be problematic; using a numerical measure to determine strength is discussed later in this course.
- Adding labels to the scatterplot that indicate different groups or categories within the data might help us gain more insight about the relationship we are exploring.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.7: Scatterplots (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.8: Introduction- Linear Relationships

## What you'll learn to do: Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

### LEARNING OBJECTIVES

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.8: Introduction- Linear Relationships is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.9: Linear Relationships (1 of 4)

## Learning Objectives

- Use a scatterplot to display the relationship between two quantitative variables. Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.

## Introduction

So far, we have visualized relationships between two quantitative variables using scatterplots. We have also described the overall pattern of a relationship by considering its direction, form, and strength. We noted that it is difficult to assess the strength of a relationship just by looking at the scatterplot. In this section, we develop a numerical measure to assess the strength.

We focus only on relationships that have a linear form. Linear forms are quite common and relatively simple to detect. More important, we have a numerical measure that can assess the strength of the linear relationship. We use this measure along with the scatterplot to describe the linear relationship between two quantitative variables.

Even though we now focus only on linear relationships, remember that not every relationship between two quantitative variables has a linear form. We have already seen several examples of relationships that are not linear. However, the measure of strength that we are about to study can be used only with linear relationships. If we use this measure with nonlinear relationships, we will draw incorrect conclusions about the relationship between the variables.

Let's start with an example. Consider the following two scatterplots.

We can see that in both cases, the direction of the relationship is *positive* and the form of the relationship is *linear*. What about the strength? Recall that the strength of a relationship is a description of how closely the data follow its form.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3465

The scale used in a scatterplot can sometimes affect our assessment of strength, so we need to develop a measure for the strength of a linear relationship between two quantitative variables.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.9: Linear Relationships (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.10: Linear Relationships (2 of 4)

## Learning Objectives

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

## The Correlation Coefficient (*r*)

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.10: Linear Relationships (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.11: Linear Relationships (3 of 4)

## Learning Objectives

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

Now we interpret the value of $r$ in the context of some familiar examples.

## Example

### Highway Sign

In a previous example, we looked at this scatterplot to investigate the relationship between the age of a driver and the maximum distance at which the driver can read a highway sign. Because the form of the relationship is linear, we can use the correlation coefficient as a measure of direction and strength of the linear relationship.

The $r$-value is $-0.793$. The $r$-value is negative ($r < 0$), which means that the linear relationship has a negative direction. We can see this in the scatterplot. Because $r$ is somewhat close to $-1$, the relationship is moderately strong.

In the context of the data, the negative correlation confirms that the maximum reading distance decreases with age. Because $r$ indicates a moderately strong linear relationship, we expect that drivers of similar age will have some (but not a lot) of variability in their maximum reading distance.

## Example

### Biology Courses

A biology department is interested in tracking the progress of its students from entry until graduation. As part of the study, the department tabulates the performance of 10 students in an introductory course and later in an upper-level course required for graduation. What is the relationship between the students' course grades in the two courses? Here are two scatterplots of the *same* data.

Both scatterplots show a relationship that is positive in direction and linear in form. The strength appears different in the two scatterplots because of the difference in scales. This illustrates why we support our visual assessment of strength with a measurement of strength. We can use the correlation coefficient as a measure of the strength of the linear relationship. The correlation coefficient is $r = 0.91$, which is close to 1. The correlation coefficient confirms that the linear relationship is very strong.

## Comment

Note that in both examples, we supplemented the scatterplot with the correlation ($r$). Now that we have the correlation, why do we still need to look at a scatterplot when examining the relationship between two quantitative variables?

The correlation coefficient can be interpreted *only* as the *measure of the strength of a linear relationship*, so we need the scatterplot to verify that the relationship indeed looks linear. This point and its importance will be clearer after we examine a few properties of $r$.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.11: Linear Relationships (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.12: Linear Relationships (4 of 4)

## Learning Objectives

- Use a correlation coefficient to describe the direction and strength of a linear relationship. Recognize its limitations as a measure of the relationship between two quantitative variables.

## Properties of *r*

## Let's Summarize

- A special case of the relationship between two quantitative variables is the **linear** relationship in which a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the correlation coefficient ($r$), which measures the *strength* and *direction* of a linear relationship between two quantitative variables. The correlation ranges between −1 and 1. Values near −1 indicate a strong negative linear relationship, values near 0 indicate a weak linear relationship, and values near 1 indicate a strong positive linear relationship.
- The correlation is an appropriate numerical measure only for linear relationships and is sensitive to outliers. Therefore, the correlation should be used only as a supplement to a scatterplot (after we look at the data).

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.12: Linear Relationships (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.13: Introduction- Association vs Causation

## What you'll learn to do: Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

### LEARNING OBJECTIVES

- Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.13: Introduction- Association vs Causation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.14: Causation and Lurking Variables (1 of 2)

## Learning Objectives

- Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

## Introduction

A common mistake people make when describing the relationship between two quantitative variables is that they confuse *association* and *causation*. This mistake is so common that we devote this entire section to clarifying the difference.

This confusion often occurs when there is a strong relationship between the two quantitative variables. In the case of a linear relationship, people mistakenly interpret an $r$-value that is close to 1 or -1 as evidence that the explanatory variable *causes* changes in the response variable. In this case, the *correct interpretation* is that there is a **statistical relationship** between the variables, not a causal link. In other words, the explanatory variable and the response variable vary together in a predictable way. There is an **association** between the variables. But this *should not* be interpreted as a cause-and-effect relationship.

Let's look at an example.

## Example

## Fire Damage

The scatterplot below shows the relationship between the number of firefighters sent to fires ($x$) and the amount of damage caused by fires ($y$) in a certain city.

The scatterplot shows a positive association with a somewhat strong curvilinear form. An increase in the number of firefighters is associated with an increase in the damage done by the fire.

Can we conclude that the increase in firefighters causes the increase in damage? Of course not.

A third variable is at play in the background – the seriousness of the fire – and is responsible for the observed relationship. More serious fires require more firefighters and also result in more damage.

The following figure will help you visualize this situation:

The seriousness of the fire is a **lurking variable**. A lurking variable is a variable that is not measured in the study. It is a third variable that is neither the explanatory nor the response variable, but it affects your interpretation of the relationship between the explanatory and response variables.

In our example, the lurking variable has an effect on both the explanatory and the response variables. This common effect creates the observed association between the explanatory and response variables even though there is no cause-and-effect link between them.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3851

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3852

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.14: Causation and Lurking Variables (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.15: Causation and Lurking Variables (2 of 2)

## Learning Objectives

- Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

In the next example, we investigate a subtle point about the confusion between association and causation. In this example, a cause-and-effect connection is logical but not justified by an observed association in a single study.

## Example

## Smoking and Lung Cancer

In this data, $x$ = cigarette consumption per capita in the United States, and $y$ = lung cancers per 100,000. To investigate the connection between cigarette consumption and lung cancers, the data is offset by 30 years because cancer takes time to develop. For example, cigarette consumption in 1945 is paired with cancer rates for 1975.

In the scatterplot, we see a fairly strong positive correlation.

Can we conclude from this data that cigarette smoking causes lung cancer? The answer is no.

The data comes from an observational study. Recall from our previous discussions in Module 1 that we can draw cause-and-effect conclusions only from randomized comparative experiments. From this study, we can say that cigarette smoking is **associated** with lung cancer. We can also say that cigarette smoking **correlates** with lung cancer. We *cannot* say that cigarette smoking **causes** lung cancer.

Yet the National Cancer Institute's website states that "cigarette smoking causes many types of cancer, including cancers of the lung" (National Cancer Institute).

How can this be? Did the National Cancer Institute conduct a randomized comparative experiment to establish this cause-and-effect relationship? Of course not. We cannot randomly assign people to smoke or not smoke. All of the studies linking smoking with cancer are observational studies. Alone, each study can show only an association.

So is it possible to draw a causal link between cigarette consumption and cancer rates? The answer is yes, well sort of. In practice, researchers use criteria such as the following to provide evidence of a causal connection from observational studies:

- There is a reasonable explanation for how one variable might cause the other.
- The association is seen in repeated studies under varying conditions.
- The effects of potential lurking variables are ruled out when we look across studies.

The point of the previous example is again that association does not imply causation. But researchers can use an *observed association as the first step in building a case for causation.*

This point is subtle but important. When experiments cannot be conducted, it can be difficult and controversial to explain an observed association between two variables. Many of the current disputes involving data and statistics involve questions of causation that we cannot investigate through an experiment. Does the death penalty reduce violent crime? Does cell phone use cause brain tumors? Does pollution cause global warming? All of these questions imply a cause-and-effect relationship in situations that are complex and involve many interacting variables. In these situations, a single observational study cannot establish a causal link between two variables. But researchers can use the observed association as a first step in building a case for causation.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3853

### Let's Summarize

- The relationship between two quantitative variables is visually displayed using the *scatterplot*, where each point represents an individual. We always plot the explanatory variable on the horizontal axis and the response variable on the vertical axis.
- When we explore a relationship using the scatterplot, we should describe the *overall pattern* of the relationship and any *deviations* from that pattern. To describe the overall pattern, consider the *direction*, *form*, and *strength* of the relationship.

- Adding labels to the scatterplot that indicate different groups or categories within the data might help us gain more insight about the relationship we are exploring.
- A special case of the relationship between two quantitative variables is the *linear* relationship. In this case, a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the *correlation coefficient (r)*, which measures the *strength* and the *direction* of a linear relationship between two quantitative variables. The correlation ranges between -1 and 1. Values near -1 indicate a strong negative linear relationship. Values near 0 can indicate a weak or no linear relationship. Values near 1 indicate a strong positive linear relationship. Remember, we use the correlation coefficient only *after* we have looked at the data and observed that there is a linear relationship. If you have no information about what the data actually looks like, then you should not use the correlation coefficient in your analysis.
- The correlation is an appropriate numerical measure only for linear relationships, and it is sensitive to outliers. Therefore, the correlation should be used only as a supplement to a scatterplot (after we look at the data).
- A *lurking variable* is a variable that is not measured in the study. It is a third variable that is neither the explanatory nor the response variable, but it affects your interpretation of the relationship between the explanatory and response variable.
- *Association does not imply causation.* Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
- An observational study alone cannot establish a causal connection between explanatory and response variables. To establish a cause-and-effect relationship, researchers must conduct a comparative randomized experiment. In reality, it is often impossible to conduct an experiment. So observational studies that show an association between two variables can be used as a first step in building a case for causation.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.15: Causation and Lurking Variables (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.16: Introduction- Linear Regression

## What you'll learn to do: For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

### LEARNING OBJECTIVES

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.16: Introduction- Linear Regression is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.17: Linear Regression (1 of 4)

## Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

So far we have used a scatterplot to describe the relationship between two quantitative variables. We described the pattern in the data by describing the direction, form, and strength of the relationship. We then focused on linear relationships. When the relationship is linear, we used correlation ($r$) as a measure of the direction and strength of the linear relationship.

Our focus on linear relationships continues here. We will

- use lines to make predictions.
- identify situations in which predictions can be misleading.
- develop a measurement for identifying the best line to summarize the data.
- use technology to find the best line.
- interpret the parts of the equation of a line to make our summary of the data more precise.

## Making Predictions

Earlier, we examined the linear relationship between the age of a driver and the maximum distance at which the driver can read a highway sign. Suppose we want to predict the maximum distance that a 60-year-old driver can read a highway sign. In the original data set, we do not have a 60-year-old driver.

How could we make a prediction using the linear pattern in the data?

Here again is the scatterplot of driver ages and maximum reading distances . (Note: Sign Legibility Distance = Max distance to read sign.) We marked 60 on the x-axis.

Of course, different 60-year-olds will have different maximum reading distances . We expect variability among individuals. But here our goal is to make a single prediction that follows the general pattern in the data. Our first step is to model the pattern in the data with a line. In the scatterplot, you see a red line that follows the pattern in the data.

To use this line to make a prediction, we find the point on the line with an $x$-value of 60. Simply trace from 60 directly up to the line. We use the $y$-value of this point as the predicted maximum reading distance for a 60-year-old. Trace from this point across to the y-axis.

We predict that 60-year-old drivers can see the sign from a maximum distance of just under 400 feet.

We can also use the equation for the line to make a prediction. The equation for the red line is

Predicted distance = 576 − 3 * Age

To predict the maximum distance for a 60-year-old, substitute Age = 60 into the equation.

Predicted distance = 576 − 3 * (60) = 396 feet

Shortly, we develop a measurement for identifying the best line to summarize the data. We then use technology to find the equation of this line. Later, in “Assessing the Fit of a Line,” we develop a method to measure the accuracy of the predictions from this “best” line. For now, just focus on how to use the line to make predictions.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3489

https://assessments.lumenlearning.co...sessments/3490

Before we leave the idea of prediction, we end with the following cautionary note:

*Avoid making predictions outside the range of the data.*

Prediction for values of the explanatory variable that fall outside the range of the data is called **extrapolation**. These predictions are unreliable because we do not know if the pattern observed in the data continues outside the range of the data. Here is an example.

Example

## Cricket Thermometers

Crickets chirp at a faster rate when the weather is warm. The scatterplot shows data presented in a 1995 issue of *Outside* magazine. Chirp rate is the number of chirps in 13 seconds. The temperature is in degrees Fahrenheit.

There is a strong relationship between chirp rate and temperature when the chirp rate is between about 18 and 45. What form does the data have? This is harder to determine. A line appears to summarize the data well, but we also see a curvilinear form, particularly when we pay attention to the first and last data points.

Both the curve and line are good summaries of the data. Both give similar predictions for temperature when the chirp rate is within the range of the data (between 18 and 45). But outside this range, the curve and the line give very different predictions. For example, if the crickets are chirping at a rate of 60, the line predicts a temperature just above 95°F. The curve predicts a much lower temperature of about 85°F.

Which is a better prediction? We do not know which is better because we do not know if the form is linear or curvilinear outside the range of the data.

If we use our model (the line or the curve) to make predictions outside the range of the data, this is an example of extrapolation. We see in this example that extrapolation can give unreliable predictions.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3491

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3861

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.17: Linear Regression (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.18: Linear Regression (2 of 4)

## Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.18: Linear Regression (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.19: Linear Regression (3 of 4)

## Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

Let's quickly revisit the list of our data analysis tools for working with linear relationships:

- Use a scatterplot and *r* to describe direction and strength of the linear relationship.
- Find the equation of the least-squares regression line to summarize the relationship.
- Use the equation and the graph of the least-squares line to make predictions.
- Avoid extrapolation when making predictions.

Now we focus on the equation of a line in more detail. Our goal is to understand what the numbers in the equation tell us about the relationship between the explanatory variable and the response variable.

Here are some of the equations of lines that we have used in our discussion of linear relationships:

$$\text{Predicted distance} = 576 - 3 * \text{Age}$$

$$\text{Predicted height} = 39 + 2.7 * \text{forearm length}$$

$$\text{Predicted monthly car insurance premium} = 97 - 1.45 * \text{years of driving experience}$$

Notice that the form of the equations is the same. In general, each equation has the form

$$\text{Predicted } y = a + b * x$$

When we find the least-squares regression line, *a* and *b* are determined by the data. The values of *a* and *b* do not change, so we refer to them as **constants**.

In the equation of the line, the constant *a* is the prediction when $x = 0$. It is called **initial value**. In a graph of the line, *a* is the ***y*-intercept**.

In the equation of the line, the constant *b* is the rate of change, called the **slope**. In a graph of the least-squares line, *b* describes how the predictions change when *x* increases by one unit. More specifically, *b* describes the average change in the response variable when the explanatory variable increases by one unit.

We can write the equation of the line to reflect the meaning of *a* and *b:*

$$\text{Predicted } y = a + b * x$$

$$\text{Predicted } y\text{-value} = (\text{initial value}) + (\text{rate of change}) * x$$

$$\text{Predicted } y\text{-value} = (y\text{-intercept}) + (\text{slope}) * x$$

The constants *a* and *b* are shown in the graph of the line below.

## Algebra review

## The algebra of a line

The general form for the equation of a line is Y = a + bX. The constants "a" and "b" can be either positive or negative. The constant "a" is the y-intercept where the line crosses the y-axis. The constant "b" is the slope. It describes the steepness of the line. In algebra we describe the slope as "rise over run". The slope is the amount that Y increases (or decreases) for each 1-unit increase in X.

## EXAMPLE

## 1

Consider the line . The intercept is 1. The slope is 1/3, and the graph of this line is, therefore:

EXAMPLE

## 2

Consider the line . The intercept is 1. The slope is -1/3, and the graph of this line is, therefore:

The simulation below allows you to see how changing the values of the slope and y-intercept changes the line. The slider on the left controls the y-intercept, a. The slider on the right controls the slope, b.

Use the simulation to draw the following lines:

$$Y = 3 + 0.67X$$
$$Y = 5 – X \text{ (which can also be written } Y = 5 – 1.0X)$$
$$Y = 2X \text{ (which can also be written } Y = 0 + 2X)$$
$$Y = 5 – 2X$$

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=146

Use the following graphs in the next activity to investigate the equation of lines.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3483

https://assessments.lumenlearning.co...sessments/3492

### Interpreting the Slope and Intercept

The constants in the equation of a line give us important information about the relationship between the predictions and *x*. In the next examples, we focus on how to interpret the meaning of the constants in the context of data.

### Example

## Highway Sign Visibility Data

Recall that from a data set of 30 drivers, we see a strong negative linear relationship between the age of a driver (*x*) and the maximum distance (in feet) at which a driver can read a highway sign. The least-squares regression line is

**Predicted *y*-value = (starting value) + (rate of change)*x**

Predicted distance = 576 − 3 * Age

Predicted distance = 576 + (−3 * Age)

The value of *b* is −3. This means that a 1-year increase in age corresponds to a predicted 3-foot decrease in maximum distance at which a driver can read a sign. Another way to say this is that there is an average decrease of 3 feet in predicted sign visibility distance when we compare drivers of age *x* to drivers of age *x* + 1.

The 576 is the predicted value when *x* = 0. Obviously, it does not make sense to predict a maximum sign visibility distance for a driver who is 0 years old. This is an example of extrapolating outside the range of the data. But the starting value is an important part of the least-squares equation for predicting distances based on age.

The equation tells us that to predict the maximum visibility distance for a driver, start with a distance of 576 feet and subtract 3 feet for every year of the driver's age.

### Example

## Body Measurements

In the body measurement data collected from 21 female community college students, we found a strong positive correlation between forearm length and height. The least-squares regression line is

Predicted height = 39 + 2.7 * forearm length

The value of $b$ is 2.7. This means that a 1-inch increase in forearm length corresponds to a predicted 2.7-inch increase in height. Another way to say this is that there is an average increase of 2.7-inches in predicted height when we compare women with forearm length of $x$ to women with forearm length of $x + 1$.

The 39 is the predicted value when $x = 0$. Obviously, it does not make sense to predict the height of a woman with a 0-inch forearm length. This is another example of extrapolating outside the range of the data. But 39 inches is the starting value in the least-squares equation for predicting height based on forearm length.

The equation tells us that to predict the height of a woman, start with 39 inches and add 2.7 inches for every inch of forearm length.

In the graph below, we see the slope $b$ represented by a triangle. An 8-inch increase in foreman length corresponds to a 21.6-inch increase in predicted height. $b = 21.6 / 8 = 2.7$. An arrow points to the starting value $a = 39$. This is the point with $x = 0$.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3863

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.19: Linear Regression (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.20: Linear Regression (4 of 4)

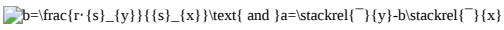
## Learning Objectives

- For a linear relationship, use the least squares regression line to model the pattern in the data and to make predictions.

In the previous activity we used technology to find the least-squares regression line from the data values.

We can also find the equation for the least-squares regression line from summary statistics for $x$ and $y$ and the correlation.

If we know the mean and standard deviation for $x$ and $y$, along with the correlation ($r$), we can calculate the slope $b$ and the starting value $a$ with the following formulas:



$$b=\frac{r\cdot s_y}{s_x}\text{ and }a=\bar{y}-b\bar{x}$$

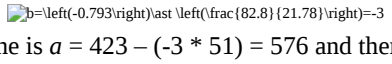
As before, the equation of the linear regression line is

$$\text{Predicted } y = a + b * x$$

## Example: Highway Sign Visibility

We will now find the equation of the least-squares regression line using the output from a statistics package.

```
> summary(data)
   Age             Distance
   Min.   :18      Min.   :280
   1st Qu.:21.8    1st Qu.:82.8
   Median :54      Median :420
   Mean   :51      Mean   :423
   3rd Qu.:71.3    3rd Qu.:467.5
   Max    :82      Max    :590
> cor(data$Age,data$Distance)
   [1] -0.793
```

- The **slope** of the line is 
$$b=\left(-0.793\right)\ast \left(\frac{82.8}{21.78}\right)=-3$$
- The **intercept** of the line is $a$ = 423 – (-3 * 51) = 576 and therefore the **least-squares regression line** for this example is Predicted distance = 576 + (-3 * Age), which can also be written as Predicted distance = 576 – 3 * Age

## Learn By Doing

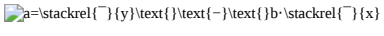
https://assessments.lumenlearning.co...sessments/3864

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3488

Now you know how to calculate the least-squares regression line from the correlation and the mean and standard deviation of $x$ and $y$. But what do these formulas tell us about the least-squares line?

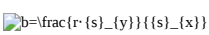
We know that the intercept $a$ is the predicted value when $x$ = 0.

The formula 
$$a=\bar{y}\text{}\text{}\text{}b\cdot\bar{x}$$
tells us that the we can find the intercept using the point: (
$$\overline{x},\overline{y}$$
).

This is interesting because it says that every least-squares regression line contains this point. In other words, the least-squares regression line goes through the mean of $x$ and the mean of $y$.

We also know that the slope of the least-squares regression line is the average change in the predicted response when the explanatory variable increases by 1 unit.

The slope formula



$$b=\frac{r\cdot s_y}{s_x}$$

tells us that the slope is related to the correlation in this way: when $x$ increases an $x$ standard deviation, the predicted $y$-value does not change by a $y$ standard deviation. Instead, the predicted $y$-value changes by less than a $y$ standard deviation. The change is a fraction of a $y$ standard deviation, and that fraction is $r$. Another way to say this is that when $x$ increases by a standard deviation in $x$, the average change in the predicted response is a fractional change of $r$ standard deviations in $y$.

It is not surprising that slope and correlation are connected. We already know that when a linear relationship is positive, the correlation and the slope are positive. Similarly, when a linear relationship is negative, the correlation and slope are both negative. But now we understand this connection more precisely.

### Let's Summarize

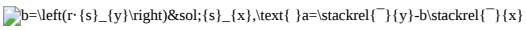
- The line that best summarizes a linear relationship is the least-squares regression line. The least-squares line is the best fit for the data because it gives the best predictions with the least amount of overall error. The most common measurement of overall error is the sum of the squares of the errors (SSE). The least-squares line is the line with the smallest SSE.
- We use the least-squares regression line to predict the value of the response variable from a value of the explanatory variable.
- Prediction for values of the explanatory variable that fall outside the range of the data is called extrapolation. These predictions are unreliable because we do not know if the pattern observed in the data continues outside the range of the data. Avoid making predictions outside the range of the data.
- The slope of the least-squares regression line is the average change in the predicted values of the response variable when the explanatory variable increases by 1 unit.
- We have two methods for finding the equation of the least-squares regression line:

$$\text{Predicted } y = a + b * x$$

**Method 1:** We use technology to find the equation of the least-squares regression line:

$$\text{Predicted } y = a + b * x$$

**Method 2:** We use summary statistics for $x$ and $y$ and the correlation. In this method we can calculate the slope $b$ and the $y$-intercept $a$ using the following:




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.20: Linear Regression (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.21: Introduction- Assessing the Fit of a Line

## What you'll learn to do: Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

### LEARNING OBJECTIVES

- Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.21: Introduction- Assessing the Fit of a Line is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.22: Assessing the Fit of a Line (1 of 4)

## Learning Objectives

- Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

## Introduction

Let's take a moment to summarize what we have done up to this point in *Examining Relationships: Quantitative Data*. Our goal from the beginning was to *examine the relationship between two quantitative variables*. We started by looking at scatterplots to see if we could see any pattern between the explanatory and response variables. We focused early in the course on identifying those cases that were *linear* in form. At the same time, we assessed how strong the linear relationship was on the basis of visual inspection. As is our usual strategy, we turned from graphs to numeric measures, and in particular, we developed the correlation coefficient, *r*, as a measure of the strength of the linear relationship we observed in the graph.

Once we established that there was a linear relationship between explanatory and response variables, the next step was to find a line that fit the data: the *best-fit line*. Here we used the least-squares method to find the regression line. Finally, we used the equation of the regression line to predict the value of the response variable for a given value of the explanatory variable.

## How Good Is the Best-Fit Line?

Now that we have a mathematical model (the least-squares regression line) that we can use to make predictions, we want to know: How good are these predictions, and how can we measure the error in a prediction?

## Example

### Highway Sign Visibility

Let's begin our investigation by predicting the maximum distance that an 18-year-old driver can read a highway sign and then determining the error in our prediction.

We use the regression line equation:

$$\text{Distance} = 576 + (-3 * \text{Age})$$

To predict the distance for an 18-year-old driver, we plug Age = 18 into the equation.

$$\text{Predicted distance} = 576 + (-3 * 18) = 522$$

Our prediction is that 522 feet is the maximum distance at which an 18-year-old driver can read a highway sign. Now let's compare our prediction to the actual data for the 18-year-old driver: (18, 510).

$$\text{The error in our prediction is } 510 - 522 = -12.$$

This tells us that the actual distance for the 18-year-old driver is 12 feet closer than the prediction. In other words, our prediction is too large. It overestimates the actual distance by 12 feet.

So in general, we have Observed data value – Predicted value = Error.

If we use $(x, y)$ to represent a typical data point and $\hat{y}$ to represent the predicted value (obtained by using the regression equation), then we have

$$\begin{array}{l}\text{observed}-\text{predicted}=\text{error}\\ y-\hat{y}=\text{error}\end{array}$$

## Learn By Doing

Using this table showing "observed" and "predicted" distances for some drivers, find the following:

| | Age | Distance (observed) | Distance (predicted) | Error observed - predicted |
|---|---|---|---|---|
| Driver 1 | 18 | 510 | 576+(-3)(18) = 522 | -12 |
| Driver 2 | 32 | 410 | 576+(-3)(32) = 480 | -70 |
| Driver 3 | 55 | 420 | 576+(-3)(55) = 411 | 9 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| Driver 30 | 82 | 360 | . | . |

https://assessments.lumenlearning.co...sessments/3497

https://assessments.lumenlearning.co...sessments/3498

https://assessments.lumenlearning.co...sessments/3499

Now let's look at the error from a different perspective. We can think of the error as a way to adjust the prediction to match the data value.

From this point of view, we rewrite $y-\hat{y}=\text{error}$ as $y=\hat{y}+\text{error}$ .

This last equation says that the observed value is the predicted value plus the error. In other words, we can think of the error as the amount that we have to add to the prediction to get the observed value. From this point of view, the error can be thought of as a *correction term*. If the error is positive, it means the prediction is too small (the prediction underestimates the actual *y*-value). If the error is negative, it means the prediction is too large (the prediction overestimates the actual *y*-value).

The prediction error is also called a **residual**. So another way to express the previous equation is

$$y=\hat{y}+\text{residual}$$

In our next example, we look at prediction error from this point of view.

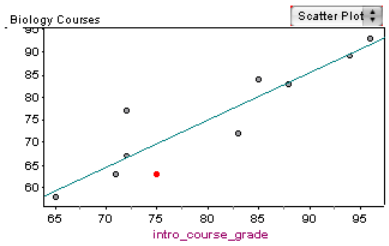
### Example

## Biology Courses

A biology department tracks the progress of students in its program. Grades in the introductory biology course have a strong linear relationship with grades in the upper-level biology courses ($r = 0.91$).

The least-squares regression equation is

$$\text{Upper course grade} = -8.9 + (1.05 * \text{Intro course grade})$$




| | Intro grade | Upper grade |
|---|---|---|
| Student 1 | 65 | 58 |
| Student 2 | 71 | 63 |
| Student 3 | 72 | 67 |
| Student 4 | 72 | 77 |
| Student 5 | 75 | 63 |
| Student 6 | 83 | 72 |
| Student 7 | 85 | 84 |
| Student 8 | 88 | 83 |
| Student 9 | 94 | 89 |
| Student 10 | 96 | 93 |

Let's look at the predicted upper course grade for a student who makes a 75% in the introductory biology course.

$$\text{Upper course grade} = -8.9 + (1.05 * 75) = 69.85 \approx 70$$

The regression line predicts that this student will make a 70% in the upper-level biology course.

The actual grade in the upper-level course for this student is 63%. The prediction is too high: it overestimates the data. To match the data value, we would need to subtract 7 from the prediction, so the error is −7.

In the scatterplot, notice that the regression line lies above the point (75, 63). Visually, we can see that the prediction is too high. This reinforces our previous observation that the prediction overestimates the data value. We would have to adjust the prediction downward to match the data value. Viewing the error as a correction term, we see the correction has to be negative.

Notice that when a point is close to the regression line, the prediction is close to the actual upper course grade, so the error is small. Another way to say this is that points close to the regression line have a small residual.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3500

https://assessments.lumenlearning.co...sessments/3501

https://assessments.lumenlearning.co...sessments/3502

https://assessments.lumenlearning.co...sessments/3503

https://assessments.lumenlearning.co...sessments/3504

https://assessments.lumenlearning.co...sessments/3505

https://assessments.lumenlearning.co...sessments/3506

https://assessments.lumenlearning.co...sessments/3507

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.22: Assessing the Fit of a Line (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.23: Assessing the Fit of a Line (2 of 4)

## Learning Objectives

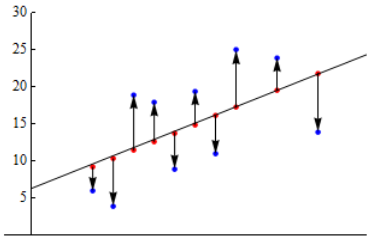
- Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

## Introduction

Now we move from calculating the residual for an individual data point to creating a graph of the residuals for all the data points. We use residual plots to determine if the linear model fits the data well.

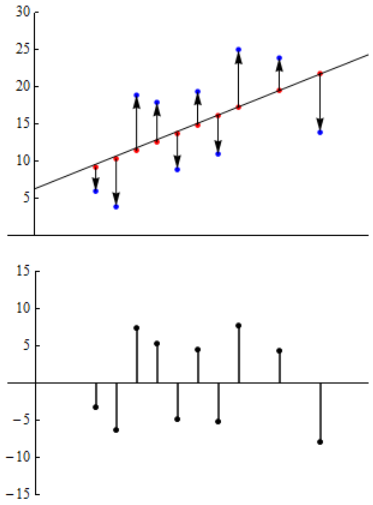
## Residual Plots

The graph below shows a scatterplot and the regression line for a set of 10 points. The blue points represent our original data set, that is, our observed values. The red points, lying directly on the regression line, are the predicted values.




The vertical arrows from the predicted to observed values represent the residuals. The up arrows correspond to positive residuals, and the down arrows correspond to negative residuals.

Now consider the following pair of graphs. The top graph is a copy of the graph we looked at above. In the graph below, we plotted the values of the residuals on their own. (The explanatory variable is still plotted on the horizontal axis, though it is not indicated this here.) This is called a **residual plot**.




In the residual plot, each point with a value greater than zero corresponds to a data point in the original data set where the observed value is greater than the predicted value. Similarly, negative values correspond to data points where the observed value is less than the predicted value.

**What are we looking for in a residual plot?**

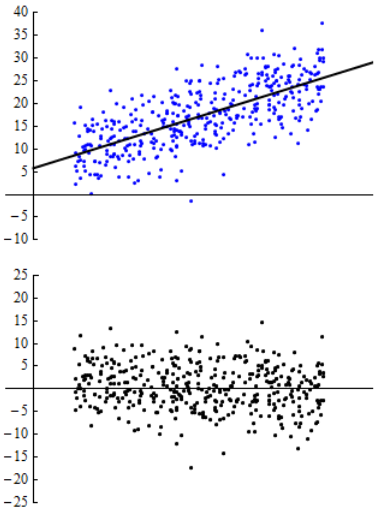
We use residual plots to determine if a linear model is appropriate. In particular, we look for any *unexpected patterns* in the residuals that may suggest that the data is not linear in form.

To help us identify an unexpected pattern, we start by looking at what we *expect* to see in a residual plot *when the form is linear.*

Example

## No Pattern in Residual Plot

Consider the pair of graphs below. Here we have a scatterplot for a data set consisting of 400 observations. The regression line is shown in the scatterplot. The residual plot is below the scatterplot.




In this example, the line in the scatterplot is a good summary of the positive linear pattern in the data. Notice that the points in the residual plot seem to be randomly scattered. As we examine the residuals from left to right, they don't appear to follow a particular path, nor does the cloud of points widen or narrow in any systematic way. We see no particular pattern. Thus, in the ideal case, when a linear model is really a good fit, we expect to see *no pattern* in the residual plot.

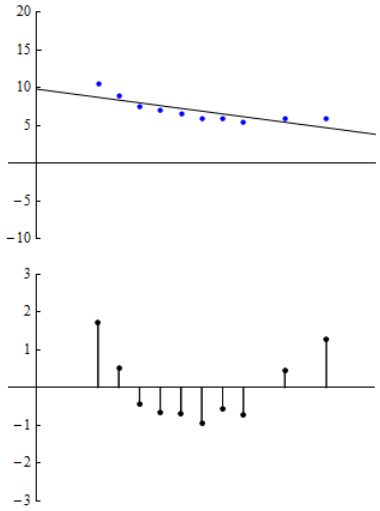
Our general principle when looking at residual plots, then, is that a residual plot with *no* pattern is good because it suggests that our use of a linear model is appropriate.

However, we must be flexible in applying this principle because what we see usually lies somewhere between the extremes of no pattern and a clear pattern. Let's look at some specific examples.

Example

## Patterns in Residual Plots

At first glance, the scatterplot appears to show a strong linear relationship. The correlation is $r = 0.84$. However, when we examine the residual plot, we see a clear U-shaped pattern. Looking back at the scatterplot, this movement of the data points above, below and then above the regression line is noticeable. The residual plot, particularly when graphed at a finer scale, helps us to focus on this deviation from linearity.

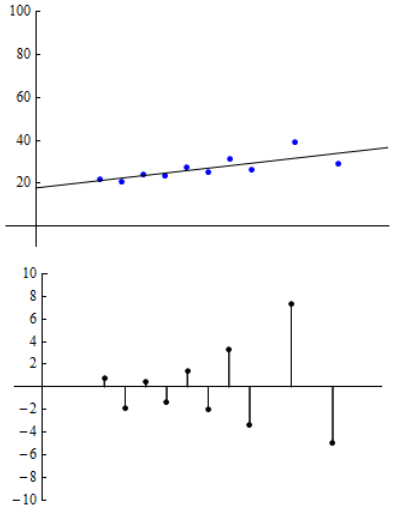



The pattern in the residual plot suggests that our linear model may not be appropriate because the model predictions will be too high for values in the middle of the range of the explanatory variable and too low for values at the two ends of that range. A model with a curvilinear form may be more appropriate.

### Example

## Patterns in Residual Plots 2

This scatterplot is based on datapoints that have a correlation of $r = 0.75$. In the residual plot, we see that residuals grow steadily larger in absolute value as we move from left to right. In other words, as we move from left to right, the observed values deviate more and more from the predicted values. Again, we have chosen a smaller vertical scale for the residual plot to help amplify the pattern to make it easier to see.

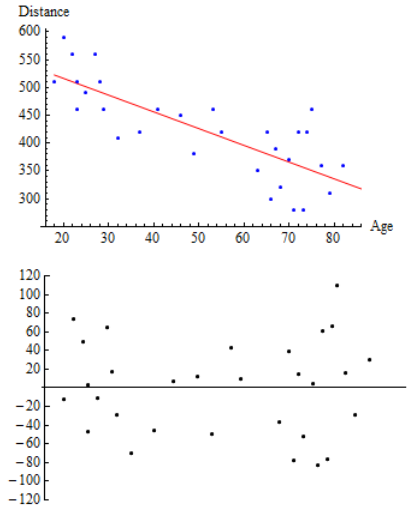



The pattern in the residual plot suggests that predictions based on the linear regression line will result in greater error as we move from left to right through the range of the explanatory variable.

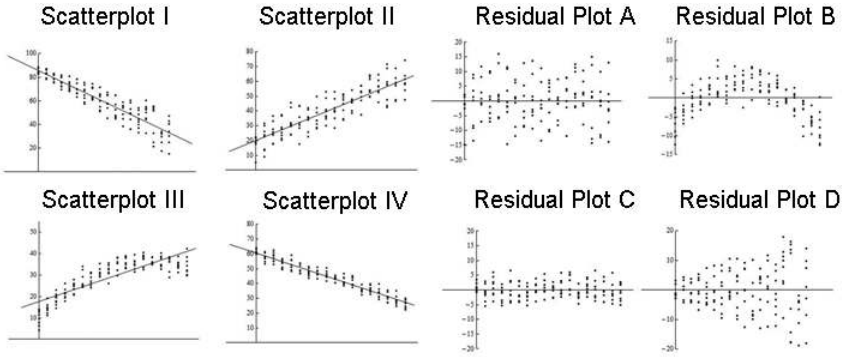
### Example

## Highway Sign Visibility

Let’s return now to our original example and take a look at what the residual plot tell us about the appropriateness of applying a linear model to this data.



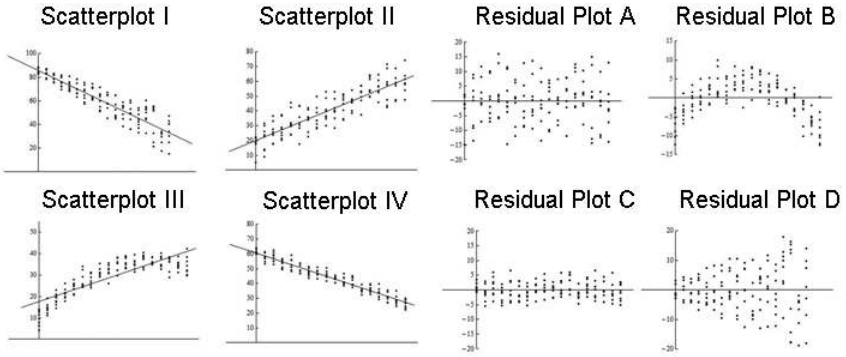

Note that the residuals are fairly randomly dispersed. However, they seem to be a bit more spread out on the left and right than they are in the middle. As we look at higher ages, there seems to be greater variation in the residuals, which suggests that we may want to be more cautious if we are trying to predict distances for older drivers. And the risks associated with extrapolation beyond the range of the data seem to be even greater here. In this case, we may still use this linear model but condition the use of it on our analysis of the residual plot.




## Learn By Doing

https://assessments.lumenlearning.co...sessments/3508

Here again are four scatterplots with regression lines shown and four corresponding residual plots.




## Learn By Doing

https://assessments.lumenlearning.co...sessments/3509

https://assessments.lumenlearning.co...sessments/3510

https://assessments.lumenlearning.co...sessments/3511

https://assessments.lumenlearning.co...sessments/3512

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.23: Assessing the Fit of a Line (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.24: Assessing the Fit of a Line (3 of 4)

## Learning Objectives

- Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

## Introduction

Here we continue our discussion of the question, *How good is the best-fit line?*

Let's summarize what we have done so far to address this question. We began by looking at how the predictions from the least-squares regression line compare to observed data. We defined a residual to be the amount of error in a prediction. Next, we created residual plots. A residual plot with no pattern reassures us that our linear model is a good summary of the data.

But how do we know if the explanatory variable we chose is really the best predictor of the response variable?

The regression line does not take into account other variables that might also be good predictors. So let's investigate the question, *What proportion of the variation in the response variable does our regression line explain?*

We begin our investigation with a scatterplot of the daily high temperature (°F) in New York City from January 1 to June 1. We have 4 years of data (2002, 2003, 2005, and 2006). The least-squares regression line has the equation $y = 36.29 + 0.25x$, where $x$ is the number of days after January 1. Therefore, January 1 corresponds to $x = 0$, and June 1 corresponds to $x = 151$.

Two things stand out as we look at this picture. First, we see a clear, positive linear relationship tracked by the regression line. As the days progress, there is an associated increase in temperature. Second, we see a substantial scattering of points around the regression line. We are looking at 4 years of data, and we see a lot of variation in temperature, so the day of the year only partially explains the increase in temperature. Other variables also influence the temperature, but the line accounts only for the relationship between the day of the year and temperature.

Now we ask the question, *Given the natural variation in temperature, what proportion of that variation does our linear model explain?*

The answer, which is surprisingly easy to calculate, is just the square of the correlation coefficient.

**The value of $r^2$ is the proportion of the variation in the response variable that is explained by the least-squares regression line.**

In the present case, we have $r = 0.73$; therefore, $\frac{\text{explained variation}}{\text{total variation}}={\text{0.73}}^{\text{2}}=\text{0.53}$. And so we say that our linear regression model explains 53% of the total variation in the response variable. Consequently, 47% of the total variation remains unexplained.

## Example

## Highway Sign Visibility

Recall that the least-squares regression line is Distance = 576 − 3 * Age. The correlation coefficient for the highway sign data set is −0.793, so $r^2 = (-0.793)^2 = 0.63$.

Our linear model uses age to predict maximum distance at which a driver can read a highway sign. Other variables may also influence reading distance. We can say the linear relationship between age and maximum reading distance accounts for 63% of the variation in maximum reading distance.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3513

https://assessments.lumenlearning.co...sessments/3868

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3514

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.24: Assessing the Fit of a Line (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.25: Assessing the Fit of a Line (4 of 4)

## Learning Objectives

- Use residuals, standard error, and $r^2$ to assess the fit of a linear model.

## Introduction

Our final investigation into assessing the fit of the regression line focuses on typical error in the predictions.

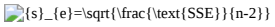
Previously, we calculated the error in a single prediction by calculating

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

But we use the regression line to make predictions even when we do not have an observed value, so we need a method for using all of the residuals to compute a typical amount of error.

We ask the question, *How do we measure the typical amount of error for predictions from the regression line?*

The most common measure of the size of the typical error is the **standard error of the regression**, which is represented by $s_e$. It is calculated using the following formula:



where *SSE* stands for the sum of the squared errors.

Finding the standard error of the regression is similar to finding the standard deviation of a distribution of data points from a single quantitative variable. In *Summarizing Data Graphically and Numerically*, we learned that the *standard deviation is roughly a measure of average distance about the mean*. Here the *standard error is roughly a measure of the average distance of the points about the regression line*.

Let's return to our example where age is used to predict the maximum distance for reading highway signs.

The residual plot for the highway sign data set is shown below. We can visualize the SSE in the formula as simply the sum of the squares of all of the vertical (residual) line segments. After dividing by $n - 2$, we have the average *squared* residual. Taking the square root then gives us a measure of the average size of the residuals.

In the case of the highway sign data, the value of $s_e$ is 51.35. In the figure below, we added horizontal lines at $y = 51.35$ and $y = -51.35$, so the red line represents the typical size of the error.

**Comment:** When we mark the $s_e$ on this residual plot, errors that fall outside of this range are larger than average. We see again that most of the errors that exceed ±51.35 are on the right. This illustrates that predictions of maximum reading distance for older drivers have larger error.

**Note:** Most statistics software computes $r$ and $r^2$ and $s_e$. Therefore, our focus is not on calculating but on understanding and interpreting.

Now let's apply the standard error of the regression as a measurement of typical error.

## Example

## Highway Sign Visibility

Let's take another look at the prediction we made earlier using the regression line equation:

$$\text{Distance} = 576 + (-3 * \text{Age})$$

In a previous example, we predicted the maximum distance that a 60-year-old driver can read a highway sign. We plugged Age = 60 into the equation and found that

$$\text{Predicted distance} = 576 + (-3 * 60) = 396$$

The question we now ask is, How good is this prediction?

Unfortunately, there is no 60-year-old driver in the original data set of 30 drivers, so we cannot calculate the residual. Instead, we use the $s_e$ as a measurement of typical error.

Technology gives $s_e = 51.35$.

So how good is the prediction for the 60-year-old driver? Based on the $s_e$ for this data, we estimate that our prediction of 396 feet is off by ±51 feet.

| | Intro grade(%) | Upper grade(%) | Predictions | Error (Residual) | Error Squared |
|---|---|---|---|---|---|
| Student 1 | 65 | 58 | 59.1 | −1.1 | 1.21 |
| Student 2 | 71 | 63 | 65.4 | −2.4 | 5.76 |
| Student 3 | 72 | 67 | 66.4 | 0.6 | 0.36 |
| Student 4 | 72 | 77 | 66.4 | 10.6 | 112.36 |
| Student 5 | 75 | 63 | 69.6 | −6.6 | 43.56 |
| Student 6 | 83 | 72 | 77.9 | −5.9 | 34.81 |
| Student 7 | 85 | 84 | 80 | 4 | 16 |
| Student 8 | 88 | 83 | 83.2 | −0.2 | 0.04 |
| Student 9 | 94 | 89 | 89.5 | −0.5 | 0.25 |
| Student 10 | 96 | 93 | 91.5 | 1.5 | 2.25 |

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3515

https://assessments.lumenlearning.co...sessments/3516

https://assessments.lumenlearning.co...sessments/3517

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3869

## Let's Summarize

- When we use a regression line to make predictions, there is error in the prediction. We calculate this error as **Observed data value − Predicted value**. A residual is another name for the prediction error.
- We use residual plots to determine whether a linear model is a good summary of the relationship between the explanatory and response variables. In particular, we look for any *unexpected patterns* in the residuals that may suggest the data is not linear in form.
- We have two numeric measures to help us judge how well the regression line models the data.
  - The square of the correlation coefficient, $r^2$, is the proportion of the variation in the response variable that is explained by the least-squares regression line.
  - The standard error of the regression, $s_e$, gives a typical prediction error based on all of the data. It roughly measures the average distance of the data from the regression line. In this way, it is similar to the standard deviation, which roughly measures average distance from the mean.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.25: Assessing the Fit of a Line (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.26: Putting It Together- Examining Relationships- Quantitative Data

## Let's Summarize

- We use a *scatterplot* to graph the relationship between two quantitative variables. In a scatterplot, each dot represents an individual. We always plot the explanatory variable on the horizontal x-axis.
- When we explore a relationship between two quantitative variables using a scatterplot, we describe the overall pattern (*direction*, *form*, and *strength*) and deviations from the pattern (*outliers*).
- When the *form of a relationship is linear*, we use the correlation coefficient, *r*, to measure the strength and direction of the linear relationship. The correlation ranges between −1 an 1. If the pattern is linear, an *r*-value near −1 indicates a strong negative linear relationship and an *r*-value near +1 indicates a strong positive linear relationship. Following are some cautions about interpreting correlation:
  - **Always make a scatterplot before interpreting *r*.** Correlation is affected by outliers and should be used only when the pattern in the data is linear.
  - **Association does not imply causation.** Do not interpret a high correlation between explanatory and response variables as a cause-and-effect relationship.
  - **Beware of lurking variables** that may be explaining the relationship seen in the data.
- The line that best summarizes a linear relationship is the *least-squares regression line*. The least-squares line is the best fit for the data because it gives the best predictions with the least amount of error. The most common measurement of overall error is the sum of the squares of the errors, SSE. The least-squares line is the line with the smallest SSE.
- We use the least-squares regression line to predict the value of the response variable from a value of the explanatory variable. **Avoid making predictions outside the range of the data.** (This is called *extrapolation*.)
- We have two methods for *finding the equation of the least-squares regression line:* Predicted $y = a + b * x$
  - We use technology to find the equation of the least-squares regression line: Predicted $y = a + b * x$
  - We use summary statistics for $x$ and $y$ and the correlation. Using this method, we can calculate the slope $b$ and the $y$-intercept $a$ using the following: b=\left(r\cdot{s}_{y}\right)/{s}_{x},\text{}a=\stackrel{¯}{y}-b\stackrel{¯}{x}
- The *slope of the least-squares regression* line is the average change in the predicted values when the explanatory variable increases by 1 unit.
- When we use a regression line to make predictions, there is error in the prediction. We calculate this error as Observed value − Predicted value. This prediction error is also called a *residual*.
- We use *residual plots* to determine whether a linear model is a good summary of the relationship between the explanatory and response variables. In particular, we look for any unexpected patterns in the residuals that may suggest that the data is not linear in form.
- We have two numeric measures to help us judge how well the regression line models the data:
  - The square of the correlation, $r^2$ , is the proportion of the variation in the response variable that is explained by the least-squares regression line.
  - The standard error of the regression, $s_e$ , gives a typical prediction error based on all of the data. It roughly measures the average distance of the data from the regression line. In this way, it is similar to the standard deviation, which roughly measures average distance from the mean.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.26: Putting It Together- Examining Relationships- Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.27: StatTutor- Academic Performance

## You are now ready to do a lab exercise for this module.

One of the first few screens in StatTutor will have a link to download the dataset for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click “Save,” which will allow you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you’re using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you’re using the TI Calculator.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=162

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.27: StatTutor- Academic Performance is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.28: Assignment- Scatterplot

In this exercise we will:

- Learn how to create a scatterplot.
- Use the scatterplot to examine the relationship between two quantitative variables.
- Learn how to create a labeled scatterplot.
- Use the labeled scatterplot to better understand the form of a relationship.

In this activity we explore the relationship between weight and height for 81 adults. We will use height as the explanatory variable. Weight is the response variable.

We will then label the men and women by adding the categorical variable gender to the scatterplot. We will see if separating the groups contributes to our understanding of the form of the relationship between height and weight.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 1:

Describe the relationship between the height and weight of the subjects. To describe the relationship write about the pattern (direction, form, and strength) and any deviations from the pattern (outliers).

So far we have studied the relationship between height and weight for all of the males and females together. It may be interesting to examine whether the relationship between height and weight is different for males and females. To visualize the effect of the third variable, gender, we will indicate in the scatterplot which observations are males and which are females.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 2:

Compare and contrast the relationship between height and weight for males and females. To compare and contrast the relationships by gender write about the pattern (direction, form, and strength) and any deviations from the pattern (outliers) for each group.

Discuss how the patterns for the two groups are similar and how they are different.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.28: Assignment- Scatterplot is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.29: Assignment- Linear Relationships

In this activity we will:

- Learn how to compute the correlation.
- Practice interpreting the value of the correlation.
- See an example of how including an outlier can *increase* the correlation.

Recall the following example: The average gestation period, or time of pregnancy, of an animal is closely related to its longevity—the length of its lifespan. Data on the average gestation period and longevity (in captivity) of 40 different species of animals have been recorded.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

Remember that the correlation is only an appropriate measure of the **linear** relationship between two quantitative variables. First produce a scatterplot to verify that gestation and longevity are nearly linear in their relationship.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

Observe that the relationship between gestation period and longevity is linear and positive. Now we will compute the correlation between gestation period and longevity.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 1:

Report the correlation between gestation and longevity and comment on the strength and direction of the relationship. Interpret your findings in context.

Now return to the scatterplot that you created earlier. Notice that there is an outlier in both longevity (40 years) and gestation (645 days). Note: This outlier corresponds to the longevity and gestation period of the elephant.

What do you think will happen to the correlation if we remove this outlier?

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

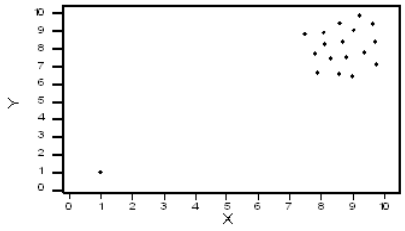
R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 2:

Report the new value for the correlation between gestation and longevity and compare it to the value you found earlier when the outlier was included. What is it about this outlier that results in the fact that its inclusion in the data causes the correlation to increase? (Hint: look at the scatterplot.)

## Comment

In the last activity, we saw an example where there was a positive linear relationship between the two variables, and including the outlier just “strengthened” it. Consider the hypothetical data displayed by the following scatterplot:




In this case, the low outlier gives an “illusion” of a positive linear relationship, whereas in reality, there is no linear relationship between X and Y.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 3.29: Assignment- Linear Relationships is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 3.30: Assignment- Linear Regression

In this activity we will:

- Find a regression line and plot it on the scatterplot.
- Examine the effect of outliers on the regression line.
- Use the regression line to make predictions and evaluate how reliable these predictions are.

## Background

The modern Olympic Games have changed dramatically since their inception in 1896. For example, many commentators have remarked on the change in the quality of athletic performances from year to year. Regression will allow us to investigate the change in winning times for one event—the 1,500 meter race.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

Observe that the form of the relationship between the 1,500 meter race's winning time and the year is linear. The least squares regression line is therefore an appropriate way to summarize the relationship and examine the change in winning times over the course of the last century. We will now find the least squares regression line and plot it on a scatterplot.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 1:

Give the equation for the least squares regression line, and interpret it in context.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 2:

Give the equation for this new line and compare it with the line you found for the whole dataset, commenting on the effect of the outlier.

## Question 3:

Our least squares regression line associates years as an explanatory variable, with times in the 1,500 meter race as the response variable. Use the least squares regression line you found in question 2 to predict the 1,500 meter time in the 2008 Olympic Games in Beijing. Comment on your prediction.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 3.30: Assignment- Linear Regression is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 4: Nonlinear Models


4.1: Why It Matters- Nonlinear Models
4.2: Introduction- Exponential Relationships
4.3: Exponential Relationships (1 of 6)
4.4: Exponential Relationships (2 of 6)
4.5: Exponential Relationships (3 of 6)
4.6: Exponential Relationships (4 of 6)
4.7: Exponential Relationships (5 of 6)
4.8: Exponential Relationships (6 of 6)
4.9: Putting It Together- Nonlinear Models


---


This page titled 4: Nonlinear Models is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.1: Why It Matters- Nonlinear Models

Before we begin *Nonlinear Models*, let's see how the new ideas in this module relate to what we learned in the previous modules, *Types of Statistical Studies and Producing Data*, *Summarizing Data Graphically and Numerically*, and *Examining Relationships: Quantitative Data*.

Recall the Big Picture:

We begin a statistical investigation with a research question. The investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data. ← **Types of Statistical Studies and Producing Data**
- Explore the Data: Analyze and summarize the data. ← **Summarizing Data Graphically and Numerically, Examining Relationships: Quantitative Data, Nonlinear Models**
- Draw a Conclusion: Use the data, probability and statistical inference to draw a conclusion about the population.

*Types of Statistical Studies and Producing Data* focused on methods for collecting reliable data. *Summarizing Data Graphically and Numerically* focused on summarizing and analyzing data for a quantitative variable. *Examining Relationships: Quantitative Data* focused on linear relationships between two quantitative variables. In *Nonlinear Models*, we focus on nonlinear relationships between two quantitative variables. In the Big Picture of Statistics, the material in this module is still part of exploratory data analysis.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.1: Why It Matters- Nonlinear Models is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.2: Introduction- Exponential Relationships

## What you'll learn to do: Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

### LEARNING OBJECTIVES

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.2: Introduction- Exponential Relationships is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.3: Exponential Relationships (1 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

In our first example of exponential relationships, we investigate a nonlinear model for growth in a population over time.

## Example

### The Return of the Bald Eagle

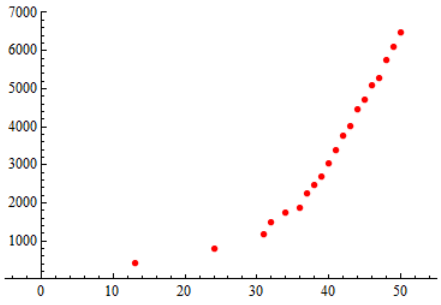
During the mid-20th century, the population of bald eagles in the lower 48 states declined substantially. A highly toxic pesticide, DDT, was the main cause of the decline. DDT causes damage to bird egg shells. By 1963, bald eagles were in danger of complete extinction. Only 417 pairs of bald eagles remained. In 1967, the bald eagle became an official endangered species. Then in 1972, the EPA banned the use of DDT in the United States. The impact of the ban was a dramatic turnaround in the fate of the bald eagle.

Here is the data. Note that in the table, we defined *t*, our explanatory variable, to be *Years after 1950*. The response variable is the number of bald eagle pairs that are mating.

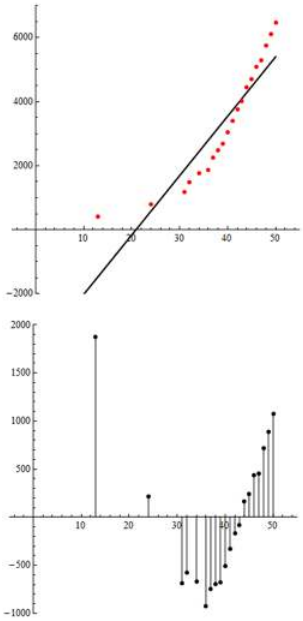
| Year | t = years after 1950 | Eagle Pairs |
|---|---|---|
| 1963 | 13 | 417 |
| 1974 | 24 | 791 |
| 1981 | 31 | 1188 |
| 1982 | 32 | 1480 |
| 1984 | 34 | 1757 |
| 1986 | 36 | 1875 |
| 1987 | 37 | 2238 |
| 1988 | 38 | 2475 |
| 1989 | 39 | 2680 |
| 1990 | 40 | 3035 |
| 1991 | 41 | 3399 |
| 1992 | 42 | 3749 |
| 1993 | 43 | 4015 |
| 1994 | 44 | 4449 |
| 1995 | 45 | 4712 |
| 1996 | 46 | 5094 |
| 1997 | 47 | 5295 |
| 1998 | 48 | 5748 |
| 1999 | 49 | 6104 |
| 2000 | 50 | 6471 |

Our goal is to find an equation to model this relationship.

Here is a scatterplot of the data. We can see that the relationship appears somewhat linear, particularly for years after 1980 ($t = 30$). The correlation coefficient for this data set is high, $r = 0.914$.




The least squares regression line is Predicted eagle pairs = −3,878.11 + 185.4*t*. Below is a scatterplot of the data with the least-squares regression line and the residual plot.

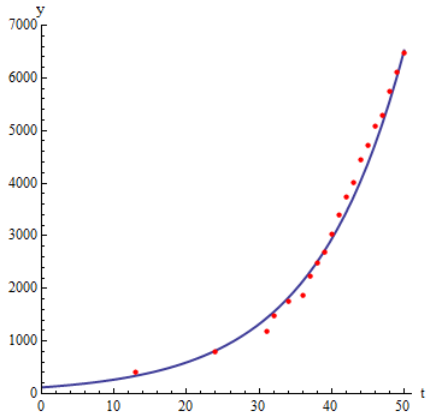



We see a clear pattern in the residuals, suggesting that a linear model does not capture patterns in the data.

Note: This is a reminder that a large *r*-value does not guarantee that a linear model is a good fit.

**Conclusion:** The data set for eagle pairs is clearly nonlinear. We need a better model.

In the scatterplot below, we fit an **exponential** model to the data. Notice how well this model describes the relationship between the variables. There is very little scatter about the exponential curve. There is a strong, positive exponential relationship between these variables.




The **equation of the exponential model** is Predicted eagle pairs = 121 $(1.083)^t$.

Note: In this equation, the *t*-variable is an exponent. Sometimes you will see this written with the caret symbol: ^. So Predicted $y = 121\ (1.083)^t$ and Predicted $y = 121(1.083)\ ^\wedge\ t$ mean the same thing.

Now we use the exponential model to make predictions about the number of bald eagle mating pairs. We also compare the predictions from the exponential model to the linear model. Because there is a strong exponential relationship and a weaker linear relationship in the data, we expect the predictions from the exponential model to be better.

In 1963 ($t = 13$), there were 417 mating pairs.

- According to the linear model: Predicted eagle pairs = –3,878.11 + 185.4 (13) = (–1,468). Obviously, a negative value does not make sense for a count of eagle pairs.
- According to the exponential model: *Predicted eagle pairs* = 121 $(1.083)^{13}$ = 341. So the exponential model underestimates by 417 – 341 = 76 mating pairs. But this is a much better prediction than we got from the linear model.

In 2000 ($t = 50$), there were 6,471 mating pairs.

- According to the linear model: Predicted eagle pairs = –3,878.11 + 185.4 (50) = 5,392. So the linear model underestimates by 6,471 – 5392 = 1,079 mating pairs.
- According to the exponential model: *Predicted eagle pairs* = 121 $(1.083)^{50}$ = 6,519. So the exponential model overestimates by 6,591 – 6,471 = 48 mating pairs. This is a much better prediction than we got from the linear model.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3518

https://assessments.lumenlearning.co...sessments/3519

https://assessments.lumenlearning.co...sessments/3520

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.3: Exponential Relationships (1 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.4: Exponential Relationships (2 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

Now we investigate what the numbers in the exponential model tell us.

## Example

## Understanding the Numbers in the Exponential Model

Our goal in this example is to understand the meaning of the numbers 121 and 1.083 in the exponential model for predicting the number of eagle mating pairs.

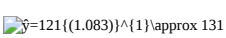
$$\text{Predicted eagle pairs} = 121\,(1.083)^t$$

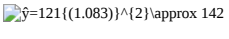
The 121 is the *initial value for the exponential model*. It is the predicted value for y when $t = 0$. It is also the ***y*-intercept**, where the exponential model crosses the y-axis.

- To see this, plug $t = 0$ into the exponential model: Predicted eagle pairs = $121\,(1.083)^0 = 121\,(1) = 121$.
- Interpretation in context: When $t = 0$, the year is 1950. In 1950, the number of eagle mating pairs is predicted to be 121.

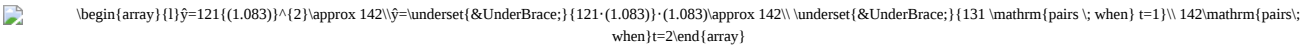
Note: A number with an exponent of 0 is equal to 1. For example, $2^0 = 1$ and $1.5120^0 = 1$. (This is not true for 0: $0^0$ is not defined.)

Now let's investigate the meaning of 1.083 in the context of eagle mating pairs.

For 1951, when $t = 1$, the model predicts  pairs of eagles that are mating.

For 1952, when $t = 2$, there are  pairs.

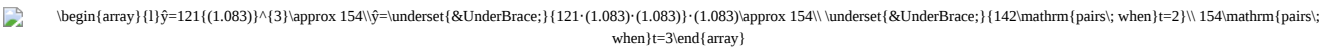
We can also view the calculation for $t = 2$ as repeated multiplication by 1.083:



$$\begin{array}{l}\hat{y}=121(1.083)^2\approx 142\\ \hat{y}=\underset{\text{when } t=1}{\underbrace{121\cdot(1.083)}}\cdot(1.083)\approx 142\end{array}$$

Note: From this viewpoint, we find the estimated number of mating pairs for 1952 by multiplying the estimated 131 pairs from the previous year by 1.083.

Here is another example: For 1953, when $t = 3$, we can rewrite $(1.083)^3$ as repeated multiplication: (1.083)(1.083)(1.083). The exponent 3 tells us to multiply the initial value 121 by 1.083 three times.



$$\begin{array}{l}\hat{y}=121(1.083)^3\approx 154\\ \hat{y}=\underset{\text{when } t=2}{\underbrace{121\cdot(1.083)\cdot(1.083)}}\cdot(1.083)\approx 154\end{array}$$

Note: We can also view this process as multiplying the estimated 142 eagle pairs from the previous year by 1.083.

In general, to find the number of mating pairs for the next year, we multiply the previous year's estimate by 1.083. We call this the **growth factor**.

We view the growth factor as containing information about the **percentage increase** in the population over the previous year. To see how this works, let's start with a hypothetical situation in which there is no change in the number of eagle mating pairs from one year to the next. Then we look at different percentages of growth for the first year to build to an understanding of the meaning of 1.083:

**No change in the number of eagle pairs:** If there is no change in a year, we have 100% of the mating pairs from the previous year. The growth factor is 1.00, which is 100% written in decimal form. This is important to understand. A growth factor of 1.00 means no growth. This makes sense because 121(1.00) = 121; there is no change when we multiply by 1.00.

**5% growth** in the first year:

- 100% of the mating pairs + **5% increase** in mating pairs = 105%.
- Convert to decimal form to find the growth factor: 105% = 1.05.
- Now multiply the growth factor by 121 to find the number of mating pairs for the next year: 121**(1.05)** = 127.

**If we multiply by 1.05, this is a 5% increase.**

**6.8% growth** in the first year:

- 100% of the mating pairs + **6.8% increase** in mating pairs = 106.8%.
- Convert to decimal form to find the growth factor: 106.8% = 1.068.
- Now multiply the growth factor by 121 to find the number of mating pairs for the next year: 121**(1.068)** = 129. **If we multiply by 1.068, this is a 6.8% increase.**

What is the meaning of 1.083 in the model Predicted eagle pairs = 121 $(1.083)^t$?

**Answer:** The 1.083 is the growth factor; as a percentage, it is 108.3%. We view 108.3% as 100% + 8.3%. There is an *estimated* 8.3% increase in the number of eagle pairs each year. (Remember, the 100% represents no change in the population.)

### Spotlight on Converting Percentages

Percent means "per 100." So a percent means "divide by 100."

To convert from a percent to the decimal form, divide by 100.

- For example, 105% means 105 ÷ 100 = 1.05, so 105% = 1.05
- 106.8% means 106.8 ÷ 100 = 1.068, so 106.8% = 1.068
- 96.8% means 96.8 ÷ 100 = .968, so 98.6% = 0.968
- Notice that dividing by 100 moves the decimal two places to the left.

To convert from a decimal form to a percent, we are converting in the opposite direction, so multiply by 100.

- To convert 1.03 to a percent: 1.03 x 100 = 103, so 1.03 = 103%
- To convert 1.083 to a percent: 1.083 x 100 = 108.3, so 1.083 = 108.3%
- To convert 0.834 to a percent: 0.834 x 100 = 83.4, so 0.834 = 83.4%
- Notice that multiplying by 100 moves the decimal two places to the right.

Note: A number is in "decimal form" when it is not a percentage. A percentage like 108.3% is not in "decimal form" even though it has a decimal in the number.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3521

https://assessments.lumenlearning.co...sessments/3522

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 4.4: Exponential Relationships (2 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.5: Exponential Relationships (3 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

Let's summarize what we have learned about exponential growth models:

The general form of an exponential growth model is $y = C \cdot b^x$.

- ***C* is the initial value**. It is the *y*-value when $x = 0$. It is also the *y*-intercept.
- ***b* is the growth factor; it will always be greater than 1 in cases of growth.** From the growth factor, we can determine the percentage increase in *y* for each additional 1 unit increase in *x*.

Let's compare the general form of an exponential growth model to the general form for a *linear model:* $y = a + bx$.

- In the linear model, *a* is the **initial value**. It is the *y*-value when $x = 0$. It is also the *y*-intercept.
- *b* is the **slope**. From the slope, we can determine the amount and direction the *y*-value changes for each additional 1 unit increase in *x*.

Now we apply what we have learned about exponential growth to find a model for a set of data.

In this activity, you use a simulation to find an exponential model that fits the population growth of Kenya.

Here are the data graphed in the scatterplot in the simulation.

| Year | t = years after 1950 | Kenya population (millions) |
|---|---|---|
| 1950 | 0 | 6.4 |
| 1960 | 10 | 8.2 |
| 1970 | 20 | 11.3 |
| 1975 | 25 | 13.5 |
| 1980 | 30 | 16.3 |
| 1985 | 35 | 19.7 |
| 1990 | 40 | 23.4 |
| 1995 | 45 | 27.5 |
| 2000 | 50 | 31.4 |
| 2005 | 55 | 35.8 |
| 2010 | 60 | 40.9 |

Notice that the Kenyan population growth has a strong positive exponential form. Use the sliders in the simulation to adjust the values of *C* and *b* to find a reasonable exponential model that fits this data.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=180

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3871

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.5: Exponential Relationships (3 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.6: Exponential Relationships (4 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

We now investigate an exponential model for **decline** in a population over time.

## Example

### Winter-run Chinook Salmon on the Sacramento River

The U.S. Endangered Species Act lists nine populations of Chinook salmon as either threatened or endangered. One such population is the winter-run Chinook salmon of the Sacramento River in northern California. The Chinook was first listed as endangered in 1994.

A number of factors contributed to the decline of the Chinook on the Sacramento River. Chief among these was the construction of the Red Bluff Diversion Dam in 1967. The dam deprived a large number of adult salmon access to necessary coldwater spawning grounds.

Researchers collected data on the Chinook population at the Red Bluff Dam. This data suggests that the population declined from about 40,000 in 1970 to below 200 in 1994. Click here to see the data.

Here is a scatterplot of the data for the years 1970 through 1994. The scatterplot has a nonlinear form with a negative association between the variables. In other words, we see that the population is declining.

Note: We defined *t*, our explanatory variable, to be *Number of years after 1970*. The response variable is the *Number of Chinook salmon present in the Sacramento River*.

Our goal is to model the decline of the Sacramento River Chinook population.

In the scatterplot below, we fit an exponential model to the data for the years 1970 through 1994. Notice that this model summarizes the pattern in the data, but the relationship is not as strong as we saw in the eagle data. There is more scatter about the exponential curve.

The equation of the exponential model is Predicted Chinook population = $49{,}304\ (0.854)^t$.

Now we use the exponential model to make predictions about the Chinook population. We expect fairly large prediction errors because the association is not strong.

In 1994 ($t = 24$), there were 186 Chinook salmon in the Sacramento River, according to the data.

According to the model: Predicted Chinook population = $49{,}304\ (0.854)^{24} = 1{,}117$.

The exponential model overestimates the number of Chinook salmon by $1{,}117 - 186 = 931$ for that year. We can tell from the graph that this prediction error is small relative to the prediction error for most of the other years. (The data point for 1994, when $t = 24$, is much closer to the curve than are other data points.)

In 2003 ($t = 33$), there were 9,757 Chinook, according to the data.

According to the model: Predicted Chinook population = $49{,}304\ (0.854)^{33} = 270$.

The exponential model underestimates the number of Chinook by $9{,}757 - 270 = 9{,}487$. This is a huge prediction error. But wait – this is an example of extrapolation: 2003 falls outside of the range of the data used to find the model. The model gives unreliable estimates for years outside the range of 1975 through 1994. If you look at the data table, you will see that the Chinook population started to increase again after 1994, with large increases in 2002 and 2003. Because the pattern changes after 1994, this exponential model gives unreliable predictions for years after 1994. (This turnaround was the result of the removal of two dams on the Sacramento River and opening the dam gates for 8 months a year at the Red Bluff Diversion Dam to allow for fish migration to winter spawning areas.)

Note: In *Examining Relationships: Quantitative Data*, we investigated techniques for assessing the fit of a linear model, such as residual plots, $r^2$ and $s_e$. We do not formally investigate residuals for exponential functions in this course. We also do not develop formal techniques for assessing the fit of an exponential model.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 4.6: Exponential Relationships (4 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.7: Exponential Relationships (5 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

The exponential model used in the Chinook salmon example showed a decline over the years and a negative association between the variables; we call such a model an **exponential decay model**. (Compare this model to the exponential growth model we investigated earlier with the eagle pairing data.)

Now we investigate the meaning of the numbers in the exponential decay model.

## Example

## Understanding the Numbers in the Exponential Decay Model

Our goal in this example is to understand the meaning of the numbers 49,304 and 0.854 in the exponential model for predicting the Chinook population.

$$\text{Predicted Chinook population} = 49{,}304\ (0.854)^t$$

As before, the 49,304 is the initial value for the exponential model. It is the predicted value for $y$ when $t = 0$.

- To see this, plug $t = 0$ into the exponential model:

  $\text{Predicted Chinook population} = 49{,}304\ (0.854)^t$

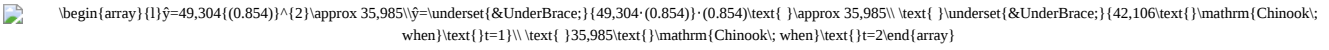
- Interpretation in context: When $t = 0$, the year is 1970. In 1970, the predicted number of Chinook is 49,304.
- It is also the $y$-intercept where the exponential model crosses the y-axis.

Now let's investigate the meaning of 0.854 in the context of Chinook population.

For 1971, when $t = 1$, the model predicts $y = 49{,}304\ (0.854)^1 = 42{,}106$ Chinook salmon in the Sacramento River.

For 1972, when $t = 2$, the model predicts $y = 49{,}304\ (0.854)^2 = 35{,}958$ Chinook.

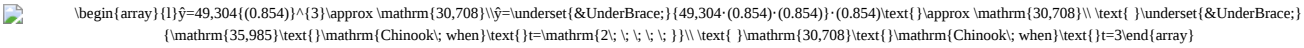
We can also view the calculation for $t = 2$ as repeated multiplication by 0.854:



$$\begin{array}{l}\hat{y}=49{,}304{(0.854)}^{2}\approx 35{,}985\\ \hat{y}=\underset{\text{when } t=1}{\underbrace{49{,}304\cdot(0.854)}}\cdot(0.854)\approx \underset{\text{when } t=2}{\underbrace{35{,}985}}\end{array}$$

Note: From this viewpoint, we find the Chinook population for 1972 by multiplying the 42,106 Chinook from the previous year by 0.854.

Here is another example: For 1973, when $t = 3$, we can rewrite $(0.854)^3$ as repeated multiplication: (0.854)(0.854)(0.854). The exponent 3 tells us to multiply the initial value 49,304 by 0.854 three times.



$$\begin{array}{l}\hat{y}=49{,}304{(0.854)}^{3}\approx 30{,}708\\ \hat{y}=\underset{\text{when } t=3}{\underbrace{49{,}304\cdot(0.854)\cdot(0.854)\cdot(0.854)}}\approx \underset{\text{when } t=3}{\underbrace{30{,}708}}\end{array}$$

Note: We can also view this process as multiplying the estimated 35,985 Chinook from the previous year by 0.854.

In general, to find the estimated number of Chinook for the next year, we multiply the previous year's estimated population by 0.854. We call this the **decay factor**.

We view the decay factor as containing information about the **percentage decrease** in the population over the previous year. To see how this works, let's start with a hypothetical situation in which there is no change in the number of Chinook from one year to the next. Then we look at different percentages of decay for the first year to build to an understanding of the meaning of 0.854. This is the same type of thinking we performed to analyze the exponential growth model previously.

**No change in the number of Chinook:** If there is no change in a year, we have 100% of the fish from the previous year, so the decay factor is 1.00, which is 100% written in decimal form. As before, this is important to understand. A decay factor of 1.00 means no decline in the population. This makes sense because 49,304 (1.00) = 49,304; there is no change when we multiply by 1.00.

**5% decay** in the first year:

- 100% of the Chinook **– 5% decrease** in the Chinook = 95% remaining.
- Convert to decimal form to find the decay factor: 95% = 0.95.
- Now multiply the decay factor by 49,304 to find the number of Chinook for the next year: 49,304(0.95) = 46,839
- So, if we multiply by 0.95, this is a *5% decrease*.

**6.8% decay** in the first year:

- 100% of the Chinook **– 6.8% decrease** in the Chinook = 93.2% remaining.
- Convert to decimal form to find the decay factor: 93.2% = 0.932.
- Now multiply the decay factor by 49,304 to find the number of Chinook for the next year: 49,304 (0.932) = 45,951
- So, if we multiply by 0.932, this is a *6.8% decrease*.

What is the meaning of 0.854 in the model Predicted Chinook population = 49,304 $(0.854)^t$?

**Answer:** The 0.854 is the decay factor; as a percentage, it is 85.4%. This tells us that 85.4% of the previous year's Chinook population are here this year. To determine the percent decrease, calculate 100% – 85.4% = 14.6%. There is an estimated 14.6% decrease in the number of Chinook each year. (Remember the 100% represents no change in the population.)

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3523

https://assessments.lumenlearning.co...sessments/3524

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3525

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3526

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 4.7: Exponential Relationships (5 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.8: Exponential Relationships (6 of 6)

## Learning Objectives

- Use an exponential model (when appropriate) to describe the relationship between two quantitative variables. Interpret the model in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.8: Exponential Relationships (6 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 4.9: Putting It Together- Nonlinear Models

## Let's Summarize

This is what we have learned about exponential models:

The general form of an *exponential model* is $y = C \cdot b^x$.

- Exponential models are nonlinear. More specifically, exponential models predict that $y$ increases or decreases by a constant percentage for each 1-unit increase in $x$.
- $C$ is the *initial value*. It is the $y$-value when $x = 0$. It is also the $y$-intercept.
- $b$ is the *growth factor* or *decay factor*. $b$ is always positive.
  - If $b$ is greater than 1, $b$ is a growth factor. In this case, the association is positive, and $y$ is increasing. This makes sense because multiplying by a number greater than 1 increases the initial value. From the growth factor, we can determine the percent increase in $y$ for each additional 1-unit increase in $x$.
  - Similarly, if $b$ is greater than 0 and less than 1, $b$ is a decay factor. In this case, the association is negative, and $y$ is decreasing. From the decay factor, we can determine the *percentage decrease* in $y$ for each additional 1-unit increase in $x$.

Let's compare the general form of an exponential model to the general form for a *linear model:* $y = a + bx$.

- In the linear model, $a$ is the *initial value*. It is the $y$-value when $x = 0$. It is also the $y$-intercept.
- $b$ is the *slope*. From the slope, we can determine the *amount* and *direction* the $y$-value changes for each additional 1-unit increase in $x$. When $b$ is positive, there is a positive association, and $y$ increases. When $b$ is negative, there is a negative association, and $y$ decreases.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 4.9: Putting It Together- Nonlinear Models is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 5: Relationships in Categorical Data with Intro to Probability


5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability

5.2: Introduction- Two-Way Tables

5.3: Two-Way Tables (1 of 5)

5.4: Two-Way Tables (2 of 5)

5.5: Two-Way Tables (3 of 5)

5.6: Two-Way Tables (4 of 5)

5.7: Two-Way Tables (5 of 5)

5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

5.9: StatTutor- Treating Depression- A Randomized Clinical Trial


---


This page titled 5: Relationships in Categorical Data with Intro to Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability

Before we begin *Relationships in Categorical Data with Intro to Probability*, it is helpful to consider how it relates to the work we have already done in previous modules.

At the start of *Summarizing Data Graphically and Numerically*, we stated the difference between quantitative and categorical variables:

- **Quantitative variables** have *numeric* values that can be averaged. A quantitative variable is frequently a measurement – for example, a person's height in inches.
- **Categorical variables** are variables that can have one of a limited number of values, or labels. Values that can be represented by categorical variables include, for example, a person's eye color, gender, or home state; a vehicle's body style (sedan, SUV, minivan, etc.); a dog's breed (bulldog, greyhound, beagle, etc.).

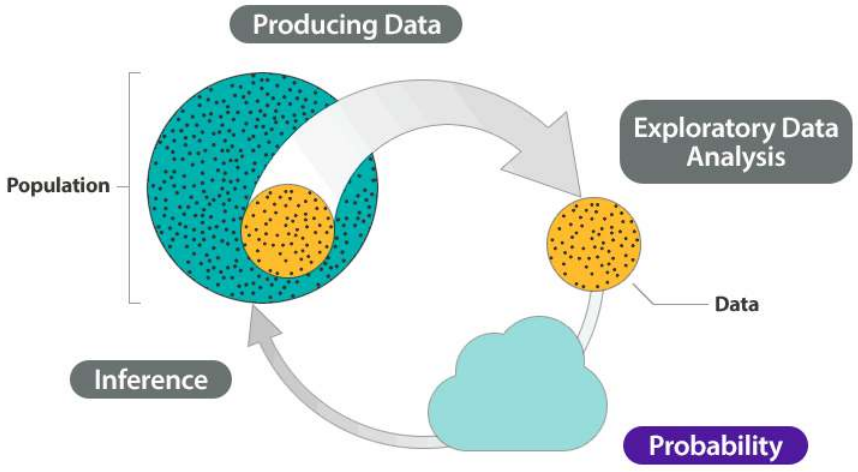
The remainder of *Summarizing Data Graphically and Numerically* focused on describing the overall pattern (shape, center, and spread) of the distribution of a quantitative variable.

In and *Examining Relationships: Quantitative Data* and *Nonlinear Models*, our goal was to identify and model the relationship between *two quantitative variables.*

Now, in this module, we turn our full attention back to categorical variables. Our objective is to study the relationship between two categorical variables. Just as in *Examining Relationships: Quantitative Data* and *Nonlinear Models*, we will be looking for patterns in the data.

As we organize and analyze data from two categorical variables, we make extensive use of **two-way tables**. Two-way tables for two categorical variables are in some ways like scatterplots for two quantitative variables: they give us a useful snapshot of all of the data organized in terms of the two variables of interest. This will be helpful in finding and comparing patterns. This part of *Relationships in Categorical Data with Intro to Probability* is exploratory data analysis in the Big Picture of Statistics.

A second important objective of this module is to introduce you to the concept of **probability**. Two-way tables give us a practical context for talking about probability. We also use two-way tables to help us visualize and solve real-world problems involving probability. This part of the module is part of probability in the Big Picture of Statistics.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.2: Introduction- Two-Way Tables

## What you'll learn to do: Analyze the relationship between two categorical variables using a two-way table.

### LEARNING OBJECTIVES

- Analyze the relationship between two categorical variables using a two-way table.
- Calculate marginal, joint, and conditional percentages and interpret them as probability estimates.
- Analyze and compare risks using conditional probabilities.
- Create a hypothetical two-way table to answer more complex probability questions.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.2: Introduction- Two-Way Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.3: Two-Way Tables (1 of 5)

## Learning Objectives

- Analyze the distribution of a categorical variable.
- Analyze the relationship between two categorical variables using a two-way table.

We begin our discussion by analyzing the distribution of a single categorical variable. Then we focus on analyzing the association between two categorical variables.

## Example

## Body Image

What is your perception of your own body? Do you feel that you are overweight, underweight, or about right? A random sample of 1,200 U.S. college students answered this question as part of a larger survey. The following table shows part of the responses:

| **Student** | **Body Image** |
|---|---|
| student 25 | overweight |
| student 26 | about right |
| student 27 | underweight |
| student 28 | about right |
| student 29 | about right |

Here are the questions we investigate:

- What percentage of students in the sample fall into each category?
- How are students divided across the three body image categories?
- Is there a pattern in the responses?
- Which response is the most common?

It is difficult to answer these questions by looking at the raw data because the raw data is a long list of 1,200 responses. We cannot see patterns easily by looking at a list, so we summarize the distribution in a table.

Recall from *Summarizing Data Graphically and Numerically* that in a graph that summarizes the distribution of a *quantitative* variable, we can see

- the possible values of the variable.
- the number of individuals with each variable value or interval of values.

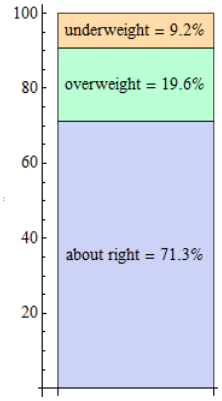
Here we use a table instead of a graph to summarize the distribution of a *categorical* variable. We create a table so we can see

- the different values (categories) the variable takes.
- how many times each value occurs (count) and, more important, how often each value occurs (by converting the counts to proportions).

Here is the table for our example:

| **Category** | **Count** | **Proportion** | **Percentage** |
|---|---|---|---|
| underweight | 110 | 110/1,200 = 0.092 | 9.2% |
| overweight | 235 | 235/1,200 = 0.196 | 19.6% |
| about right | 855 | 855/1,200 = 0.713 | 71.3% |

We can use a stacked bar chart to display the distribution of the body image variable. Note that this distribution is completely described by the three percentages 9.2%, 19.6%, and 71.3%, which correspond to the three categories of the body image variable: “underweight,” “overweight,” and “about right.” The percentages add to 100% because all 1,200 individual responses fall into one of these three categories. (Note that the percentages actually add up to 99.9% because we rounded percentages to three decimal places.)




Now that we have summarized the distribution of values in the body image variable, let’s go back and interpret the results in the context of the questions we posed.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3956

https://assessments.lumenlearning.co...sessments/3957

### Example

## Two-Way Table for Body Image and Gender

Once we’ve interpreted the results, another interesting question arises: *If we separate our sample by gender and compare the male and female responses, will we find a similar distribution across body image categories? Or is there a difference based on gender?*

Answering these questions requires us to examine the relationship between two categorical variables: gender and body image. We want to determine if gender explains the differences in body image responses. Therefore,

- the *explanatory* variable is gender, and
- the *response* variable is body image.

Here is part of the raw data for body image and gender of each student:

| Student | Gender | Body Image |
|---|---|---|
| student 25 | M | overweight |
| student 26 | M | about right |
| student 27 | F | underweight |
| student 28 | F | about right |
| student 29 | M | about right |

Once again, the raw data is a long list of 1,200 responses. We need to organize the information in a table so we can more easily compare the results for females and males. To summarize the relationship between two categorical variables, we create a display called a **two-way table**.

Here is the two-way table for our example:

| | About Right | Overweight | Underweight | Row Totals |
|---|---|---|---|---|
| **Female** | 560 | 163 | 37 | 760 |
| **Male** | 295 | 72 | 73 | 440 |
| **Column Totals** | 855 | 235 | 110 | 1,200 |

Let's take a closer look at this table:

The table helps us to compare females to males because there is a row for each gender. The body image categories are the columns. As we move across a particular *row*, all of the individuals are of the *same gender*. And as we move down a particular *column*, all of the individuals have the *same body image*.

We also added a row at the bottom and a column at the right, which we call the **margins** of the table. The numbers in the margins are totals for each row or column.

In the following table, look at the numbers in the Female row and note that their sum, **560 + 163 + 37 = 760**, is displayed in the margin at the right labeled Row Totals. There are 760 females in the sample.

| | About Right | Overweight | Underweight | Row Totals |
|---|---|---|---|---|
| **Female** | **560** | **163** | **37** | **760** |
| **Male** | 295 | 72 | 73 | 440 |
| **Column Totals** | 855 | 235 | 110 | 1,200 |

Likewise, in the next table, look at the numbers in the Overweight column and note that their sum, **163 + 72 = 235**, is displayed in the margin at the bottom of the table labeled Column Totals. There are 235 students in the sample who answered "overweight" to the body image question.

| | About Right | Overweight | Underweight | Row Totals |
|---|---|---|---|---|
| **Female** | 560 | **163** | 37 | 760 |
| **Male** | 295 | **72** | 73 | 440 |
| **Column Totals** | 855 | **235** | 110 | 1,200 |

Where a row and column cross, we see the number of individuals who fit both descriptions: a particular gender and a particular body image. It may be helpful to think of the six inner cells as six rooms filled with the 1,200 students from the sample. For example, in one room are the 72 males who think of themselves as overweight. In another room, we have 37 females who think of themselves as underweight. (Maybe they should have a potluck and get to know each other.)

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3527

https://assessments.lumenlearning.co...sessments/3873

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3528

https://assessments.lumenlearning.co...sessments/3529

So far we have organized the raw data in a much more informative display – the two-way table. But we have not answered our primary question: Is body image related to gender?

Exploring the relationship between two categorical variables (in this case, body image and gender) amounts to *comparing the distributions of the response variable* (in this case, body image) *for different values of the explanatory variable* (in this case, male vs. female).

We do this in the next example.

### Example

## Is Body Image Related to Gender?

Here we have removed the column totals from the table because gender is the explanatory variable. We compare females with particular body image responses to males with the same response, so we need to know the total numbers of females and males. We no longer need to know the total number of students for each body image category.

Compare these distributions! ⇒

| | about right | overweight | underweight | Row Totals |
|---|---|---|---|---|
| female | 560 | 163 | 37 | 760 |
| male | 295 | 72 | 73 | 440 |

Note that there are more females than males, so when we compare females to males, it is misleading to compare raw counts in each body image category. For example, it is misleading to say, “Five-hundred sixty females responded ‘about right’ compared to only 295 males,” because the sample includes a lot more females than males. Instead, we compare the percentage of females who responded “about right” to the percentage of males who responded “about right”:

- Of the 760 females, 560 responded “about right”: 560 ÷ 760 = 0.737 = 73.7%
- Of the 440 males, 295 responded “about right”: 295 ÷ 440 = 0.67 = 67%

We can interpret percentages as “a number out of 100,” so by converting to percentages, we are reporting the results as though there are 100 females and 100 males. We can see that a higher percentage of females feel “about right” about their body weight.

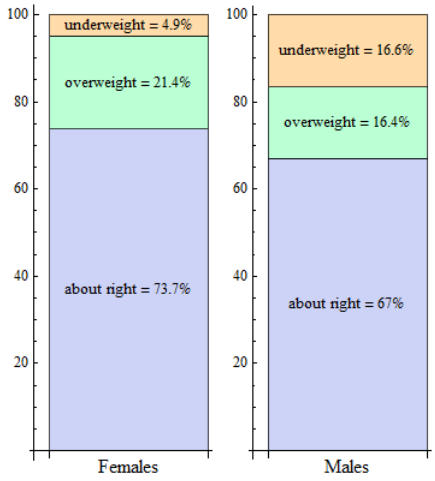
In general, we need to supplement our display, the two-way table, with numeric summaries that allow us to compare the distributions. Therefore, we always convert counts to percentages.

Note: It is important to identify the *explanatory* variable because we always use the totals for the explanatory variable to calculate the percentages.

In our example, we look at each gender separately and convert the counts to percentages within each gender. In the Female row, we divide each count by **760**, the total number of females. In the Male row, we divide each count by **440**, the total number of males. The resulting percentages are shown in the following table: green for females, black for males. We call these **conditional percentages**. The percentages in green are the distribution of body image based on the *condition that students are female*. The percentages in black are the distribution of body image based on the *condition that students are male*. Thus, our two sets of conditional percentages form two *conditional distributions* for body image.

| | **About Right** | **Overweight** | **Underweight** | **Row Totals** |
|---|---|---|---|---|
| **Female** | 560/760 = **73.7%** | 163/760 = **21.4%** | 37/760 =**4.9%** | 760/760 = **100%** |
| **Male** | 295/440 = **67%** | 72/440 = **16.4%** | 73/440 = **16.6%** | 440/440 = **100%** |

Here is a side-by-side display comparing the conditional body image distributions for females and males.




Now that we summarized the relationship between the categorical variables gender and body image, we use the next activity to interpret the results in the context of the questions we posed.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3530

https://assessments.lumenlearning.co...sessments/3531

https://assessments.lumenlearning.co...sessments/3874

At the start of this example, we asked the following questions:

*If we separate our sample by gender and compare the male and female responses, will we find a similar distribution across body image categories? Or is there a difference based on gender?*

As a result of our analysis, we know that the conditional distributions for males and females for body image are not the same. And there is enough of a difference to believe that these two categorical variables are in fact related.

In the next activity, we practice investigating the relationship between two different categorical variables.

We investigate this question in the next activity: *Is there a relationship between smoking rates and college programs?* Researchers sent an online health behavior survey to 25,000 college students in 2009. The following table summarizes results based on 6,055 student responses. (C. J. Berg, C. M. Klatt, J. L. Thomas, J. S. Ahluwalia, and L. C. An, "The Relationship of Field of Study to Current Smoking Status among College Students," *College Student Journal* 43(3):744–754, 2009.)

| | Smoked in Last 30 Days | Did Not Smoke in Last 30 Days | |
|---|---|---|---|
| Art, design, performing arts | 149 | 336 | 485 |
| Humanities | 197 | 454 | 651 |
| Communication, languages | 233 | 389 | 622 |
| Education | 56 | 170 | 226 |
| Health Sciences | 227 | 717 | 944 |
| Math, engineering, sciences | 245 | 924 | 1,169 |
| Social science, human services | 306 | 593 | 899 |
| Independent study | 134 | 260 | 394 |
| Undeclared | 176 | 489 | 665 |
| | 1,723 | 4,332 | 6,055 |

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3532

https://assessments.lumenlearning.co...sessments/3533

https://assessments.lumenlearning.co...sessments/3534

https://assessments.lumenlearning.co...sessments/3535

In the next activity, we investigate whether health insurance coverage differs by geographic region. The U.S. government collects information on Americans who do not have health insurance. Here is the data:

| Region | Uninsured | Insured | Row Totals |
|---|---|---|---|
| **Northeast** | 6,782 | 47,043 | 53,825 |
| **Midwest** | 7,757 | 57,135 | 64,892 |
| **South** | 19,090 | 85,800 | 104,890 |
| **West** | 11,676 | 55,427 | 67,103 |
| **Column Totals** | 45,305 | 245,405 | 290,710 |

## Let's Summarize

The relationship between two categorical variables is summarized using

- Data display: Two-way table, supplemented by
- Numeric summaries: Conditional percentages.

Conditional percentages are calculated separately for each value of the explanatory variable. When we try to understand the relationship between two categorical variables, we compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.3: Two-Way Tables (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.4: Two-Way Tables (2 of 5)

## Learning Objectives

- Calculate marginal, joint, and conditional percentages and interpret them as probability estimates.

In the previous section, we used the information in a two-table to examine the relationship between two categorical variables. Our goal was to answer the big question: *Are the variables related?*

In this section, we continue to work with two-way tables, but we ask a different set of questions.

## Example

## Community College Enrollment

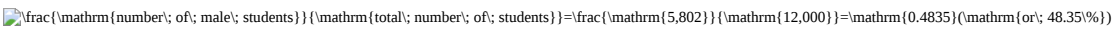
The following table summarizes the full-time enrollment at a community college located in a West Coast city. There are a total of 12,000 full-time students enrolled at the college. The two categorical variables here are *gender* and *program*. The programs include academic and vocational programs at the college. Assume that a student can enroll in only one program.

| | **Arts-Sci** | **Bus-Econ** | **Info Tech** | **Health Science** | **Graphics Design** | **Culinary Arts** | **Row Totals** |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

Let's consider a few preliminary questions to get familiar with this new data set.

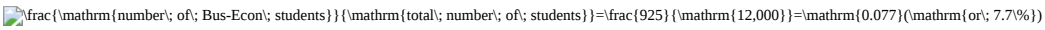
1. *What proportion of the total number of students are male students?*

**Answer**



2. *What proportion of the total number of students are Bus-Econ students?*

**Answer**



Note that to calculate this proportion, we used two numbers in the margin that relate to just one of the categorical variables (program). This calculation is therefore called a **marginal proportion**.

Note: This proportion does not help us determine if gender is related to program because it involves only one of the variables.

Now consider the following question:

> *If we choose one student at random from among all 12,000 students at the college, how likely is it that this student will be in the Bus-Econ program?*

From our previous calculation, we know that only about 8% (7.7%) of the students at the college are in the Bus-Econ program. That's a fairly low number, so it is not very likely that our random student will be a Bus-Econ student.

One way to state our conclusion is to say:

> There is about an **8% chance** of picking a Bus-Econ major.

This means that if we selected 100 students at random, we would expect on average that 8 of them would be in the Bus-Econ program.

Here is another way to state this conclusion:

There is about an **0.08 probability** of picking a Bus-Econ major.

Because this probability is exactly the same as the marginal proportion we calculated earlier, we call it a **marginal probability**.

**Note: *P* for Probability**

It is customary to use the capital letter *P* to stand for probability. So instead of writing "The probability that a student is in Bus-Econ program equals 0.08," we can write *P*(student is in Bus-Econ) = 0.08.

The following table is used for the next Learn By Doing and Did I Get This? activities.

| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3536

https://assessments.lumenlearning.co...sessments/3537

### Example

## Conditional Probability

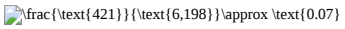
Here is the same community college enrollment data.

| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

Here is our first question:

*If we select a female student at random, what is the probability that she is in the Health Sciences program?*

**Answer** Of the 6,198 female students at the college, **421** are enrolled in Health Sciences. (Find these numbers in the table.) The probability we are looking for is:



Therefore, the probability that a female student is in the Health Sciences program is approximately 0.07.

**Focus on Language**

We need to pause here and be very careful about the language we use in describing this situation.

Note that we *start* with a female student and *then* ask what is the probability that this female student is in the Health Sciences department.

In this case, our *starting point is that the student is a female*. This information sets the conditions for calculating the probability. Once the condition *(student is female)* is set, we focus on the female student population. In terms of the two-way table, it means that the only numbers we will be using are in the Female row: 421 and 6,198.

### What Is a Conditional Probability?

The probability we calculated earlier is an example of a **conditional probability**. In general, a conditional probability is one that is based on a given condition. Here the *given condition* is that the student is female.

Here is the notation we use for a conditional probability:

- Original question: *If we select a female student at random, what is the probability that she is in the Health Sciences program?*
- Notation: *P*(student is in Health Sciences **given that** student is female).
- We also write this as *P*(Health Sciences **given** female).

An even shorter way of writing this is to use a vertical bar | in place of *given:P*(Health Sciences | female).

The following table is used for the next Learn By Doing and Did I Get This? activities.

| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3538

https://assessments.lumenlearning.co...sessments/3539

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.4: Two-Way Tables (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.5: Two-Way Tables (3 of 5)

## Learning Objectives

- Calculate marginal, joint, and conditional percentages and interpret them as probability estimates.

At this point, we know how to determine *marginal probabilities*, such as the probability that a randomly selected student is female: *P*(female).

And we know how to calculate *conditional probabilities*, such as the probability that a randomly selected female student is in the Health Science program: *P*(Health Science | female)

But we do not know how to calculate **joint probabilities**, such as the probability that a randomly selected student is both a female *and* in the Health Sciences program.

We write this joint probability as *P*(female and Health Sciences).

The following example illustrates how to calculate a joint probability.

## Example

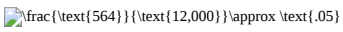
## Joint Probability

**Question:** *If we select a student at random, what is the probability that the student is both a male **and** in the Info Tech program?*

**Answer** This question involves male students who are in the Info Tech program, but it is NOT a conditional probability. We are picking a student at random from the *entire population of 12,000 students*, so there is no condition. Our shorthand notation for this probability is:

*P*(male **and** Info Tech)

Since 564 of the 12,000 students enrolled at the college are both male and in the Info Tech program (see table), the probability *P*(male and Info Tech) is:



| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | **564** | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | **12,000** |

We call this calculation a **joint probability**. Note that when we calculate a joint probability, we divide the count from an inner cell of the table by the overall total count in the lower right corner.

The following table is used for the next Learn By Doing and Did I Get This? activities.

| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

Learn By Doing

https://assessments.lumenlearning.co...sessments/3540

https://assessments.lumenlearning.co...sessments/3541

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 5.5: Two-Way Tables (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.6: Two-Way Tables (4 of 5)

## Learning Objectives

- Analyze and compare risks using conditional probabilities.

When we calculate the probability of a **negative outcome** like a heart attack, we often refer to the probability as a **risk**. For example, we talk about the *probability* of winning the lottery but the *risk* of getting struck by lightning. Whenever you see the word *risk*, keep in mind it's just another word for *probability*.

## Example

## Risk and the Physicians' Health Study

Researchers in the Physicians' Health Study (1989) designed a randomized clinical trial to determine whether aspirin reduces the risk of heart attack. Researchers randomly assigned a large sample of healthy male physicians (22,071) to one of two groups. One group took a low dose of aspirin (325 mg every other day). The other group took a placebo. This was a double-blind experiment. Here are the final results.

| | **Heart Attack** | **No Heart Attack** | **Row Totals** |
|---|---|---|---|
| **Aspirin** | 139 | 10,898 | 11,037 |
| **Placebo** | 239 | 10,795 | 11,034 |
| **Column Totals** | 378 | 21,693 | 22,071 |

Note that the categorical variables in this case are

- *Explanatory variable:* Treatment (aspirin or placebo)
- *Response variable:* Medical outcome (heart attack or no heart attack)

**Question:** *Does aspirin lower the risk of having a heart attack?*

To answer this question, we compare two conditional probabilities:

- The probability of a heart attack given that aspirin was taken every other day.
- The probability of a heart attack given that a placebo was taken every other day.

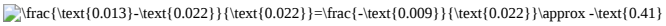
From the table we have

- *P*(heart attack | aspirin) = 139 / 11,037 = 0.013
- *P*(heart attack | placebo) = 239 / 11,034 = 0.022

The result shows that taking aspirin reduced the risk from 0.022 to 0.013.

We often compare two risks by calculating the **percentage change**. We calculate the difference (how much the risk changed) and divide by the risk for the placebo group.

Here is the calculation:

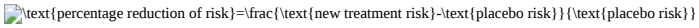


**Therefore, we conclude that taking aspirin results in a 41% reduction in risk.**

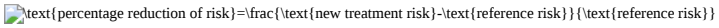
As reported in the *New England Journal of Medicine*, "This trial of aspirin for the primary prevention of cardiovascular disease demonstrates a conclusive reduction in the risk of myocardial infarction (heart attack)." *(Source: "Final Report on the Aspirin Component of the Ongoing Physicians' Health Study," New England Journal of Medicine 321(3):129–35, 1989.)*

**Comment**

In the preceding example, we compared the difference in risk (how much the risk changed) to the risk for the placebo (nontreatment) group:



In general, we are interested in determining how much a new treatment reduces the risk compared to a **reference** risk. The reference may be nontreatment (e.g., use of a placebo), or it could be an existing treatment that we hope to improve on. So we have:



The following table is used for the next Learn By Doing activity.

| | **Nonfatal** | **Fatal** | **Row Totals** |
|---|---|---|---|
| **Seat Belt** | 412,368 | 510 | 412,878 |
| **No Seat Belt** | 162,527 | 1,601 | 164,128 |
| **Column Totals** | 574,895 | 2,111 | 577,006 |

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3542

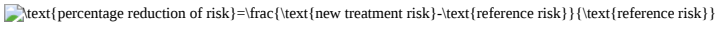
https://assessments.lumenlearning.co...sessments/3543

https://assessments.lumenlearning.co...sessments/3544

Let's summarize our work with probability. We defined three kinds of probabilities related to a two-way table.

- A *marginal probability* is the probability of a categorical variable taking on a particular value *without regard to the other categorical variable*. For example, *P*(Health Sciences) is the probability that a student is enrolled in the Health Sciences program. In calculating the probability, we use overall student data contained in the margins of the table. We do not take into account the other categorical variable: gender.
- A *conditional probability* is the probability of a categorical variable taking on a particular value *given the condition that the other categorical variable has some particular value*. For example, *P*(Health Sciences given female) is the probability that a student is enrolled in Health Sciences given that we know the student is female. In calculating the probability, we use only a subset of the data. The subset used is determined by the given condition: if our condition relates to female students, then we consider only the information in the table pertaining to females.
- A *joint probability* is the probability that the *two categorical variables each take on a specific value*. For example: *P*(male and Info Tech) is the probability that a student is both a male and in the Info Tech program. In calculating this probability, we divide the count in one inner cell of the table by the overall total count (in the lower right corner).

When we calculate the probability of a negative outcome like a heart attack, we often refer to the probability as a *risk*. We compare risk by calculating the percentage change:



CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.6: Two-Way Tables (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.7: Two-Way Tables (5 of 5)

## Learning Objectives

- Create a hypothetical two-way table to answer more complex probability questions.

In our previous work with probability, we computed probabilities using a two-way table of data from a large sample. Now we create a hypothetical two-way table to answer more complex probability questions.

## Example

## Will It Be a Boy or a Girl?

A pregnant woman often opts to have an ultrasound to predict the gender of her baby.

Assume the following facts are known:

- Fact 1: 48% of the babies born are female.
- Fact 2: The proportion of girls correctly identified is 9 out of 10.
- Fact 3: The proportion of boys correctly identified is 3 out of 4.

*(Source: Keeler, Carolyn, and Steinhorst, Kirk. "New Approaches to Learning Probability in the First Statistics Course," Journal of Statistics Education 9(3):1–24, 2001.)*

Here are the questions we want to answer:

**Question 1:** *If the examination predicts a girl, how likely is it that the baby will be a girl?*
**Question 2:** *If the examination predicts a boy, how likely is it that the baby will be a boy?*

Let's consider what the possibilities are.

- The ultrasound examination predicts a girl, and either (a) a girl is born or (b) a boy is born.
- The ultrasound exam predicts a boy, and either (a) a girl is born or (b) a boy is born.

Let's represent these four possible outcomes in a two-way table. On the left we have the categorical variable *prediction*, and on the top the categorical variable *gender of baby*.

| | **Girl** | **Boy** | |
|---|---|---|---|
| **Predict Girl** | | | |
| **Predict Boy** | | | |

Now we find ourselves in an interesting situation. A two-way table without data!

The key idea is to create a two-way table consistent with the stated facts, then use the table to answer our questions.

To get started, let's assume we have ultrasound predictions for 1,000 random babies. We could have picked any number here, but 1,000 will make our calculations easier to keep track of.

Starting with this number, we work backwards with our three facts to fill in this "hypothetical" table.

The first step is to put 1,000 as the overall total in the bottom right corner.

| | **Girl** | **Boy** | **Row Totals** |
|---|---|---|---|
| **Predict Girl** | | | |
| **Predict Boy** | | | |
| **Column Totals** | | | **1,000** |

Let's consider Fact 1: 48% of the babies born are female.

The bottom row gives the distribution of the categorical variable *gender of baby*. We can use this fact to compute the total number of girls and boys.

- 48% girls means that 0.48 (1,000) = 480 are girls.
- 52% are boys. (If 48% are girls, then 100% − 48% = 52% are boys.) So, 0.52(1,000) = 520 boys.

Fill these values into the bottom row of table.

- Note: These are marginal totals.
- You can check your work: These numbers should add to 1,000. If we add all the girls and boys together, we get the total number of babies.

| | **Girl** | **Boy** | **Row Totals** |
|---|---|---|---|
| **Predict Girl** | | | |
| **Predict Boy** | | | |
| **Column Totals** | 0.48(1,000) = **480** | 0.52(1,000) = **520** | 1,000 |

Now let's move on to Fact 2: The proportion of girls correctly identified is 9 out of 10.

- 9 out of 10 is 90% (9 ÷ 10 = 0.90 = 90%).
- 90% of the girls are correctly identified: 0.90(480) = 432.
- 10% of the girls are misidentified (predicted to be a boy): 0.10(480) = 48.

Fill these values into the table.

- You can check your work: These numbers should add to the total number of girls.
- (Girls who are correctly identified as girls ) + (Girls who are misidentified as boys) = Total girls

| | **Girl** | **Boy** | **Row Totals** |
|---|---|---|---|
| **Predict Girl** | 0.90(480)= **432** | | |
| **Predict Boy** | 0.10(480) = **48** | | |
| **Column Totals** | 480 | 520 | 1,000 |

Finally, we use Fact 3: The proportion of boys correctly identified is 3 out of 4.

- 3 out of 4 is 75% (3 ÷ 4 = 0.75 = 75%).
- 75% of the boys are correctly identified: 0.75(520) = 390.
- 25% of the boys are misidentified (predicted to be a girl): 0.25(520) = 130.

Fill these values into the table.

- You can check your work: These numbers should add to the total number of boys.
- (Boys who are correctly identified as boys ) + (Boys who are misidentified as girls) = Total boys

| | **Girl** | **Boy** | **Row Totals** |
|---|---|---|---|
| **Predict Girl** | 432 | 0.25(520) = **130** | |
| **Predict Boy** | 48 | 0.75(520) = **390** | |
| **Column Totals** | 480 | 520 | 1,000 |

Filling in the Row Totals, we now have a complete hypothetical two-way table based on our given information.

| | Girl | Boy | Row Totals |
|---|---|---|---|
| **Predict Girl** | 432 | 130 | 562 |
| **Predict Boy** | 48 | 390 | 438 |
| **Column Totals** | 480 | 520 | 1,000 |

We are now in a position to answer our two questions:

**Question 1:** *If the examination predicts a girl, how likely is it that the baby will be a girl?*

**Answer:** We are asked to find the probability of a girl **given that** the examination predicts a girl.

This is the conditional probability: *P*(girl | predict girl).

So our answer to Question 1 is *P*(girl | predict girl) = 432 / 562 = 0.769.

**Question 2:** *If the examination predicts a boy, how likely is it that the baby will be a boy?*

**Answer:** We are asked to find the probability of a boy **given that** the examination predicts a boy.

This is the conditional probability: *P*(boy | predict boy).

So our answer to Question 2 is *P*(boy | predict boy) = 390 / 438 = 0.890.

**Conclusion:** If an ultrasound examination predicts a girl, the prediction is correct about 77% of the time. In contrast, when the prediction is a boy, it is correct 89% of the time.

**Comment**

Are you surprised at the answers to these questions? Looking just at the three given facts, you might have intuitively expected a different result. This is exactly why a two-way table is so useful. It helps us organize the relevant information in a way that permits us to carry out a logical analysis. When it comes to probability, sometimes our intuition needs some help.

Use the following context for the next Learn By Doing activity.

A large company has instituted a mandatory employee drug screening program. Assume that the drug test used is known to be 99% accurate. That is, if an employee is a drug user, the test will come back positive (“drug detected”) 99% of the time. If an employee is a non-drug user, then the test will come back negative (“no drug detected”) 99% of the time. Assume that 2% of the employees of the company are drug users.

In constructing the hypothetical two-way table, it is convenient to start by assuming that the company has 10,000 employees (10,000 is a large enough number to ensure that all calculations result in whole numbers).

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3545

https://assessments.lumenlearning.co...sessments/3546

https://assessments.lumenlearning.co...sessments/3547

https://assessments.lumenlearning.co...sessments/3548

https://assessments.lumenlearning.co...sessments/3549

https://assessments.lumenlearning.co...sessments/3550

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.7: Two-Way Tables (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

## Let's Summarize

To summarize the relationship between two categorical variables, use:

- A data display: A two-way table
- Numerical summaries: Conditional percentages

When we investigate the relationship between two categorical variables, we use the values of the explanatory variable to define the comparison groups. We then compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

> For example, we investigated the relationship between body image and gender. We compared males to females. For each gender, we determined the percentage who felt their body weight was about right, overweight, or underweight. *P*(body image "about right" | male) is compared to *P*(body image "about right" | female).

**Keys Ideas from Our Work with Probability**

We defined three kinds of probabilities related to a two-way table:

- A **marginal probability** is the probability of a categorical variable taking on a particular value *without regard to the other categorical variable*. For example, *P*(Health Sciences) is the probability that a student is enrolled in the Health Sciences program. In calculating the probability, we use overall student data contained in the margins of the table. A marginal probability is a row or column total divided by the table total.
- A **conditional probability** is the probability of a categorical variable taking on a particular value *given the condition that the other categorical variable has some particular value*. For example, *P*(Health Sciences **given** female) means we look first at all females, then identify the female students who are Health Science students. In calculating the probability, we use only a subset of the data. The condition determines the subset of data we use. If our condition relates to female students, then we consider only the information in the table pertaining to females.
- A **joint probability** is the probability that the *two categorical variables each take on a specific value*. For example: *P*(male **and** Info Tech) is the probability that a student is both a male and in the Info Tech program. In calculating this probability, we divide the count from one inner cell of the table by the overall total count (in the lower right corner.)

When we calculate the probability of a **negative outcome**, we often refer to the probability as a **risk**. We compare risk by calculating the percentage change (divide difference in risks by risk in placebo group).

Finally, we created hypothetical two-way tables to compute complex probabilities, such as the probability of a positive drug test for someone who does not use drugs.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 5.9: StatTutor- Treating Depression- A Randomized Clinical Trial

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=208

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 5.9: StatTutor- Treating Depression- A Randomized Clinical Trial is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 6: Probability and Probability Distributions

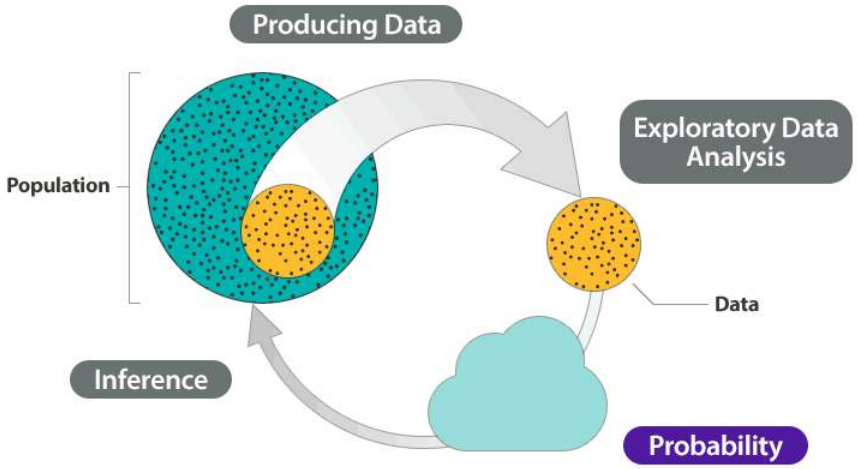

6.1: Why It Matters- Probability and Probability Distributions
6.2: Introduction- Another Look at Probability
6.3: Another Look at Probability (1 of 2)
6.4: Another Look at Probability (2 of 2)
6.5: Introduction- Probability Rules
6.6: Probability Rules (1 of 3)
6.7: Probability Rules (2 of 3)
6.8: Probability Rules (3 of 3)
6.9: Introduction- Discrete Probability Distribution
6.10: Discrete Random Variables (1 of 5)
6.11: Discrete Random Variables (2 of 5)
6.12: Discrete Random Variables (3 of 5)
6.13: Discrete Random Variables (4 of 5)
6.14: Discrete Random Variables (5 of 5)
6.15: Introduction- Continuous Probability Distribution
6.16: Continuous Probability Distribution (1 of 2)
6.17: Continuous Probability Distribution (2 of 2)
6.18: Introduction- Normal Random Variables
6.19: Normal Random Variables (1 of 6)
6.20: Normal Random Variables (2 of 6)
6.21: Normal Random Variables (3 of 6)
6.22: Normal Random Variables (4 of 6)
6.23: Normal Random Variables (5 of 6)
6.24: Normal Random Variables (6 of 6)
6.25: Putting It Together- Probability and Probability Distribution


---


This page titled 6: Probability and Probability Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.1: Why It Matters- Probability and Probability Distributions

Recall the Big Picture—the four-step process that encompasses statistics (as it is presented in this course):




So far, we've discussed the first two steps:

**Producing data**—how data are obtained and what considerations affect the data production process.

**Exploratory data analysis**—tools that help us get a first feel for the data by exposing their features using graphs and numbers.

Our eventual goal is **inference**—drawing reliable conclusions about the population on the basis of what we've discovered in our sample. To really understand how inference works, though, we first need to talk about **probability**, because it is the underlying foundation for the methods of statistical inference. We use an example to try to explain why probability is so essential to inference.

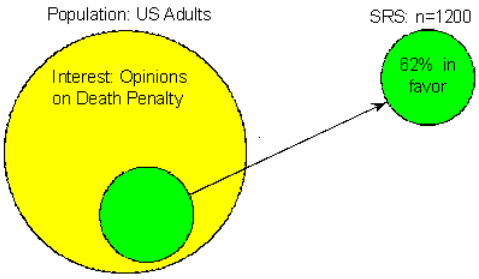
First, here is the general idea: As we all know, the way statistics works is that we use a sample to learn about the population from which it was drawn. Ideally, the sample should be random so that it represents the population well.

Recall from *Types of Statistical Studies and Producing Data* that when we say a random sample represents the population *well*, we mean that there is no inherent bias in this sampling technique. It is important to acknowledge, though, that this does not mean all random samples are necessarily "perfect." Random samples are still random, and therefore no random sample will be exactly the same as another. One random sample may give a fairly accurate representation of the population, whereas another random sample might be "off" purely because of chance. Unfortunately, when looking at a particular sample (which is what happens in practice), we never know how much it differs from the population. This uncertainty is where probability comes into the picture. We use probability to quantify how much we expect random samples to vary. This gives us a way to draw conclusions about the population in the face of the uncertainty that is generated by the use of a random sample. The following example illustrates this important point.

### Example

## Death Penalty

Suppose we are interested in estimating the percentage of U.S. adults who favor the death penalty. To do so, we choose a random sample of 1,200 U.S. adults and ask their opinion: either in favor of or against the death penalty. We find that 744 of the 1,200, or 62%, are in favor. (Although this is only an example, 62% is quite realistic given some recent polls). Here is a picture that illustrates what we have done and found in our example:




Our goal is to do inference—to learn and draw conclusions about the opinions of the entire population of U.S. adults regarding the death penalty on the basis of the opinions of only 1,200 of them.

Can we conclude that 62% of the population favors the death penalty? Another random sample could give a very different result, so we are uncertain. But because our sample is random, we know that our uncertainty is due to chance, not to problems with how the sample was collected. So we can use probability to describe the likelihood that our sample is within a desired level of accuracy. For example, probability can answer the question, *How likely is it that our sample estimate is no more than 3% from the true percentage of all U.S. adults who are in favor of the death penalty?*

Answering this question (which we do using probability) is obviously going to have an important impact on the confidence we can attach to the inference step. In particular, if we find it quite unlikely that the sample percentage will be very different from the population percentage, then we have a lot of confidence that we can draw conclusions about the population on the basis of the sample.

In this module, we discuss probability more generally. Then we begin to develop the probability machinery that underlies inference.
CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 6.1: Why It Matters- Probability and Probability Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.2: Introduction- Another Look at Probability

## What you'll learn to do: Interpret (in context) a probability as a long-run relative frequency of an event.

### LEARNING OBJECTIVES

- Interpret (in context) a probability as a long-run relative frequency of an event.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.2: Introduction- Another Look at Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.3: Another Look at Probability (1 of 2)

## Learning Objectives

- Interpret (in context) a probability as a long-run relative frequency of an event.

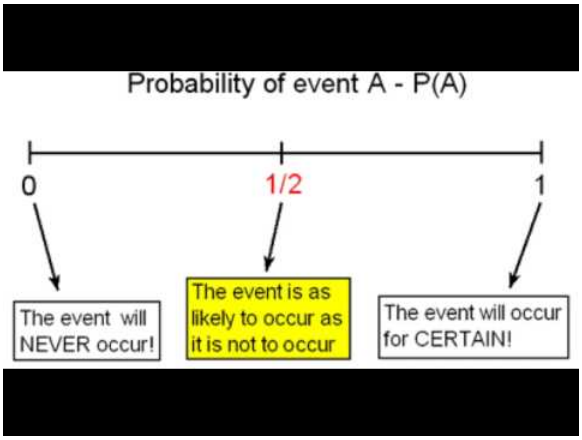
In the module *Relationships in Categorical Data with Intro to Probability*, we used the word *probability* to mean "likelihood" or "chance." We used data to make statements about

- the likelihood that a randomly selected student from a specific college is a Health Science major.
- the risk associated with not wearing a seat belt.
- the chance of a positive drug test for someone who does not use drugs when the test is 94% accurate.

For each of these probability statements, we used a notation $P(A)$ where $A$ is the description of an event. We used the following notation to represent probability statements like the preceding ones:

- $P$(Health Science)
- $P$(fatal accident given that the person was not wearing a seatbelt) = $P$(fatal accident | not wearing a seatbelt)
- $P$(a person is not a drug user given that the person had a positive test result) = $P$(not a drug user | positive test result)

In each case, the probability was a number between 0 and 1. What does this number tell us about the likelihood of an event occurring?




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=216

In summary, the probability that an event will occur is a number between (and including) 0 and 1. We write this idea in mathematical notation as $0 \le P(A) \le 1$.

## Example

## A Closer Look at How We Calculate Probabilities

You may have had some experience with probability using coins, cards, and dice.

*What is the probability that when you flip a coin you get heads?*
*What is the probability that when you roll the dice you get doubles?*

We can answer these types of probability questions without collecting data. In situations where the outcomes are equally likely, we can use mathematics to calculate the probability instead of collecting data. For example, what is the probability of getting heads when you toss a coin? There are two equally likely outcomes: heads or tails. So $P(\mathrm{heads})=\frac{1}{2}$. This is the **theoretical probability** of getting a head when you toss a coin. We determine the number of ways an event can occur and divide by the total number of possible outcomes. No experiments or data collection is necessary.

What is the probability that a community college student is female? Like tossing a coin, this event also has two outcomes: female or male. But is $P(\mathrm{female})=\frac{1}{2}$ ? To estimate this probability, we have to collect data. We can use the data from the West

Coast college that we saw in *Relationships in Categorical Data with Intro to Probability* and estimate that $P(\mathrm{female})=\frac{\mathrm{6,198}}{\mathrm{12,000}}=0.5165$ . Of course, this estimate assumes this college is a representative sample of community colleges. Data from 2010 enrollments at Los Medanos College in California give a different estimate: $P(\mathrm{female})=\frac{\mathrm{5,581}}{\mathrm{9,966}}=0.56$ . Neither estimate is equal to 0.5 because there appear to be more women than men attending community college. These are examples of empirical probabilities.

**Empirical probability** of an event is an estimate, using data, of the likelihood that the event will happen. We can view the probabilities we calculated in *Relationships in Categorical Data with Intro to Probability* as empirical probabilities.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.3: Another Look at Probability (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.4: Another Look at Probability (2 of 2)

## Learning Objectives

- Interpret (in context) a probability as a long-run relative frequency of an event.

## What Is the Relationship between Theoretical and Empirical Probability?

## Learn By Doing

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.4: Another Look at Probability (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.5: Introduction- Probability Rules

## What you'll learn to do: Reason from probability distributions, using probability rules, to answer probability questions.

### LEARNING OBJECTIVES

- Reason from probability distributions, using probability rules, to answer probability questions.
- Use conditional probability to identify independent events.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.5: Introduction- Probability Rules is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.6: Probability Rules (1 of 3)

## Learning Objectives

- Interpret (in context) a probability as a long-run relative frequency of an event.

In our previous discussions of probability, we focused on determining the probability of one event at a time. For example, we used two-way tables in *Relationships in Categorical Data with Intro to Probability* to find the probability that a randomly selected female student from a community college is a Health Science major.

Now we shift our focus to describing the probabilities of all possible outcomes instead of the probability of just one outcome.

We think of all possible outcomes as variable values. Each variable value has a probability. The variable values together with their probabilities are a **probability distribution**.

### Example

## Probability Distribution for Blood Type

Consider the variable *blood type*. This is a categorical variable with variable values A, B, AB, or O. Using relative frequencies from large samples of randomly chosen individuals, we can estimate the probability of choosing a person with a given blood type. Using relative frequencies, the Stanford University's Blood Center (bloodcenter.stanford.edu) gives the probabilities of human blood types in the United States as follows:

| Blood Type | O | A | B | AB |
|---|---|---|---|---|
| Probability | 0.45 | 0.41 | 0.10 | 0.04 |

This table is an example of a probability distribution. Each variable value is assigned a probability.

Notice the following important fact about this probability distribution:

*The sum of all of the probabilities is 1*. This makes sense because we have listed all the outcomes. Since each probability is a relative frequency, these outcomes make up 100% of the observations.

We can use the probability distribution to answer probability questions:

> **Question:** People with blood type O can donate blood to people with any other blood type. For this reason, people with blood type O are called universal donors. *What is the probability that a randomly selected person from the United States is a universal donor?*
> **Answer***P*(universal donor) = *P*(blood type O) = 0.45. There is a 45% chance that a randomly selected person in the United States is a universal donor.

### Example

## Probability Distribution for Boreal Owl Eggs

Boreal owls are common in Canada and Alaska. They are fairly small, averaging 10 inches in length and weighing from 4 to 6 oz. They often make their nests in woodpecker holes. The number of eggs in a boreal owl nest generally ranges from 4 to 6 eggs. Using relative frequencies from large field observations, we can estimate the probability of a nest containing a certain number of eggs.

The variable is *Boreal owl eggs in a nest*. This is a quantitative variable with values 0, 1, 2, 3, 4, 5, or 6 eggs. The probability distribution gives the probability that a nest will have from 0 to 6 eggs.

| Number of Eggs | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Probability | 0.2 | 0.1 | 0.1 | 0.25 | 0.25 | 0.05 | 0.05 |

This table is also an example of a probability distribution. Each variable value is assigned a probability.

Note: *The sum of all of the probabilities is 1*. This is always true for a probability distribution.

We can use the probability distribution to answer probability questions:

> **Question:** *Which is more likely: (1) To find a boreal owl nest with 3 eggs, or (2) To find a boreal owl nest with 4 eggs.*
> **Answer** Both of these events are equally likely. $P(3 \text{ eggs}) = P(4 \text{ eggs}) = 0.25$. There is a 25% chance that if you find a boreal owl nest, it will have 3 eggs. You are equally likely to find a boreal owl nest with 4 eggs.

Notice the following important facts about probability distributions:

- *The outcomes are random events.* When we randomly choose a person, we do not know their blood type. But there is a predictable pattern in the outcomes that is described by the relative frequencies. When we randomly select a boreal owl nest, we do not know how many eggs it will contain, but there is a predictable pattern in the outcomes that is described by the relative frequencies.
- *All outcomes are assigned a probability.*
- *The probabilities are numbers between 0 and 1.* This makes sense because each probability is a relative frequency.
- *The sum of all of the probabilities is 1.* This makes sense because we have listed all the outcomes. Since each probability is a relative frequency, these outcomes make up 100% of the observations.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3552

https://assessments.lumenlearning.co...sessments/3553

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.6: Probability Rules (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.7: Probability Rules (2 of 3)

## Learning Objectives

- Reason from probability distributions, using probability rules, to answer probability questions.

Here we continue to use probability distributions to answer probability questions. We look for some patterns that suggest general rules for determining probabilities.

## Example

## When Can We Add Probabilities?

Compare these two questions. What do the solutions have in common?

**Question 1:** A person with blood type A can receive blood from individuals with type A or O blood. *What is the probability that a randomly selected person from the United States can donate blood to someone with type A blood?*

| Blood Type | O | A | B | AB |
|---|---|---|---|---|
| Probability | 0.45 | 0.41 | 0.10 | 0.04 |

**Answer** $P(\text{donate to A}) = P(\text{blood type A or blood type O}) = 0.45 + 0.41 = 0.86$. There is an 86% chance that a randomly selected person in the United States can donate blood to someone with type A blood.

**Question 2:** *What is the probability that a randomly chosen boreal owl nest will either be empty or contain only 1 egg?*

| Number of Eggs | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Probability | 0.2 | 0.1 | 0.1 | 0.25 | 0.25 | 0.05 | 0.05 |

**Answer** $P(\text{no eggs or 1 egg}) = P(\text{no egg}) + P(\text{1 egg}) = 0.2 + 0.1 = 0.3$. There is a 30% chance that a randomly selected boreal owl nest will be empty or contain only one egg.

What do these solutions have in common?

In each case, we have two events and we want to find the probability that either event A *or* event B occurs. In each case, we added the probabilities. This works because the events have no outcomes in common. When two events have no outcomes in common, they are **disjoint**.

The events “type A blood” and “type O blood” are disjoint. These events cannot both happen at the same time for a single person. A person cannot have both type A blood and type O blood.

The events “no eggs” and “1 egg” are disjoint. These outcomes cannot both happen at the same time for a single nest. A nest cannot contain no eggs and at the same time contain 1 egg.

If two events are disjoint, then we can add their individual probabilities. We write this fact as a rule:

$P(A \text{ or } B) = P(A) + P(B)$

## Comment

We stated the addition rule as a formal rule. A rule is a concise way to summarize a general principle from specific examples. This is one advantage of a rule. One disadvantage of a rule is that sometimes it discourages us from just thinking through a problem. Students often have the experience that they misremember a rule or forget the conditions required for the rule to work. This leads to mistakes that we can avoid if we just think through the problem without worrying about rules. We encourage you to think through probability problems whenever possible without resorting to rules. If you use a rule, be careful to check that the situation meets the conditions required for using the rule.

This addition rule for probabilities only works when the events are disjoint. If the events are not disjoint, the rule does not work. Here is an example of when the rule does not work because the events are not disjoint.

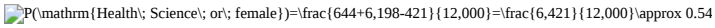
### Example

## When Can We NOT Add Probabilities?

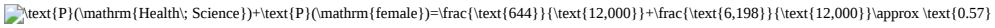
| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

**Question:** *What is the probability that a randomly selected student is either a Health Science major or a female?*

**Answer** There are 644 Health Science majors and 6,198 females, but 421 students are counted twice because they are both Health Science majors and female. We must subtract these students before calculating the relative frequency:



Now let's calculate the individual probabilities and see if the rule works:



**Main point:** $P(\text{Health Science or female}) \neq P(\text{Health Science}) + P(\text{female})$. In other words, the addition rule does not work here. Why not? The two events "Health Science" and "female" are not disjoint. The data set contains people who are both in the Health Science program and female.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3878

https://assessments.lumenlearning.co...sessments/3879

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3880

### Example

## Do We Ever Subtract Probabilities?

Compare these two questions. What do the solutions have in common?

**Question 1:** People with blood type O can donate blood to people with any other blood type. For this reason, people with blood type O are called universal donors. *What is the probability that a randomly selected person from the United States is* ***not*** *a universal donor?*

| Blood Type | O | A | B | AB |
|---|---|---|---|---|
| Probability | 0.45 | 0.41 | 0.10 | 0.04 |

**Answer** $P(\text{NOT a universal donor}) = P(\text{blood type is not type O}) = P(\text{blood type A, B, or AB}) = 0.41 + 0.10 + 0.04 = 0.55$. There is a 55% chance that a randomly selected person in the United States is not a universal donor.

Here is another way we can solve this problem. We can use the idea that all of the probabilities together make up 100% of the possibilities. If we add up all the probabilities in the table, we get 1. We can subtract the probability that someone is type O from 1 to find the probability that the person is not type O:

$P(\text{NOT a universal donor}) = P(\text{blood type is not type O}) = 1 - P(\text{type O}) = 1 - 0.45 = 0.55$

**Question 2:** *What is the probability that a randomly selected boreal owl nest is* ***not*** *empty?*

| Number of Eggs | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Probability | 0.2 | 0.1 | 0.1 | 0.25 | 0.25 | 0.05 | 0.05 |

**Answer***P*(nest is not empty) = *P*(at least one egg) = *P*(1, 2, 3, 4, 5, or 6 eggs) = 0.1 + 0.1 + 0.25 + 0.25 + 0.05 + 0.05 = 0.80. There is an 80% chance that the nest you observe has at least one egg.

Here is another approach:

*P*(nest is not empty) = *P*(at least one egg) = 1 – *P*(0 eggs) = 1 – 0.2 = 0.8.

What do these solutions have in common?

In each case, we have an event that can be interpreted as a "not" statement. The probability that a person is not a universal donor means the person is *not type O*. The probability that a boreal owl nest is empty means the nest *does not contain 0 eggs*. In each case, the easy way to compute the probability is to use the **complement** event. The complement of event A is the event composed of outcomes that are "not *A*." In our examples, the complement of "type O blood" is the event composed of "blood types A, B, or AB." The complement of "0 eggs" is the event composed of "1, 2, 3, 4, 5, or 6 eggs." When two sets of events are complements, their probabilities add to 1.

When one event is the complement of another, then we can use the complement rule:

$$P(\text{not } A) = 1 - P(A)$$

We can use this rule to find probabilities only when the two events are complements. Two events are complements when their probabilities add to 1.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3881

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3882

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.7: Probability Rules (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.8: Probability Rules (3 of 3)

## Learning Objectives

- Use conditional probability to identify independent events.

**Independence and Conditional Probability**

Recall that in the previous module, *Relationships in Categorical Data with Intro to Probability*, we introduced the idea of the conditional probability of an event.

Here are some examples:

- the probability that a randomly selected female college student is in the Health Science program: *P*(Health Science | female)
- *P*(a person is not a drug user given that the person had a positive test result) = *P*(not a drug user | positive test result)

Now we ask the question, *How can we determine if two events are independent?*

## Example

## Identifying Independent Events

*Is enrollment in the Health Science program independent of whether a student is female? Or is there a relationship between these two events?*

| | Arts-Sci | Bus-Econ | Info Tech | Health Science | Graphics Design | Culinary Arts | Row Totals |
|---|---|---|---|---|---|---|---|
| **Female** | 4,660 | 435 | 494 | 421 | 105 | 83 | 6,198 |
| **Male** | 4,334 | 490 | 564 | 223 | 97 | 94 | 5,802 |
| **Column Totals** | 8,994 | 925 | 1,058 | 644 | 202 | 177 | 12,000 |

To answer this question, we compare the probability that a randomly selected student is a Health Science major with the probability that a randomly selected *female* student is a Health Science major. If these two probabilities are the same (or very close), we say that the events are independent. In other words, independence means that being female does not affect the likelihood of enrollment in a Health Science program.

To answer this question, we compare:

- the unconditional probability: *P*(Health Sciences)
- the conditional probability: *P*(Health Sciences | female)

If these probabilities are equal (or at least close to equal), then we can conclude that enrollment in Health Sciences is **independent** of being a female. If the probabilities are substantially different, then we say the variables are **dependent**.



$$\begin{array}{l}P(\mathrm{Health\; Science})=\frac{644}{12,000}\approx 0.054\text{ }(\mathrm{marginal\; probability};\text{ an\; unconditional\; probability})\\ P(\mathrm{Health\; Science}\text{ }|\text{ }\mathrm{female})=\frac{421}{6,198}\approx 0.068\text{ }(\mathrm{conditional\; probability})\end{array}$$

Both conditional and unconditional probabilities are small; however, 0.068 is relatively large compared to 0.054. The ratio of the two numbers is 0.068 / 0.054 = 1.25. So the conditional probability is 25% larger than the unconditional probability. It is much more likely that a randomly selected *female* student is in the Health Science program than that a randomly selected student, without regard for gender, is in the Health Science program. There is a large enough difference to suggest a relationship between being female and being enrolled in the Health Science program, so these events are **dependent**.

## Comment:

To determine if enrollment in the Health Science program is independent of whether a student is female, we can also compare the probability that a student is female with the probability that a *Health Science* student is female.



$$\begin{array}{l}P(\mathrm{female})=\frac{\mathrm{6,198}}{12,000}\approx 0.517\\ P(\mathrm{female}\text{ }|\text{ }\mathrm{Health\; Science})=\frac{421}{644}\approx 0.654\end{array}$$

We see again that the probabilities are not equal. Equal probabilities will have a ratio of one. The ratio is $\frac{\text{0.517}}{\text{0.654}}\approx \text{0.79}$, which is not close to one. It is much more likely that a randomly selected *Health Science student* is female than that a randomly selected student is female. This is another way to see that these events are dependent.

To summarize:

If $P(A \mid B) = P(A)$, then the two events A and B are independent.To say two events are independent means that the occurrence of one event makes it neither more nor less probable that the other occurs.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3883

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3555

https://assessments.lumenlearning.co...sessments/3556

https://assessments.lumenlearning.co...sessments/3557

In *Relationships in Categorical Data with Intro to Probability*, we explored marginal, conditional, and joint probabilities. We now develop a useful rule that relates marginal, conditional, and joint probabilities.

### Example

## A Rule That Relates Joint, Marginal, and Conditional Probabilities

Let's consider our body image two-way table. Here are three probabilities we calculated earlier:

Marginal probability: $P(\mathrm{about\; right})=\frac{855}{\mathrm{1,200}}$

Conditional probability: $P(\mathrm{female}|\mathrm{about\; right})=\frac{560}{855}$

Joint probability: $P(\mathrm{femaleand\; about\; right})=\frac{560}{\mathrm{1,200}}$

Note that these three probabilities only use three numbers from the table: 560, 855, and 1,200. (We grayed out the rest of the table so we can focus on these three numbers.)

Now observe what happens if we multiply the marginal and conditional probabilities from above.

$$\mathrm{P}(\mathrm{about\; right})\mathrm{}\cdot\mathrm{P}(\mathrm{female\; |about\; right})=\frac{855}{1200}\cdot\frac{560}{855}=\frac{560}{1200}$$

The result 560 / 1200 is exactly the value we found for the joint probability.

When we write this relationship as an equation, we have an example of a general rule that relates joint, marginal, and conditional probabilities.

$$\begin{array}{l}P(\mathrm{about\; right})\mathrm{}\cdot P(\mathrm{female}|\mathrm{about\; right})=P(\mathrm{female\; and\; about\; right})\\ \mathrm{marginal\; probability}\cdot\mathrm{conditional\; probability}=\mathrm{joint\; probability}\end{array}$$

In words, we could say:

- The joint probability equals the product of the marginal and conditional probabilities

This is a general relationship that is always true. In general, if *A* and *B* are two events, then

$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$This rule is always true. It has no conditions. It always works.

**When the events are independent**, then $P(B \mid A) = P(\mathrm{B})$. So our rule becomes

$P(A \text{ and } B) = P(A) \cdot P(B)$This version of the rule only works when the events are independent. For this reason, some people use this relationship to identify independent events. They reason this way:

If $P(A \text{ and } B) = P(A) \cdot P(B)$ is true, then the events are independent.

### Comment:

Here we want to remind you that it is sometimes easier to think through probability problems without worrying about rules. This is particularly easy to do when you have a table of data. But if you use a rule, be careful that you check the conditions required for using the rule.

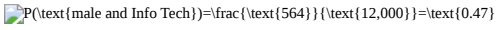
### Example

## Relating Marginal, Conditional, and Joint Probabilities

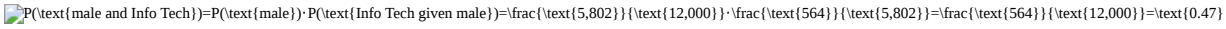
What is the probability that a student is both a male and in the Info Tech program?

There are two ways to figure this out:

(1) Just use the table to find the joint probability:



(2) Or use the rule:



### Learn By Doing

https://assessments.lumenlearning.co...sessments/3884

All of the examples of independent events that we have encountered thus far have involved two-way tables. The next example illustrates how this concept can be used in another context.

### Example

## A Coin Experiment

Consider the following simple experiment. You and a friend each take out a coin and flip it. What is the probability that both coins come up heads?

Let's start by listing what we know. There are two events, each with probability ½.

*P*(your coin comes up heads) = ½
*P*(your friend's coin comes up heads) = ½

We also know that these two events are independent, since the probability of getting heads on either coin is in no way affected by the result of the other coin toss.

We are therefore justified in simply multiplying the individual probabilities:

$$(½)\ (½) = ¼$$

**Conclusion:** There is a 1 in 4 chance that both coins will come up heads.

If we extended this experiment to three friends, then we would have three independent events. Again we would multiply the individual probabilities:

$$(½)\ (½)\ (½) = ⅛$$

**Conclusion:** There is a 1 in 8 chance that all three coins will come up heads.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 6.8: Probability Rules (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.9: Introduction- Discrete Probability Distribution

## What you'll learn to do: Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

### LEARNING OBJECTIVES

- Distinguish between discrete random variables and continuous random variables.
- Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.9: Introduction- Discrete Probability Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.10: Discrete Random Variables (1 of 5)

## Learning Objectives

- Distinguish between discrete random variables and continuous random variables.

In our previous discussion of probability distributions, we did not distinguish between probability distributions for categorical and quantitative variables. Our focus was on developing the rules of probability. We looked at the probability distribution for the categorical variable *blood type*. We also looked at the probability distribution for the quantitative variable *number of boreal owl eggs in a nest*. The probability rules apply in both situations.

Now we focus more closely on probability distributions for quantitative variables. These distributions will be very important when we study statistical inference. Examples of such variables are:

- number of boreal owl eggs in a nest
- number of times a college student changes major
- shoe size
- weight of a student
- foot lengths for adults

*When the outcomes are quantitative, we call the variable a random variable*. In this section, we discuss the probability distributions of discrete random variables and random variables.

**Discrete random variables** have numeric values that can be listed and often can be counted. For example, the variable *number of boreal owl eggs in a nest* is a discrete random variable. Shoe size is also a discrete random variable. Blood type is not a discrete random variable because it is categorical.

**Continuous random variables** have numeric values that can be any number in an interval. For example, the (exact) weight of a person is a continuous random variable. Foot length is also a continuous random variable. Continuous random variables are often measurements, such as weight or length. We view measurements as continuous even though the limitations of a ruler or a scale give discrete measurements. For example, imagine weighing yourself on a digital scale that gives weights to the nearest tenth of a pound. You will get measurements that are rounded to the nearest tenth, such as 152.3 or 165.8. Actual weights could theoretically be any value in an interval, such as 152.345612555 or something like that. So with a discrete variable, you can count the possible values for the variable without rounding off. With a continuous variable, you cannot.

## Comment

The word *random* here means that the outcomes are uncertain in the short run but have a regular distribution or predictable pattern in the long run. In statistics, we reserve the term *random variable* for quantitative variables. This can be a bit confusing because categorical variables can also describe random outcomes.

Now we investigate the probability distributions for discrete random variables.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.10: Discrete Random Variables (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.11: Discrete Random Variables (2 of 5)

## Learning Objectives

- Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

## Probability Distribution for Discrete Random Variables

In this section, we work with probability distributions for discrete random variables. Here is an example:

## Example

Consider the random variable *the number of times a student changes major*.

(For convenience, it is common practice to say: Let $X$ be the random variable *number of changes in major*, or $X$ = number of changes in major, so that from this point we can simply refer to $X$, with the understanding of what it represents.)

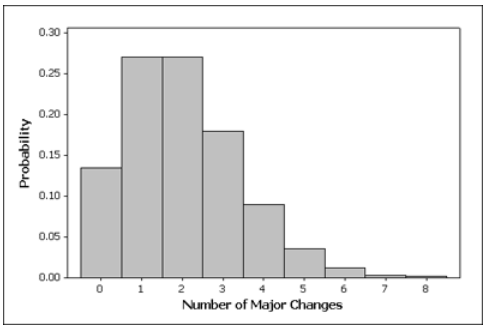
Here is the probability distribution of the random variable $X$:

| $X$=# changes in major | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Probability | 0.135 | 0.271 | 0.271 | 0.180 | 0.090 | 0.036 | 0.012 | 0.003 | 0.002 |

Here is what it tells us:

For a randomly selected student, we cannot predict how many times he or she will change majors, but there is a predictable pattern described by the probability distribution (or model) above. So this is a random variable for which we are assuming the values range from 0 to 8. (In reality, a negligible proportion of students change majors more than 8 times.) The table provides a way to assign probabilities to outcomes. Note that if we add up the probabilities of all possible outcomes (0.135 + 0.271 + … + 0.002), we get exactly 1, which is not surprising (because one of the possible outcomes 0, 1, … , 8 will occur for sure).

Another way to represent the probability distribution of a random variable is with a probability histogram.




The horizontal axis accounts for the range of all possible values of the random variable (in our case, 0–8), and the vertical axis represents the probabilities of those values.

The heights of the bars add to 1, which is not surprising since the heights represent probabilities.

Let's summarize the features of a probability distribution:

- The outcomes described by the model are **random**. This means that individual outcomes are uncertain, but there is a regular, predictable distribution of outcomes in a large number of repetitions.
- The model provides a way of assigning probabilities to all possible outcomes.
- The probability of each possible outcome can be viewed as the relative frequency of the outcome in a large number of repetitions, so like any other probability, it can be any value between 0 and 1.
- The sum of the probabilities of all possible outcomes must be 1.

## Comment

*Where do these probability distributions come from?* Recall that probability distributions can come from data, such as the distribution of boreal owl eggs. Scientists observe thousands of nests and record the number of eggs in each nest. Then they

calculate the relative frequency of each outcome. The relative frequency of each outcome represents the empirical probability for that outcome.

We can also use a mathematical formula to represent a probability distribution. In this case, we make assumptions about how outcomes will be distributed. In other words, we use a mathematical formula to describe the predicted relative frequencies for all possible outcomes. We do not look at mathematical formulas for probability distributions in this course, but we want you to be aware that not all probability distributions come from data.

### Example

Recall the probability distribution of the random variable $X$ = number of changes in major:

| *X=# changes in major* | *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|---|
| *Probability* | *0.135* | *0.271* | *0.271* | *0.180* | *0.090* | *0.036* | *0.012* | *0.003* | *0.002* |

Let's see what kinds of probability questions we can answer using it.

1. *What is the probability that a college student will change majors at most once?*

The phrase "at most once" means either the student never changes majors ($X = 0$) or the student changes majors once ($X = 1$). Therefore, to find this probability, we need to add the probabilities that are highlighted in the table:

| *X=# changes in major* | *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|---|
| *Probability* | *0.135* | *0.271* | *0.271* | *0.180* | *0.090* | *0.036* | *0.012* | *0.003* | *0.002* |

So,

$P$(a college student changes majors at most once) = $P(X = 0) + P(X = 1) = 0.135 + 0.271 = 0.406$

The probability that a randomly selected college student will change majors at most once is about 0.406. We can also say that about 40.6% of the time, a randomly selected college student will change majors at most once.

2. John's parents are concerned that he has decided to change his major for the second time. John claims that he is not unusual. *What is the probability that a randomly selected college student will change his major as often as or more often than John?*

To answer the question about John, we need know the probability that a randomly selected student will change his major 2 or more times. We need to add together the probabilities shaded in the table.

| *# changes in major* | *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Probability* | *0.135* | *0.271* | *0.271* | *0.180* | *0.090* | *0.036* | *0.012* | *0.003* | *0.002* | *0.000* |

$P$(change major 2 or more times) = $P(X = 2) + P(X = 3) + \ldots + P(X = 8) = 0.594$

Here is another way to figure this out. We can use the idea that all of the probabilities together make up 100% of the possibilities. So if we add up all the probabilities in the table we should get 1. Now if we figure out the probability that someone changes majors 0 or 1 times, we can just subtract this from 1 to find the probability that someone changes majors 2 or more times. As we learned previously, this is the complement rule.

$P$(change major 2 or more times) = $1 - [P(X = 0) + P(X = 1)] = 1 - [0.135 + 0.271] = 0.594$

Do you think John has given a convincing argument that he is not unusual? Yes! Fifty-nine percent of the time, a college student will change majors as often as or more often than John did. Stating this same result in terms of probability, we might say, "There is a 59% probability that a randomly selected college student will change majors 2 or more times while in college."

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3875

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3876

We found that changing a major 2 or more times is not very unusual, since it happens about 59% of the time. So…

3. *How often would John need to change his major to be considered unusual?*

One way to answer this question is to just a make a judgment call about what we might consider “unusual” based on the table. For example, we might notice that the probability that a student will change majors 5 or more times is about 5%.

| | |
|---|---|
| $P$(change majors 5 or more times) | $= P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)$ |
| | $= 0.036 + 0.012 + 0.003 + 0.002 = 0.053$ |

An event that occurs only 5% of the time is pretty unusual.

Are there other ways to more definitively determine what might be considered unusual? Well, we might use a measure of center, such as the mean, to determine a “typical” number of times that students change majors. Values that are 2 standard deviations above the mean could be used to identify unusual behavior. We will come back to this question after we have developed an understanding of mean and standard deviation for a probability distribution.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.11: Discrete Random Variables (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.12: Discrete Random Variables (3 of 5)

## Learning Objectives

- Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

**The Mean and Standard Deviation of a Discrete Random Variable**

We now focus on the mean and standard deviation of a discrete random variable. We discuss how to calculate these measures of center and spread for this type of probability distribution, but in general we will use technology to do these calculations.

## Example

## The Mean of a Discrete Random Variable

At Rushmore Community College, there have been complaints about how long it takes to get food from the college cafeteria. In response, a study was conducted to record the total amount of time students had to wait to get their food. The following table gives the total times (rounded to the nearest 5 minutes) to get food for 200 randomly selected students.

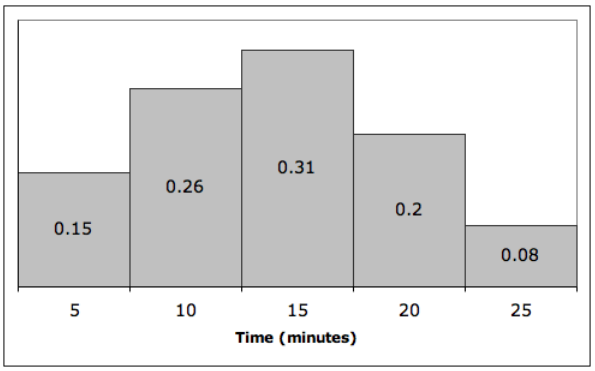
Here is the **frequency table**.

| Time (minutes) | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| Number of students | 30 | 52 | 62 | 40 | 16 |

Using this data, we can create a **probability distribution** for the random variable $X$ = "time to get food." As we have done before, we divide each frequency (count) by the total number of observations. For example, to calculate the probability that a student will have to wait 10 minutes to get their food we divide: (the number of students in the sample that waited 10 minutes) by (the total number of students in the sample) = 52 / 200 = 0.26.

| $X$ = Time (minutes) | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| $P(X)$ | 30 / 200 = **0.15** | 52 / 200 = **0.26** | 62 / 200 = **0.31** | 40 / 200 = **0.20** | 16 / 200 = **0.08** |

Here is the corresponding probability histogram:




**A comment on probability histograms**

In this probability histogram, the area, instead of the height, is the probability. In general, when we work with probability histograms, the area will represent the probability, so we will not worry about the units on the y-axis. Since the area represents the probabilities, the total area is 1.

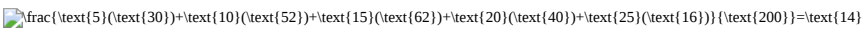
Because in this case we have the actual data in the first table, we start by using that table of actual counts to calculate the mean. However, usually all we have is the probability distribution, so we will also consider how to calculate the mean directly from this information alone.

**Calculating the Mean from the Frequency Table**

| Time (minutes) | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| Number of students | 30 | 52 | 62 | 40 | 16 |

We have 200 observations that are summarized in this table. We have 30 students with a time of 5 minutes, 52 students with a time of 10 minutes, 62 students with a time of 15 minutes, and so on.

To calculate the mean (that is the average), we have to add 30 fives + 52 tens + 62 fifteens + 40 twenties + 16 twenty-fives and then divide by 200. Here is that calculation:

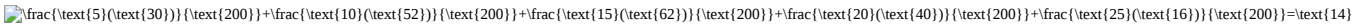


So the mean time for students to get their food in the cafeteria is 14 minutes.

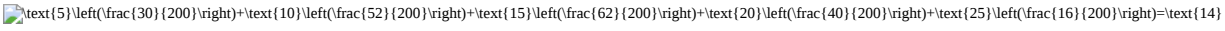
**Calculating the Mean from the Probability Distribution**

Now let's take a closer look at the calculation we just did.

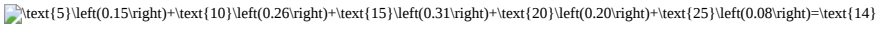
Notice that the large fraction on the left could be broken up into a sum of five smaller fractions all with the denominator 200:



Okay, we are almost there. The last thing to do is rewrite each of these fractions like this:



Here is the same equation with the fractions expressed as decimals:



Look closely at the terms we are adding. In each case, we have the product of one of the possible values of $X$ and its corresponding probability:

| $X$ = Time (minutes) | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| $P(X)$ | 30 / 200 = **0.15** | 52 / 200 = **0.26** | 62 / 200 = **0.31** | 40 / 200 = **0.20** | 16 / 200 = **0.08** |

As we can see, the mean is just a **weighted average**. That is, the mean is the weighted sum of all the possible values of the random variable $X$, where each value is weighted by its probability.

## Comment

**Why Is the Mean a Weighted Average?**

The mean of a discrete random variable $X$ should give us a measure of the long-run average value for $X$. It therefore makes sense to count more heavily those values of $X$ that have a high probability, because they are more likely to occur and will consequently influence the long-run average. On the other hand, those values of $X$ with low probability will not occur very often, so they will have little effect on the long-run average. It therefore makes sense to not give them much weight in our calculation.

## The Formula for the Mean of a Discrete Random Variable

Earlier in the course, when we calculated the mean of a data set, we used the symbol (x-bar) to represent that value. We do not use to represent the mean of a random variable; instead we use (pronounced "mu-sub-x").

Here is the formula that we have come up with for the mean of a discrete random variable. Note that represents the probability of $x$, where $x$ is a value of the random variable $X$.

$$\mu_x = \sum x \cdot p(x)$$

The **mean** equals the **sum** of all the **values of x times their probabilities.**

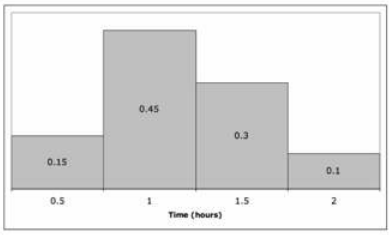
Another term often used to describe the mean is **expected value**. It is a useful term because it reminds us that the mean of a random variable is not calculated on a fixed data set. Rather, the mean (expected value) is a measure of the expected long-term behavior of the random variable.

## Learn By Doing

Drivers entering the short-term parking facility at an airport are given the option to purchase a parking permit for one of four possible time periods: ½ hour, 1 hour, 1½ hours, or 2 hours. Thus, for each driver who enters the parking facility, we can consider their choice of parking time as a discrete random variable. In this case, the random variable $X$ has four possible values: 0.5, 1, 1.5, and 2.

Assume that the probability distribution for $X$ is given by the following table.

| X = parking time (hours) | 0.5 | 1.0 | 1.5 | 2.0 |
|---|---|---|---|---|
| P(X) | 0.15 | 0.45 | 0.30 | 0.10 |




For example, reading from this table, it appears that there is a 15% chance that the next driver entering the parking facility will opt for a ½-hour permit. In the probability histogram, the area of each rectangle (not the height) is the probability of the corresponding $x$-value occurring.

https://assessments.lumenlearning.co...sessments/3558

https://assessments.lumenlearning.co...sessments/3559

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.12: Discrete Random Variables (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.13: Discrete Random Variables (4 of 5)

## Learning Objectives

- Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

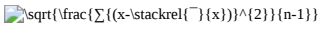
**The Standard Deviation for a Discrete Random Variable**

The mean of a discrete random variable gives us a measure of the long-run average but it gives us no information at all about how much variability to expect. For example, earlier we found that the average cafeteria wait time at Rushmore Community College was 14 minutes. Put in terms of our random variable, this means over the long run, if we continued to keep track of wait times for students entering the cafeteria, their times would average 14 minutes. Some students would get their food in less than 14 minutes, and some would have to wait longer.

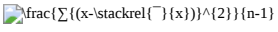
Is that all we need to know? Suppose on the one hand the average time was 14 minutes, but we knew that it was most likely that times would range from 8 to 20 minutes. Compare that to a situation where again the average time was 14 minutes, but it was most likely that times would range only from 13 to 15 minutes. That would give us a different picture of what the problem at the cafeteria might be. What we need is a measure of how much variability to expect in a random variable $X$ over the long run. The standard deviation is that measure.

Just as we need both the mean and standard deviation to get a full picture of the shape of a data set, we need both the mean and standard deviation of a random variable to understand its likely long-term behavior.

In *Summarizing Data Graphically and Numerically*, we used the following formula to compute the standard deviation of a data set.

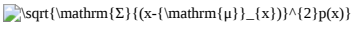


As you may recall, the most important part of this formula is the term inside the square root, which we call the *average of the squares of the deviations from the mean*.

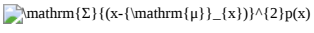


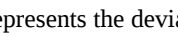
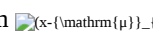
As we will see, the formula for the standard deviation for a discrete random variable has a lot in common with this formula.

Here is the formula for the standard deviation of a discrete random variable. Note that represents the probability of $x$, where $x$ is a value of the random variable $X$. And again stands for the mean of $X$.



Again, we focus on the term inside the square root:



The term  here represents the deviation of each value of the random variable $X$ from the mean , just as the term  represents the deviation of each observation of the data set from the mean .

In both cases, we proceed to sum the squares of these deviations. In the case of a data set, we divide by $n - 1$ to find the average squared deviation. However, in the case of a discrete random variable, we again use a weighted average. Why? Because we don't want to give undue weight to values of $X$ that are unlikely to occur. So those values of $X$, even if far from the mean , will not contribute much to the standard deviation if their probability is low. On the other hand, values of $X$ with large probabilities will count more in our calculation of the standard deviation of $X$.

## Example

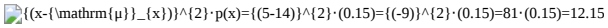
## Cafeteria Wait Times

Let's revisit the problem about wait times in the cafeteria at Rushmore Community College. Recall the following probability distribution.

| $X$ = Time (minutes) | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| $P(X)$ | **0.15** | **0.26** | **0.31** | **0.20** | **0.08** |

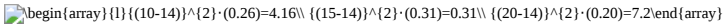
On the previous page, we found that the average wait time is 14 minutes. Now we will compute the standard deviation of wait times and think a bit about what it tells us.

We start by computing the squared deviations from the mean and weighting them by the probability. For the first value of $X$, we have



$$(x-\mu_x)^2 \cdot p(x)=(5-14)^2 \cdot (0.15)=(-9)^2 \cdot (0.15)=81 \cdot (0.15)=12.15$$

Performing the same operation on the next three values of $X$ will give us



$$\begin{array}{l}(10-14)^2 \cdot (0.26)=4.16 \\ (15-14)^2 \cdot (0.31)=0.31 \\ (20-14)^2 \cdot (0.20)=7.2\end{array}$$

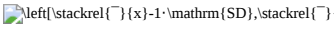
### Learn By Doing

https://assessments.lumenlearning.co...sessments/3560

The next step of the formula is to add up the weighted square deviations from the mean, as follows:

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3561

Recall that in *Summarizing Data Graphically and Numerically* we used the standard deviation of a quantitative data set to give a range of typical values. This range of typical values was formed by blocking off an interval 1 standard deviation to the right and left of the mean. In other words, the range of typical values was  . Exactly the same thing can be done in the current context of random variables.

$$\left[\bar{x}-1 \cdot \mathrm{SD}, \bar{x}+1 \cdot \mathrm{SD}\right]$$

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3562

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 6.13: Discrete Random Variables (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.14: Discrete Random Variables (5 of 5)

## Learning Objectives

- Use probability distributions for discrete and continuous random variables to estimate probabilities and identify unusual events.

Here is another example of how to use the mean and standard deviation of a discrete random variable to identify unusual values for a random variable.

## Example

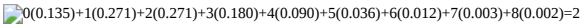
## Changing Majors

Here we have again the probability distribution of the number of changes in major.

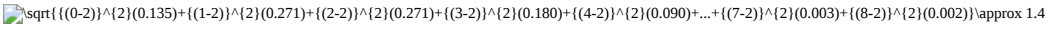
| $X$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $P(X)$ | 0.135 | 0.271 | 0.271 | 0.180 | 0.090 | 0.036 | 0.012 | 0.003 | 0.002 |

*How often do we expect a college student to change majors?*

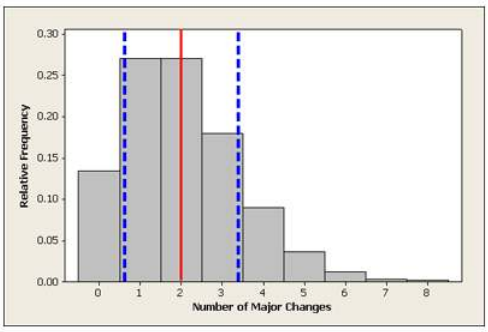
This question is asking for the expected value, which is the mean of the probability distribution. So we calculate the weighted average, as before:




*What is the standard deviation of the probability distribution?*



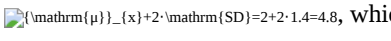

We have drawn lines to show the mean and 1 standard deviation above and below the mean.




Recall that earlier, we discussed what would be considered an unusual (and not unusual) number of changes in major, and we used probability calculations to assess that. For example, we found that changing majors 5 or more times occurs only about 5% of the time and therefore can be considered unusual.

Another way to think about defining "unusual" is to look at outcomes relative to the mean. We might consider outcomes more than 2 standard deviations above the mean as unusual.

What values are more than 2 standard deviations above the mean of 2?

Mean + 2 (standard deviation) = 

, which rounds to 5.

We conclude from this line of reasoning that a college student who changes majors 5 or more times is "unusual."

In *Summarizing Data Graphically and Numerically*, we used the standard deviation to identify usual, or typical, values. We said that a typical range of values falls within 1 standard deviation of the mean. We can use a similar idea here.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3877

## Example

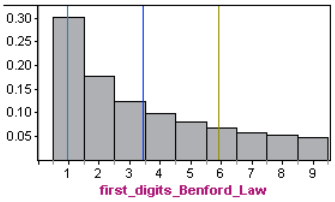
## Detecting Fraud

Legitimate records often display a surprising pattern that is not present in faked tax returns or other fraudulent accounting records. In legitimate records, the distribution of first digits can be modeled using Benford's law. For example, suppose the total income recorded on a tax return is $20,712. The first digit is 2. Now we examine a very large number of tax returns and record the first digit of total income for all of the returns. The relative frequency of each first digit will behave according to Benford's law.

| *First digit from legit tax records* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* |
|---|---|---|---|---|---|---|---|---|---|
| *Probability predicted by Benford's Law* | *0.301* | *0.176* | *0.125* | *0.097* | *0.079* | *0.067* | *0.058* | *0.051* | *0.046* |

Benford's law can also be described using a mathematical formula, but we will not go into that here. Instead, let's double-check that this distribution meets the criteria for a probability distribution of a discrete random variable. For a randomly selected tax return, we cannot predict what the first digit will be, but the first digits behave according to a predictable pattern described by Benford's law. The model assigns probabilities to all possible values for a first digit (notice that the first digit cannot be zero). All possible outcomes taken together have a probability of 1. You can verify this by adding together the probabilities in the table.

Here is the probability distribution for first digits based on Benford's law shown in a histogram. The mean is approximately 3.4, with a standard deviation of about 2.5 (calculations not shown).

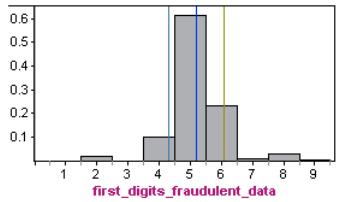



Now, let's compare this distribution to real data.

The second line in the following table is the probability distribution for the first significant digit in true tax data collected by Mark Nigrini from 169,662 IRS model files. You can see that relative frequencies of first digits in the legitimate tax records follow Benford's law very closely.

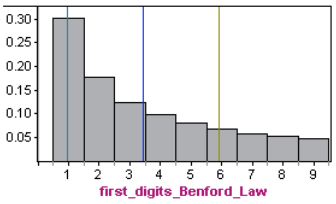
| *First digit from legit tax records* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* |
|---|---|---|---|---|---|---|---|---|---|
| *Relative frequency of first digits in legit tax records* | *0.305* | *0.178* | *0.126* | *0.096* | *0.078* | *0.066* | *0.056* | *0.050* | *0.045* |
| *Probability predicted by Benford's Law* | *0.301* | *0.176* | *0.125* | *0.097* | *0.079* | *0.067* | *0.058* | *0.051* | *0.046* |

By comparison, here is the probability distribution for first digits in fraudulent tax records from a study of fraudulent cash disbursement and payroll expenditures conducted in 1995 by the district attorney's office in Kings County, New York. For fraudulent data, the mean is approximately 5.2, with a standard deviation of about 0.9.




Obviously, the relative frequencies of first digits from the fraudulent data do not follow Benford's law (shown again below). The distributions have very different shapes, means, and standard deviations. Compared to legitimate data, in fraudulent data, we are

much more likely to see numbers with a first digit of 5 and much less likely to see numbers with a first digit of 1, 2, or 3.




**Comment**

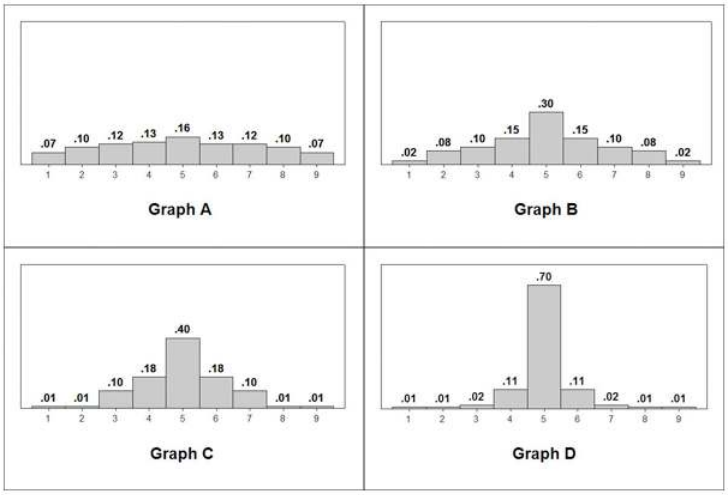
When we compare the two distributions above, we can get a better understanding of the standard deviation of a random variable. The distribution in which it is more likely to find values that are further from the mean will have a larger standard deviation.

Likewise, the distribution in which it is less likely to find values that are further from the mean will have a smaller standard deviation.

In the fraudulent distribution, values like 1 or 2 that are far from the mean are very unlikely. On the other hand, in the Benford's law distribution, the values 1 and 2 are quite likely. Indeed, the standard deviation of the Benford law is 2.5, which is larger than the standard deviation of 0.9 in the fraudulent distribution.

### Learn By Doing

Use the following histograms to answer the activity question:




https://assessments.lumenlearning.co...sessments/3563

### Let's Summarize

- The probability of an event is a measure of the likelihood that the event occurs.
- Probabilities are always between 0 and 1. The closer the probability is to 0, the less likely the event is to occur. The closer the probability is to 1, the more likely the event is to occur.
- The two ways of determining probabilities are empirical and theoretical.
  - Empirical methods use a series of trials that produce outcomes that cannot be predicted in advance (hence the uncertainty). The probability of an event is approximated by the relative frequency of the event.
  - Theoretical methods use the nature of the situation to determine probabilities. Probability rules allow us to calculate theoretical probabilities.
- Some common probability rules:
  - The probability of the complement of an event $A$ can be found by subtracting the probability of $A$ from 1: $P(\text{not } A) = 1 - P(A)$

- Events are called disjoint or mutually exclusive if they have no events in common. If $A$ and $B$ are disjoint events, then $P(A \text{ or } B) = P(A) + P(B)$.
- When the knowledge of the occurrence of one event $A$ does not affect the probability of another event B, we say the events are independent. If $A$ and $B$ are independent events, then $P(A \text{ and } B) = P(A) \cdot P(B)$.

- When we have a quantitative variable with outcomes that occur as a result of some random process (e.g., rolling a die, choosing a person at random), we call it a *random variable*.
- There are two types of random variables:
  - Discrete random variables have numeric values that can be listed and often can be counted.
  - Continuous random variables can take any value in an interval and are often measurements. This type of random variable will be discussed in section 6.2.
- A probability distribution of a random variable tells us the probabilities of all the possible outcomes (for discrete random variables) of the variable or ranges of values (for continuous random variables). A probability distribution shows us the regular, predictable distribution of outcomes in a large number of repetitions of a random variable.
- For a discrete random variable, the probabilities of values are areas of the corresponding regions of the probability histogram for the variable.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.14: Discrete Random Variables (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.15: Introduction- Continuous Probability Distribution

## What you'll learn to do: Use a probability distribution for a continuous random variable to estimate probabilities and identify unusual events.

### LEARNING OBJECTIVES

- Use a probability distribution for a continuous random variable to estimate probabilities and identify unusual events.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.15: Introduction- Continuous Probability Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.16: Continuous Probability Distribution (1 of 2)

## Learning Objectives

- Use a probability distribution for a continuous random variable to estimate probabilities and identify unusual events.

In the previous section, we learned about discrete probability distributions. We used both probability tables and probability histograms to display these distributions. In this section, we shift our focus from discrete to continuous random variables. We start by looking at the probability distribution of a discrete random variable and use it to introduce our first example of a probability distribution for a continuous random variable.

## Example

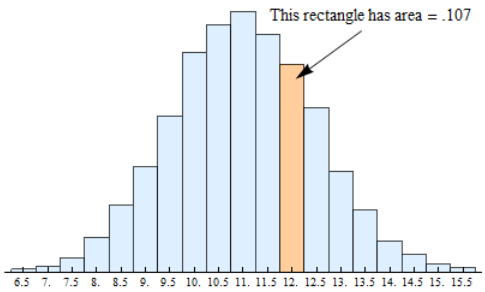
## Shoe Size

Let $X$ = the shoe size of an adult male. $X$ is a discrete random variable, since shoe sizes can only be whole and half number values, nothing in between. For this example we will consider shoe sizes from 6.5 to 15.5. So the possible values of $X$ are 6.5, 7.0, 7.5, 8.0, and so on, up to and including 15.5. Here is the probability table for $X$:

| $X$ | 6.5 | 7 | 7.5 | 8 | 8.5 | 9 | 9.5 | 10 | 10.5 | 11 | 11.5 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P(X)$ | 0.001 | 0.003 | 0.007 | 0.018 | 0.034 | 0.054 | 0.080 | 0.113 | 0.127 | 0.134 | 0.122 | 0.107 |

| $X$ | 12.5 | 13 | 13.5 | 14 | 14.5 | 15 | 15.5 |
|---|---|---|---|---|---|---|---|
| $P(X)$ | 0.085 | 0.052 | 0.032 | 0.016 | 0.009 | 0.004 | 0.002 |

And here is the probability histogram that corresponds to the table.



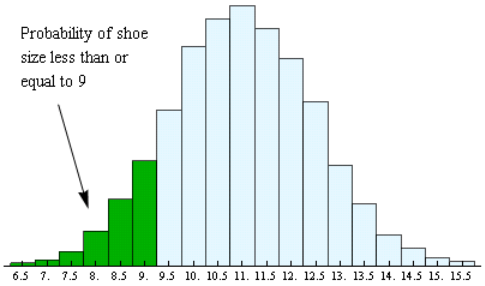

As is always the case for probability histograms, the area of the rectangle centered above each value is equal to the corresponding probability. For example, in the preceding table, we see that the probability for $X = 12$ is 0.107.

In the probability histogram, the rectangle centered above 12 has area = 0.107.

We write this probability as $P(X = 12) = 0.107$.

And finally, as is the case for all probability histograms, because the sum of the probabilities of all possible outcomes must add up to 1, the sums of the areas of all of the rectangles shown must also add up to 1.

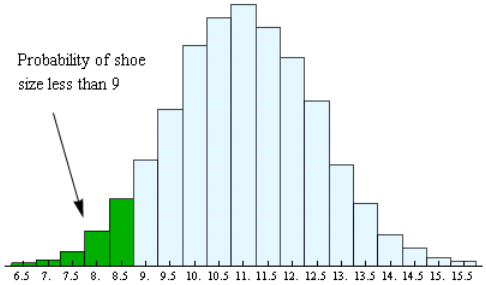
Now we can find the probability of shoe size taking a value in any interval just by finding the area of the rectangles over that interval. For instance, the area of the rectangles up to and including 9 shows the probability of having a shoe size less than or equal to 9.




We can find this probability (area) from the table by adding together the probabilities for shoe sizes 6.5, 7.0, 7.5, 8.0, 8.5 and 9. Here is that calculation:

0.001 + 0.003 + 0.007 + 0.018 + 0.034 + 0.054 = 0.117Total area of the six green rectangles = 0.117 = probability of shoe size less than or equal to 9. We write this probability as $P(X \leq 9) = 0.117$.

Recall that for a discrete random variable like shoe size, the probability is affected by whether or not we include the end point of the interval. For example, the area – and corresponding probability – is reduced if we consider only shoe sizes strictly less than 9:




This time when we add the probabilities from the table, we exclude the probability for shoe size 9 and just add together the probabilities for shoe sizes 6.5, 7.0, 7.5, 8.0, and 8.5:

0.001 + 0.003 + 0.007 + 0.018 + 0.034 = 0.063

Total area of the five rectangles in green = 0.063 = probability of shoe size less than 9. We write this probability as

$P(X < 9) = 0.063$

## Spotlight on Inequality Notation

Here is a review of inequality notation:

### The symbol “<”means “less than”

- Here is a correct use of this symbol: 3 < 12. We read this left to right as 3 is less than 12.
- You can think of the “less than” symbol as an arrow pointing to the smaller number.
- Some students remember the “less than” symbol from elementary school as a hungry alligator that is eating the larger number:

3 12 is 3 < 12

- X < 12 means X is any number less than 12. If X represents shoe sizes, this includes whole and half sizes smaller than size 12.
- P(X < 12) is the probability that X is less than 12.

### The symbol “≤”means “less than or equal to”

- X ≤ 12 means X can be 12 or any number less than 12. If X is shoe sizes, this includes size 12 as well as whole and half sizes less than size 12.
- We often say “**at most** 12” to indicate X ≤ 12.

- P(X ≤ 12) is the probability that X is 12 or less than 12.

### The symbol ">"means "greater than"

- Here is a correct use of this symbol: 15 > 12. We read this left to right as 15 is greater than 12.
- You can also think of the "greater than" symbol as an arrow pointing (as before) to the smaller number.
- Or you can use the hungry alligator idea. The hungry alligator that is still eating the larger number:

- X > 12 means X is any number greater than 12. If X is shoe sizes, this includes whole and half sizes larger than size 12.
- P(X > 12) is the probability that X is greater than 12.

### The symbol "≥" means "greater than or equal to"

- X ≥ 12 means X can be 12 or any number greater than 12. If X is shoe sizes, this includes size 12 as well as whole and half sizes greater than size 12.
- We often say "**at least** 12" to indicate X ≥ 12.
- P(X ≥ 12) is the probability that X is 12 or greater than 12.

### To indicate an interval we combine "less than" and "greater than" symbols:

- To indicate the interval between 9 and 12, we write 9 < X < 12. This interval says " 9 is less than X and X is also less than 12." So this interval includes numbers greater than 9 but also less than 12. For example, 10 is in this interval but 13 is not. Also, 9 and 12 are **not** in this interval.
- P(9 < X < 12) is the probability that X is between 9 and 12.
- P(9 ≤ X ≤ 12) is the probability that X is the same interval except that the interval also includes 9 and 12.

### Transition to Continuous Random Variables

Now we will make the transition from discrete to continuous random variables. Instead of shoe size, let's think about foot length. Unlike shoe size, this variable is not limited to distinct, separate values, because foot lengths can take any value over a *continuous* range of possibilities. In other words, foot length, unlike shoe size, can be measured as precisely as we want to measure it. For example, we can measure foot length to the nearest inch, the nearest half inch, the nearest quarter of an inch, the nearest tenth of an inch, etc. Therefore, foot length is a *continuous random variable*.

What happens to the probability histogram when we measure foot length with more precision? When we increase the precision of the measurement, we will have a larger number of bins in our histogram. This makes sense because each bin contains measurements that fall within a smaller interval of values. For example, if we measure foot lengths in inches, one bin will contain measurements from 6-inches up to 7-inches. But if we measure foot lengths to the nearest half-inch, then we now have two bins: one bin with lengths from 6 up to 6.5-inches and the next bin with lengths from 6.5 up to 7-inches.

You can use the following simulation to see what happens to the probability histogram as the width of intervals decrease. Change the interval width by clicking on 0.5 in., 0.25 in., or 0.1 in.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=242

At the bottom of the simulation is an option to add a curve. This curve is generated by a mathematical formula to fit the shape of the probability histogram. Check "Show curve" and click through the different bin widths. Notice that as the width of the intervals gets smaller, the probability histogram gets closer to this curve. More specifically, the area in the histogram's rectangles more closely approximates the area under the curve. If we continue to reduce the size of the intervals, the curve becomes a better and better way to estimate the probability histogram. We'll use smooth curves like this one to represent the probability distributions of continuous random variables. This idea is discussed in more detail on the next page.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.16: Continuous Probability Distribution (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.17: Continuous Probability Distribution (2 of 2)

## Learning Objectives

- Use a probability distribution for a continuous random variable to estimate probabilities and identify unusual events.

Previously, we examined the probability distribution for foot length. For foot length and for all other *continuous random variables*, the probability distribution can be approximated by a smooth curve called a *probability density curve*.

Recall that these smooth curves are mathematical models. We use a mathematical model to describe a probability distribution so that we can use technology and the equation of this model to estimate probabilities. (As we mentioned earlier, we do not study the equation for this curve in this course, but every statistical package uses this equation, and the area under the corresponding curve, to estimate probabilities.)

As in a probability histogram, the total area under the density curve equals 1, and the curve represents probabilities by area. To find the probability that $X$ is in an interval, find the area above the interval and below the density curve.

For example, if $X$ is foot length, let's find $P(10 < X < 12)$, the probability that a randomly chosen male has a foot length anywhere between 10 and 12 inches. This probability is the area above the interval $10 < X < 12$ and below the curve. We shaded this area with green in the following graph.

If, for example, we are interested in $P(X < 9)$, the probability that a randomly chosen male has a foot length of less than 9 inches, we have to find the area shaded in green below:

## Comments

1. We have seen that for a *discrete* random variable like shoe size, $P(X < 9)$ and $P(X \leq 9)$ have different values. In other words, including the endpoint of the interval changes the probability. In contrast, for a *continuous* random variable like foot length, the probability of a foot length of less than or equal to 9 will be the same as the probability of a foot length of strictly less than 9. In other words, $P(X < 9) = P(X \leq 9)$. Visually, in terms of our density curve, the area under the curve up to and including a certain point is the same as the area up to and excluding the point. This is because there is no area over a single point. There are infinitely many possible values for a continuous random variable, so technically the probability of any single value occurring is zero!
2. It should be clear now why the total area under any probability density curve must be 1. The total area under the curve represents $P$ ($X$ gets a value in the interval of its possible values). Clearly, according to the rules of probability, this must be 1, or always true.
3. Density curves, like probability histograms, may have any shape imaginable as long as the total area underneath the curve is 1. Each density curve is a mathematical model with an equation that is used to find the area underneath the curve.

## Let's Summarize

The probability distribution of a continuous random variable is represented by a probability density curve. The probability that $X$ has a value in any interval of interest is the area above this interval and below the density curve.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.17: Continuous Probability Distribution (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.18: Introduction- Normal Random Variables

## What you'll learn to do: Use a normal probability distribution to estimate probabilities and identify unusual events.

### LEARNING OBJECTIVES

- Use a normal probability distribution to estimate probabilities and identify unusual events.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.18: Introduction- Normal Random Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.19: Normal Random Variables (1 of 6)

## Learning Objectives

- Use a normal probability distribution to estimate probabilities and identify unusual events.

In *Summarizing Data Graphically and Numerically*, we encountered data sets, such as height and weight, with distributions that are fairly symmetric with a central peak. We call these *bell-shaped*.

Many variables, such as weight, shoe sizes, foot lengths, and other human physical characteristics, exhibit these properties. The symmetry indicates that the variable is just as likely to take a value a certain distance below its mean as it is to take a value that same distance above its mean. The bell shape indicates that values closer to the mean are more likely, and it becomes increasingly unlikely to take values far from the mean in either direction.

We use a mathematical model with a smooth bell-shaped curve to describe these bell-shaped data distributions. These models are called **normal curves** or **normal distributions**. They were first called "normal" because the pattern occurred in many different types of common measurements.

The general shape of the mathematical model used to generate a normal curve looks like this:

**Observations of Normal Distributions**

There are many normal curves. Even though all normal curves have the same bell shape, they vary in their center and spread.

Because normal curves are mathematical models, we use Greek letters to represent the mean and standard deviation of a normal curve. The *mean* of a normal distribution locates its *center*. We use the Greek letter μ (pronounced "mu" ) to represent the mean. We use the Greek letter σ (pronounced "sigma") to represent the standard deviation of a normal distribution. The *standard deviation* determines the *spread* of the distribution. In fact, the shape of a normal curve is completely determined by specifying its standard deviation. As we will see, if two normal distributions have the same standard deviation, then the shapes of their normal curves will be identical.

Following are some observations we can make as we look at the figure above:

- The black and the red normal curves have means or centers at μ = 10. However, the red curve is more spread out and thus has a larger standard deviation. Notice that the red normal curve is also shorter. This makes sense because these curves are probability density curves, so the area under each curve has to be 1.
- The black and the green normal curves have the same standard deviation or spread.

## Comment

- We use $\stackrel{¯}{x}$ to represent the mean of data in a sample. We use μ to represent the mean of a density curve defined by a mathematical model.
- We use SD or $s_x$ to represent the standard deviation of data in sample. We use σ to represent the standard deviation of a density curve defined by a mathematical model.

The normal curve has a central role in statistical inference, as we'll see in *Linking Probability to Statistical Inference*. Understanding the normal distribution is an important step in the direction of our overall goal, which is to relate sample means or proportions to population means or proportions. The goal of this section is to help you better understand normal random variables and their distributions.

All normal curves share a basic geometry. While the mean locates the center of a normal curve, it is the standard deviation that is in control of the geometry. To see how, let's examine a few pictures of normal curves to see what they reveal.

## Example

## One Standard Deviation on Each Side of the Mean

Let's start with a random variable $X$ that has a normal distribution with **mean = 10** and **standard deviation = 2**. Let's practice our new notation. Here we would write μ = 10 and σ = 2 .

The normal curve for $X$ is shown below.

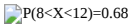
As expected, the mean $\mu = 10$ is located at the center of the normal curve. The other two arrows point to values 1 standard deviation on each side of the mean.

The point *1 standard deviation less than the mean* is represented by $\mu - \sigma$ . Since $\mu = 10$ and $\sigma = 2$, this point is located at $10 - 2 = 8$, as shown.

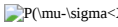
The point *1 standard deviation more than the mean* is represented by $\mu + \sigma$ . Since $\mu = 10$ and $\sigma = 2$, this point is located at $10 + 2 = 12$, as shown.

You will notice we have indicated that the area of the green region is 0.68. So we can say that the probability of $X$ being between 8 and 12 equals 0.68.

Or, using our probability notation, we could write:




Now here is an interesting fact. If we took *any* normal distribution and drew a similar picture, the probability that a value falls within 1 standard deviation of the mean is always the same. Here are several ways to express this idea:

- For any normal curve, the central area within 1 standard deviation of the mean equals 0.68.
- Roughly 68% of the time we will expect $X$ to have a value within 1 standard deviation of the mean.
- 

.

This is a big deal. It is one of the things that makes normal curves special. In general, probability density curves for continuous random variables with different shapes don't have this special property.

Let's put this idea in context. If the weight of babies at birth follows a normal distribution with mean $\mu = 3{,}500$ grams and standard deviation $\sigma = 600$ grams, then we can conclude that most babies – that is, about 68% – will weigh somewhere between 2,900 grams (i.e., $3{,}500 - 600 = 2{,}900$) and 4,100 grams (i.e., $3{,}500 + 600 = 4{,}100$).

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.19: Normal Random Variables (1 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.20: Normal Random Variables (2 of 6)

## Learning Objectives

- At these points, the curve changes the direction of its bend and goes from bending upward to bending downward, or vice versa. A point like this on a curve is called an **inflection point**. Every normal curve has inflection points at exactly 1 standard deviation on each side of the mean.
- Click here to open this simulation in its own window.

  An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=250

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3564

Now we extend this idea to look at the probability of a value falling within 2 standard deviations of the mean or 3 standard deviations of the mean.

If $X$ is a normal random variable with mean $\mathrm{\mu}$ and standard deviation $\mathrm{\sigma}$, then

1. To summarize using probability notation:

$$\begin{array}{l}\mathrm{1.}P(\mu-\sigma<X<\mu+\sigma)=0.68\\ \mathrm{2.}P(\mu-2\sigma<X<\mu+2\sigma)=0.95\\ \mathrm{3.}P(\mu-3\sigma<X<\mu+3\sigma)=0.997\end{array}$$

### Comment

Let's take a moment to look a bit deeper at what the empirical rule tells us.

- https://assessments.lumenlearning.co...sessments/3565
  CC licensed content, Shared previously
  - Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.20: Normal Random Variables (2 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.21: Normal Random Variables (3 of 6)

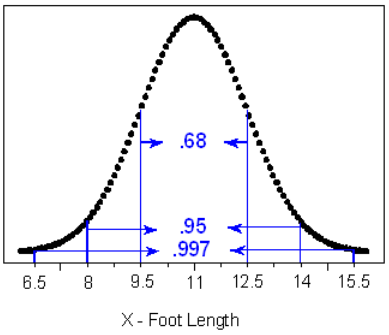
## Learning Objectives

- Use a normal probability distribution to estimate probabilities and identify unusual events.

## Example

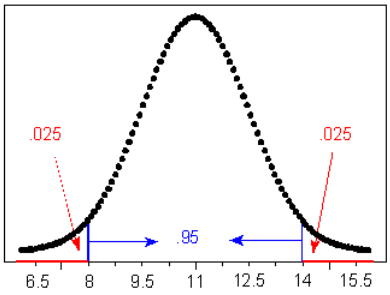
## The Empirical Rule in a Context

Suppose that foot length of a randomly chosen adult male is a normal random variable with mean $\mu=11$ and standard deviation $\sigma=1.5$ . Then the empirical rule lets us sketch the probability distribution of $X$ as follows:




**(a)** What is the probability that a randomly chosen adult male will have a foot length between 8 and 14 inches?
**Answer** 0.95, or 95%
**(b)** An adult male is almost guaranteed (0.997 probability) to have a foot length between what two values?
**Answer** 6.5 and 15.5 inches
**(c)** The probability is only 2.5% that an adult male will have a foot length greater than how many inches?
**Answer** 14 inches

Ninety-five percent of the area is within 2 standard deviations of the mean, so 2.5% of the area is in the tail above 2 standard deviations. The $x$-value 2 standard deviations above the mean is 14 inches.




Now you should try a few: questions (d), (e), and (f) are presented in the Learn By Doing activity. Use the figure preceding question (a) to help you.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3567

**Comment**

Notice that there are two types of problems we may want to solve: those like (a) and, from the Learn By Doing activity, (d) and (e), in which a particular interval of values of a normal random variable is given and we are asked to find a probability; and those like (b), (c), and, from the Learn By Doing, (f), in which a probability is given and we are asked to identify values of the normal random variable.

CC licensed content: Shared previously
• Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.21: Normal Random Variables (3 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.22: Normal Random Variables (4 of 6)

## Learning Objectives

- Use a normal probability distribution to estimate probabilities and identify unusual events.

Let's go back to our example of foot length: How likely or unlikely is it for a male's foot length to be more than 13 inches?

Because 13 inches doesn't happen to be exactly 1, 2, or 3 standard deviations away from the mean, we could give only a very rough estimate of the probability at this point. Clearly, the empirical rule only describes the tip of the iceberg, and although it serves well as an introduction to the normal curve and gives us a good sense of what would be considered likely and unlikely values, it is very limited in the probability questions it can help us answer.

Here is another familiar normal distribution:

Suppose we are interested in knowing the probability that a randomly selected student will score 633 or more on the math portion of her SAT (this is represented by the red area). Again, 633 does not fall exactly 1, 2, or 3 standard deviations above the mean. Notice, however, that a SAT score of 633 and a foot length of 13 are both about one-third of the way between 1 and 2 standard deviations. As you continue to read this page, you'll realize that this positioning relative to the mean is the key to finding probabilities.

## Finding Probabilities for a Normal Random Variable

As we saw, the empirical rule is very limited in helping us answer probability questions. It is limited to questions involving values that fall exactly 1, 2, and 3 standard deviations away from the mean.

We can approach the answering of probability questions in two possible ways: a table and technology. In the next section, you will learn how to use technology to convert between *z*-scores and probabilities.

## Standardizing Values

The first step to assessing a probability associated with a normal value is to determine the *relative* value with respect to all the other values taken by that normal variable. This is accomplished by determining how many standard deviations below or above the mean that value is.

### Example

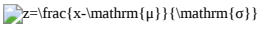
## Foot Length

How many standard deviations below or above the mean male foot length is 13 inches? Since the mean is 11 inches, 13 inches is 2 inches above the mean. Since a standard deviation is 1.5 inches, this would be 2 / 1.5 = 1.33 standard deviations above the mean. Combining these two steps, we could write:

(13 in. − 11 in.) / (1.5 in. per standard deviation) = (13 − 11) / 1.5 standard deviations = +1.33 standard deviations

In the language of statistics, we have just found the *z*-score for a male foot length of 13 inches to be $z = +1.33$. Or, to put it another way, we have *standardized* the value of 13. In general, the standardized value *z* tells how many standard deviations below or above the mean the original value is. It is calculated as follows:

*z*-score = (value – mean) / standard deviationThe convention is to denote a value of our normal random variable $X$ with the letter $x$. Since the mean is written μ and the standard deviation σ, we may write the standardized value as




Notice that since σ is always positive, for values of $x$ above the mean (μ), $z$ will be positive; for values of $x$ below μ, $z$ will be negative.

### Example

## Standardizing Foot Measurements

Let's go back to our foot length example and answer some more questions.

**(a)** What is the standardized value for a male foot length of 8.5 inches? How does this foot length relate to the mean?
$z = (8.5 - 11) / 1.5 = -1.67$. This foot length is 1.67 standard deviations *below* the mean.
**(b)** A man's standardized foot length is +2.5. What is his actual foot length in inches? If $z = +2.5$, then his foot length is 2.5 standard deviations above the mean. Since the mean is 11 and each standard deviation is 1.5, we get that the man's foot length is $11 + 2.5(1.5) = 14.75$ inches.

Thez-score also allows us to compare values of different normal random variables. Here is an example:

**(c)** In general, women's foot length is shorter than men's. Assume that women's foot length follows a normal distribution with a mean of 9.5 inches and standard deviation of 1.2. Ross's foot length is 13.25 inches, and Candace's foot length is only 11.6 inches. Which of the two has a longer foot relative to his or her gender group?

To answer this question, let's find the *z*-score of each of these two normal values, bearing in mind that each value comes from a different normal distribution.

Ross: *z*-score = (13.25 − 11) / 1.5 = 1.5 (Ross's foot length is 1.5 standard deviations above the mean foot length for men).

Candace: *z*-score = (11.6 − 9.5) / 1.2 = 1.75 (Candace's foot length is 1.75 standard deviations above the mean foot length for women).

Note that even though Ross's foot is longer than Candace's, Candace's foot is longer relative to their respective genders.

To Sum Up…Problem (c) illustrates how *z*-scores become crucial when you want to **compare distributions**.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.22: Normal Random Variables (4 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.23: Normal Random Variables (5 of 6)

## Learning Objectives

- Use a normal probability distribution to estimate probabilities and identify unusual events.

We now know that the empirical rule gives probabilities for values that lie exactly 1, 2, and 3 standard deviations away from the mean. But how do we determine the probability that a value lies some fraction of a standard deviation away from the mean? In this situation, we will use technology to find the probability.

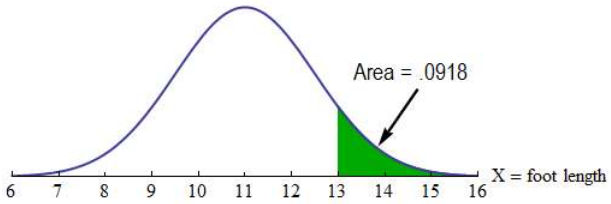
For now, we use a simulation to find the probability based on a *z*-score. Statistical packages can also be used to find probabilities associated with a normal curve.

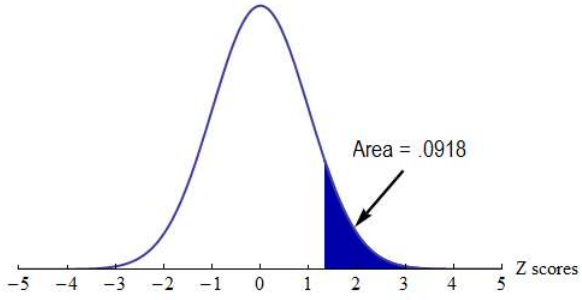
## Example

## The Distribution of z-Scores

Recall that we use the area under a normal density curve to find a probability. If we convert the *x*-values into *z*-scores, the distribution of *z*-scores is also a normal density curve. This curve is called the **standard normal distribution**.

Here we compare the normal density curve for the foot lengths to the standard normal curve:




The normal curve pictured on top is the model for the distribution of foot lengths. Note that the values on the axis are foot lengths. The distribution has a mean of 11 inches and a standard deviation of 1.5 inches. A foot length of 13 inches is marked. The shaded area is 0.0918. This is the probability that a randomly selected male will have a foot length greater than 13 inches: $P(X > 13) = 0.0918$.

The normal curve pictured on bottom is the standard normal distribution. This represents the distribution of *z*-scores. Note that the values on the axis are *z*-scores. The mean is 0 and the standard deviation is 1. The *z*-score corresponding to $X = 13$ inches is marked.

$$Z=\frac{x-\mathrm{\mu}}{\mathrm{\sigma}}=\frac{13-11}{1.5}=1.33$$

The shaded area here is the same, 0.0918. This is the probability that a *z*-score is greater than 1.33: $P(Z > 1.33) = 0.0918$.

Here is the main idea: Since the areas are the same, *we use the standard normal curve to find the probabilities associated with any normal density curve*.

Note: The standard normal distribution *always* has a mean = 0 and a standard deviation = 1. To understand this, recall that a *z*-score is the number of standard deviations *X* is above (or below) the mean.

- When the *x*-value is the mean, the *z*-score is 0.

- We can illustrate this for the foot lengths: If $X$ is the mean, then

$$\begin{array}{l}X=11\\ Z=\frac{x-\mathrm{\mu }}{\mathrm{\sigma }}=\frac{11-11}{1.5}=\frac{0}{1.5}=0\end{array}$$

- When $X$ is 1 standard deviation above the mean, the $z$-score is 1.
- We can illustrate this for the foot lengths: If $x$-value is 1 standard deviation above the mean, $X = 11 + 1.5 = 12.5$.

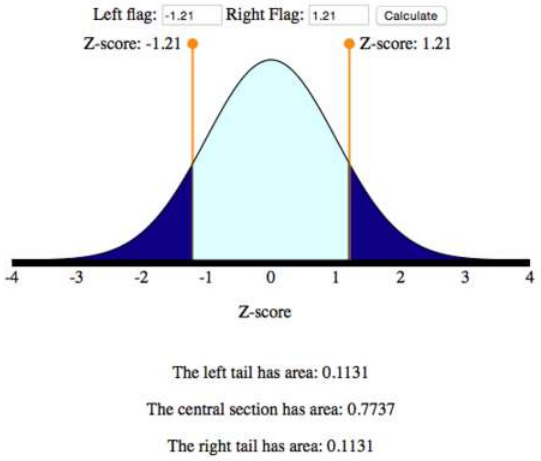
$$Z=\frac{x-\mathrm{\mu }}{\mathrm{\sigma }}=\frac{12.5-11}{1.5}=\frac{1.5}{1.5}=1$$

- Similarly, the $z$-score for the $x$-value that is 1 standard deviation below the mean is $-1$.

### Example

## Using the Standard Normal Curve to Find Probabilities

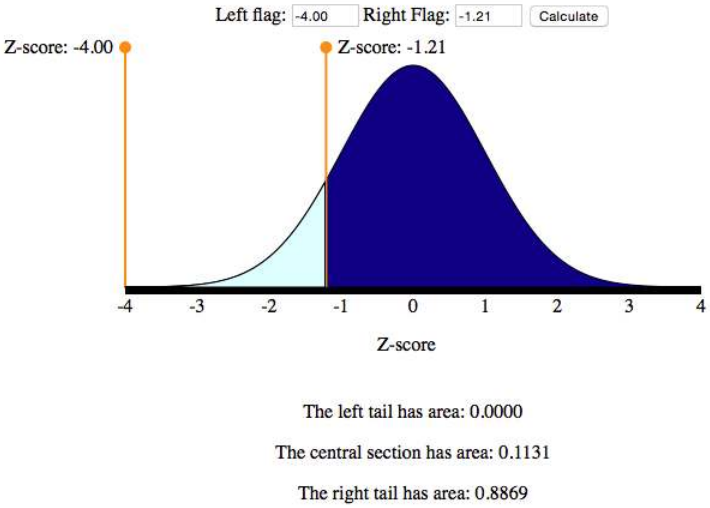
We use a simulation based on the standard normal distribution to find probabilities. In this simulation, the numbers on the *horizontal* axis are $z$-scores. The areas are rounded to four decimal places, so these areas are not exact values.




Use this image from the simulation to find the probabilities:

- Find $P(-1.21 < Z < 1.21)$. In words, we want to find the probability that the $z$-score is between $-1.21$ and 1.21. This probability is the light blue area. So the answer is 0.7737.
- Find $P(Z > 1.21)$. In words, we want to find the probability that the $z$-score is greater than 1.21. This probability is the blue area to the right of $Z = 1.21$. So the answer is 0.1131.
- Find $P(Z < -1.21)$. In words, we want to find the probability that the $z$-score is less than $-1.21$. This probability is the dark blue area to the left of $Z = -1.21$. So the answer is 0.1131. Note that this is the same as $P(Z > 1.21)$ because of the symmetry in the normal distribution. Both tails have the same area.
- Find $P(Z > -1.21)$. In words, we want to find the probability that the $z$-score is greater than $-1.21$. This probability is the area to the right of $-1.21$. This is the sum of the green area and the dark blue area. So the answer is 0.7737 + 0.1131 = 0.8868.

Note: Here is another way to use the simulation to find $P(Z > -1.21)$. Here we moved one slider as far to the left as possible, then located the other slider at $Z = -1.21$.




Notice that the area is 0.8869, not 0.8868. This discrepancy is due to rounding. The simulation uses a mathematical model to find the areas. These areas are rounded to four decimal places. Don't worry about this small difference. Either answer is acceptable.

We now practice using the simulation based on the standard normal curve to find probabilities. Later we use this same simulation to find probabilities for any normal distribution.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=256

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3568

https://assessments.lumenlearning.co...sessments/3569

Click here to open the above simulation in its own window.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3570

https://assessments.lumenlearning.co...sessments/3571

https://assessments.lumenlearning.co...sessments/3572

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 6.23: Normal Random Variables (5 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.24: Normal Random Variables (6 of 6)

## Learning Objectives

- Use a normal probability distribution to estimate probabilities and identify unusual events.

Now we use the simulation and the standard normal curve to find the probabilities associated with any normal density curve.

## Example

## Length of Human Pregnancy

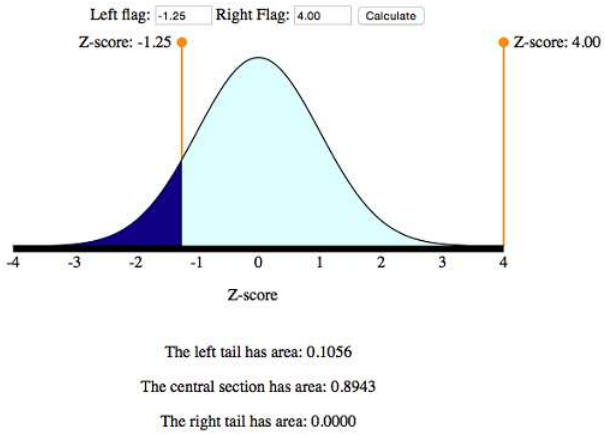
The length (in days) of a randomly chosen human pregnancy is a normal random variable with μ = 266, σ = 16. So $X$ = length of pregnancy (in days)

**(a)** What is the probability that a randomly chosen pregnancy will last less than 246 days?

We want $P(X < 246)$. To find this probability, we first convert $X = 246$ to a $z$-score:

$$Z=\frac{246-266}{16}=\frac{-20}{16}=-1.25$$

Now we can use the simulation to find $P(Z < -1.25)$. This is the area under the normal probability curve to the *left* of $Z = -1.25$.




The probability that a randomly chosen pregnancy lasts less than 246 days is 0.1056. In other words, there is an 11% chance that a randomly selected pregnancy will last less than 246 days.

**(b)** Suppose a pregnant woman's husband has scheduled his business trips so that he will be in town between the 235th and 295th days of her pregnancy. What is the probability that the birth will take place during that time?

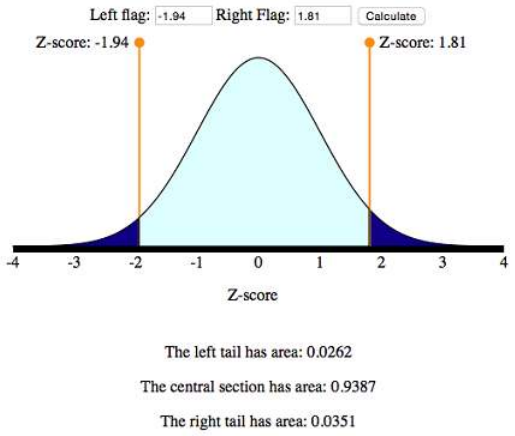
Compute the $z$-scores for each of these $x$-values:

$$Z=\frac{235-266}{16}=\frac{-31}{16}=-1.94$$

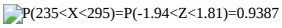
and

$$Z=\frac{295-266}{16}=\frac{29}{16}=1.81$$

Use the simulation to find the area under the standard normal curve *between* these two $z$-scores.




So the desired probability is 0.9387.



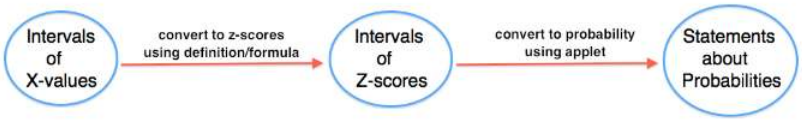

There is about a 94% probability that he will be home for the birth. Looks like he planned well.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3573

The previous examples all followed the same general form: Given values of a normal random variable, we found an associated probability. The two basic steps in the solution process were as follows:

1. Convert *x*-value to a *z*-score.
2. Use the simulation to find associated probability.




The next example is a different type of problem: Given a probability, we will find the associated value of the normal random variable. The solution process will go in reverse order.

1. Use a new simulation to convert statements about probabilities to statements about *z*-scores.
2. Convert *z*-scores to *x*-values.

These types of problems are informally called “work-backwards” problems. We will use a new simulation for these types of problems. The new simulation requires us to enter a probability and then gives us the associated z-score. This is backwards from the simulation we worked with previously where we entered a z-score to find a probability. We will use this simulation in the next example.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=258

### Example

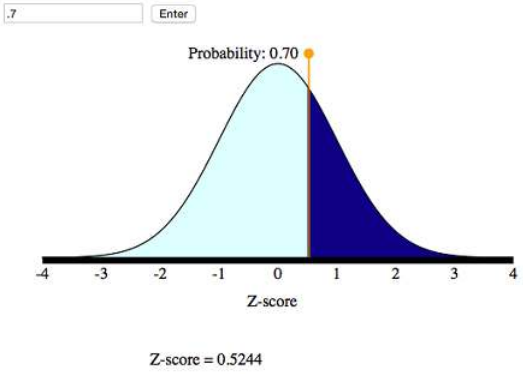
## Work Backwards to Find X

Foot length (in inches) of a randomly chosen adult male is a normal random variable with a mean of 11 and standard deviation of 1.5. So $X$ = foot length (inches).

**(a)** Suppose that an XL sock is designed to fit the largest 30% of men's feet. What is the smallest foot length that fits an XL sock?

**Step 1:** Use the simulation to convert the probability to a statement about *z*-scores.

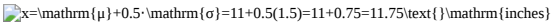
We want to mark off the largest 30% of the distribution, so the probability to the *right* of the *z*-score is 30%. This means that 70% of the area is to the left of the *z*-score.




From the simulation, we can see that the corresponding *z*-score is 0.52.

**Step 2:** Now we need to convert this *z*-score to a foot length.

Before we calculate the length, note that the *z*-score is about 0.5, so the *x*-value will be about 0.5 standard deviations above the mean.



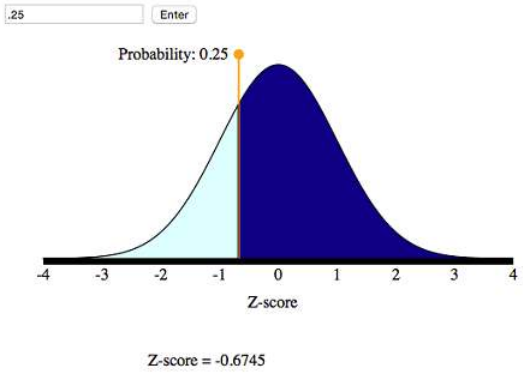
$$x=\mu+0.5\cdot\sigma=11+0.5(1.5)=11+0.75=11.75\text{ inches}$$

**Conclusion:** A foot length of 11.75 inches is the shortest foot for an XL sock.

**(b)** What is the first quartile for the men's foot lengths?

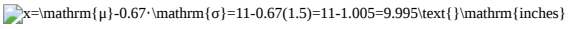
**Step 1:** Use the simulation to convert this probability into a statement about *z*-scores.

We want to mark off the smallest 25% of the distribution, so the probability to the *left* of the *z*-score is 25%.




From the simulation, we can see that the corresponding *z*-score is −0.67.

**Step 2:** Convert this *z*-score to a foot length. If *X* is the foot length we seek, then *X* is 0.67 standard deviations *below* the mean. That is,



$$x=\mu-0.67\cdot\sigma=11-0.67(1.5)=11-1.005=9.995\text{ inches}$$

**Conclusion:** The first quartile mark is 9.995 inches, so about 25% of the men's feet are shorter than 10 inches.

## Comments

In the preceding example (specifically step 2), we found the *x*-value by reasoning about the meaning of the *z*-score. We can also develop a formula for this process.

Recall the definition of *z*-score. In words, the *z*-score of an *x*-value is the number of standard deviations *X* is away from the mean. As a formula, this is

$$Z=\frac{x-\mathrm{\mu }}{\mathrm{\sigma }}$$

We can solve this equation for X as follows:

$$\begin{array}{l}\frac{x-\mathrm{\mu }}{\mathrm{\sigma }}=Z\\ x-\mathrm{\mu }=Z\cdot \mathrm{\sigma }\\ x=\mathrm{\mu }+Z\cdot \mathrm{\sigma }\end{array}$$

This gives us a formula for finding *X* from *Z*. You can use this formula in step 2 of a work-backwards problem.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3888

## Let's Summarize

- In "Continuous Random Variables," we made the transition from discrete to continuous random variables. A continuous random variable is not limited to distinct values. It is a measurement such as foot length. We cannot display the probability distribution for a continuous random variable with a table or histogram. We use a density curve to assign probabilities to intervals of *x*-values. We use the *area under the density curve to find probabilities*.
- We use a *normal density curve* to model the probability distribution for many variables, such as weight, shoe sizes, foot lengths, and other human physical characteristics. Normal curves are mathematical models. We use **μ** to represent the mean of a normal curve and **σ** to represent the standard deviation of a normal curve. We use Greek letters to remind us that the normal curve is not a distribution of real data. It is a mathematical model based on a mathematical equation. We use this mathematical model to represent the perfect bell-shaped distribution.
- For a normal curve, the *empirical rule for normal curves* tells us that 68% of the observations fall within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations of the mean.
- To compare *x*-values from different distributions, we standardize the values by finding a *z*-score: $Z=\frac{x-\mathrm{\mu }}{\mathrm{\sigma }}$
- A *z*-score measures how far *X* is from the mean in standard deviations. In other words, the *z*-score is the number of standard deviations *X* is from the mean of the distribution. For example, $Z = 1$ means the *x*-value is 1 standard deviation above the mean.
- If we convert the *x*-values into *z*-scores, the distribution of *z*-scores is also a normal density curve. This curve is called the **standard normal distribution**. We use a simulation with the standard normal curve to find probabilities for any normal distribution.
- We can also work backwards and find the *x*-value for a given probability. We used a different simulation to work backwards from probabilities to *x*-values. With this simulation, we found *x*-values corresponding to quartiles and percentiles.

## Are You Ready for the Checkpoint?

If you completed all of the exercises in this module, you should be ready for the Checkpoint. To make sure that you are ready for the Checkpoint, use the My Response link below to evaluate your understanding of the learning objectives for this module and to submit questions that you may have.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.24: Normal Random Variables (6 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 6.25: Putting It Together- Probability and Probability Distribution

## Let's Summarize

Here is a summary of the key concepts developed in this module:

- The *probability* of an event is a measure of the likelihood that the event occurs. Probabilities are always between 0 and 1. The closer the probability is to 0, the less likely the event is to occur. The closer the probability is to 1, the more likely the event is to occur.
- The two ways of determining probabilities are empirical and theoretical.
  - *Empirical* methods are based on data. The probability of an event is approximated by the relative frequency of the event.
  - *Theoretical* methods use the nature of the situation to determine probabilities.
- Following are some common probability rules:
  - $P(\text{not } A) = 1 - P(A)$.
  - When two events have no outcomes in common, they are disjoint. If A and B are disjoint events, $P(A \text{ or } B) = P(A) + P(B)$.
  - When the knowledge of the occurrence of one event $A$ does not affect the probability of another event $B$, we say the events are independent. If $A$ and $B$ are **independent events**, $P(A \text{ and } B) = P(A) \cdot P(B)$.
- When we have a quantitative variable with outcomes that occur as a result of some random process (e.g., rolling a die, choosing a person at random), we call it a *random variable*. There are two types of random variables:
  - *Discrete* random variables have numeric values that can be listed and often can be counted. We find probabilities using areas in a probability histogram.
  - *Continuous* random variables can take any value in an interval and are often measurements. We use a density curve to assign probabilities to intervals of $x$-values. We use the *area under the density curve to find probabilities*.
- We use a *normal density curve* to model the probability distribution for many variables, such as weight, shoe sizes, foot lengths, and other physical characteristics. For a normal curve, the empirical rule for normal curves tells us that 68% of the observations fall within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations of the mean.
- To compare $x$-values from different distributions, we standardize the values by finding a $z$-score: $Z=\frac{x-\mathrm{\mu}}{\mathrm{\sigma}}$
- A $z$-score measures how far $X$ is from the mean in standard deviations. In other words, the $z$-score is the number of standard deviations $X$ is from the mean of the distribution. For example, $Z = 1$ means the $x$-value is one standard deviation above the mean.
- If we convert the $x$-values into $z$-scores, the distribution of $z$-scores is also a normal density curve. This curve is called the *standard normal distribution*. We use a simulation with the standard normal curve to find probabilities for any normal distribution.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 6.25: Putting It Together- Probability and Probability Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 7: Linking Probability to Statistical Inference


7.1: Why It Matters- Linking Probability to Statistical Inference
7.2: Parameters vs. Statistics
7.3: Introduction- Statistical Inference
7.4: Distribution of Sample Proportions (1 of 6)
7.5: Distribution of Sample Proportions (2 of 6)
7.6: Distribution of Sample Proportions (3 of 6)
7.7: Distribution of Sample Proportions (4 of 6)
7.8: Distribution of Sample Proportions (5 of 6)
7.9: Distribution of Sample Proportions (6 of 6)
7.10: Introduction- Distribution of Sample Proportions
7.11: Statistical Inference (1 of 3)
7.12: Statistical Inference (2 of 3)
7.13: Statistical Inference (3 of 3)
7.14: Putting It Together- Linking Probability to Statistical Inference



This page titled 7: Linking Probability to Statistical Inference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.1: Why It Matters- Linking Probability to Statistical Inference

This module introduces our study of inference. Before we begin *Linking Probability to Statistical Inference*, let's look at how the remainder of the course relates to the Big Picture of Statistics.

Recall that we start a statistical investigation with a research question. The investigation proceeds with the following steps:

- Produce Data: Determine what to measure, then collect the data. ← **Types of Statistical Studies and Producing Data**
- Explore the Data: Analyze and summarize the data. ← **Summarizing Data Graphically and Numerically, Examining Relationships: Quantitative Data, Nonlinear Models, Relationships in Categorical Data with Intro to Probability**
- Draw a Conclusion: Use the data, probability, and statistical inference to draw a conclusion about the population. ← **Relationships in Categorical Data with Intro to Probability, Probability and Probability Distributions,Linking Probability to Statistical Inference,Inference for One Proportion, Inference for Two Proportions, Inference for Means, Chi-Square Tests**

In the Big Picture of Statistics, we are about to start the last step: Inference. We use data from a sample to "infer" something about the population in this and the upcoming modules. Inference is based on probability.

### Example

At the end of April 2005, *ABC News* and the *Washington Post* conducted a poll to determine the percentage of U.S. adults who support the death penalty.

**Research question:** *What percentage of U.S. adults support the death penalty?*

Steps in the statistical investigation:

1. **Produce Data:** Determine what to measure, then collect the data.The poll selected 1,082 U.S. adults at random. Each adult answered this question: "Do you favor or oppose the death penalty for a person convicted of murder?"
2. **Explore the Data:** Analyze and summarize the data.In the sample, 65% favored the death penalty.
3. **Draw a Conclusion:** Use the data, probability, and statistical inference to draw a conclusion about the population.

Our goal is to determine the percentage of the U.S. adult population that support the death penalty. We know that different samples give different results. What are the chances that a sample reflects the opinions of the population within 3%? Probability describes the likelihood that a sample is this accurate, so we can say with 95% confidence that between 62% and 68% of the population favor the death penalty.

We illustrated the Big Picture of Statistics with an example about an inference made from survey data. From a random sample of U.S. adults, we estimated the percentage of all U.S. adults who support the death penalty. We saw that probability describes the likelihood that an estimate is within 3% of the true percentage with this opinion in the population. For this example, there is a 95% chance that a random sample is within 3% of the true population percentage.

Because random samples vary, inference always involves uncertainty. This uncertainty is captured by probability statements that are part of our conclusions. We emphasize this point in the following example where we look more closely at the process of statistical investigation with a real court case. Our goal is to identify arguments that use statistical inference to draw a conclusion as well as arguments that do not use inference.

### EXAMPLE

## Using Inference to Detect Cheating – A Real Case

How can a prosecutor use data to detect cheating? The details of this case appeared in *Chance* magazine in 1991.

**The Case:** During an exam at a university in Florida in 1984, the proctor suspected that one student, whom we will call Student C, was copying answers from another student, whom we will call Student A. The proctor accused Student C of cheating, and the case went to the university's supreme court.

**The Evidence:** At the trial, the prosecution introduced evidence based on data. Here is the evidence: On the 16 questions missed by both Student A and Student C, 13 of the answers were the same.

**The Argument:** The prosecutor used the data to draw an inference based on probability. He asked the question: Could 13 out of 16 matches be due to chance? He argued that a match of 13 out of 16 by chance alone was very unlikely. The probability of this occurring is very small. So there had to be another explanation besides chance, and the prosecutor said the explanation was cheating. Based on this evidence, Student C was found guilty of academic dishonesty.

**The Role of Random Chance:** To decide if we agree with this argument, we need to understand if chance might explain this result. We need to determine if it would be unusual to get 13 matches on 16 questions by chance alone. To determine if 13 out of 16 is unusual, we have to look at what happens in the long run if students just guess on 16 multiple choice questions.

Let's assume that each question had four options: a, b, c, or d. We use a computer program to randomly assign answers to each question, which mimics what happens when someone randomly guesses. Using software to imitate chance behavior is called **simulation**.

Here you see a representation of answers from Student A and Student C as well as three randomly generated answer sets for the 16 questions missed by both Student A and Student C. We highlighted matches with Student A's answers in green.

Notice that Student C has 13 answers that match Student A's answers: As a proportion, this is $13 \div 16 = 0.8125$. It means that about 81% of the time, Student C's answers matched Student A's answers on questions that they both missed.

For the random answers generated by "guessing," we see the proportion of matches differs:

- Set 1 has 6 matches with Student A: $6 \div 16 = 0.375$ as a proportion, or about 38%.
- Set 2 has 3 matches with Student A: $3 \div 16 = 0.1875$ as a proportion, or about 19%.
- Set 3 has 5 matches with Student A: $5 \div 16 = 0.3125$ as a proportion, or about 31%.

The proportion of matches in the randomly generated answers vary quite a bit, from 0.1875 to 0.3125. But none are close to the proportion of matches seen in Student C's paper, 0.8175. The proportion of matches for the three sets of results from the random guesses are graphed in the following figure. The blue line is the proportion of matches for Student C.

Now we repeat the simulation of random sets of answers on the 16 questions, each time determining the proportion of matches to the wrong answers on Student A's paper. On the left is a dotplot for 100 random sets of answers. Each dot represents a set of 16 answers generated by random guessing. On the right is a histogram of 1,025 random sets of answers, which shows the long-run pattern.

**Analysis:** In the histogram, we see that typical results fall between about 0.1 and 0.4. More specifically, if a student randomly guessed on these 16 questions, it would not be surprising to see from 2 to 6 matches with Student A's wrong answers. Translated into proportions, this is $2 \div 16 = 0.125$ to $6 \div 16 = 0.375$.

Is Student C unusual? Yes. Notice that no randomly generated set of answers comes close to the proportion of matches on Student C's exam. A proportion of 0.8125 would be very unusual if guessing, so we conclude that Student C was not guessing.

**Conclusion:** The prosecution argued that a match of 13 out of 16 wrong answers (a proportion of 0.8125) was unusual and could not be explained by random chance. Our simulation agrees with this observation. When we created answer sets by randomly guessing, we never saw more than 9 matches out of 16, which is a proportion of 0.5625. We agree with the prosecution that there has to be another explanation besides chance.

*However, could there be another explanation besides cheating?*

Yes. If you don't know the answer to a question on an exam, you rarely guess at random. It is more likely that you will make an educated guess. Some wrong answers might be more logical than others. This could also explain the large proportion of matches on wrong answers between the two students. So this evidence is not convincing evidence that Student C cheated, but we know that he did not just guess.

EXAMPLE

## Follow-up Argument Based on Exploratory Data Analysis

Student C appealed his case. A second trial was held. This time the prosecution made a different argument using data. This argument did not use statistical inference.

The prosecution created a new measurement. They compared every student's paper to Student A's paper. For the 40-multiple choice questions on the test, they counted the number of matches to Student A's paper and divided by 40. This new measurement was also a proportion.

**Analysis:** There were 88 students. Here are the results:

Each dot represents a student who took the exam. A dot with a proportion of 0.6 means that 60% of this student's answers matched Student A's answers.

Note: Student A is included in the data. Of course, Student A is the dot with the proportion of 40 out of 40 = 1.0 This makes sense because all of Student A's answers matched Student A. (This does not mean that Student A did well on the exam.)

We see that many of the proportions fall between between 0.40 (which is 16 matches out of 40) and 0.60 (which is 24 matches out of 40). So it is not surprising if between 40% and 60% of a student's answers matched Student A's answers.

Student C had 32 matches out of 40, which is a proportion of 0.80. (This does not mean that Student C made an 80% on the exam. It means that 80% of Student C's answers matched Student A's answers.) Student C is once again an unusual data point.

**Conclusion:** This time it is harder to argue that Student C is not cheating. When we compare him to the rest of the class, his paper had an unusually high number of matches with Student A's answers. This data together with the proctor's testimony is fairly convincing evidence for the prosecution's claim that Student C cheated by copying from Student A's paper.

*(Source: Phillip J. Boland and Michael Proschan, "The Use of Statistical Evidence in Allegations of Exam Cheating," Chance 3(3):10-14, 1991.)*

### What's the Main Point?

Statistical inference always involves an argument based on probability.

In this court case, the prosecution used two different types of arguments to provide evidence of cheating. The first argument is an example of statistical inference because it is based on probability. We set up a simulation to reflect an assumption that the prosecutor made. The assumption is that answer sets come from random guessing. We simulated over a thousand answer sets with randomly chosen answers to investigate the long-run behavior of simulated answer sets. We then compared Student C to the distribution of randomly generated answer sets. Student C was unusual. We concluded that Student C was not randomly guessing.

The second argument is an example of exploratory data analysis with no statistical inference. The prosecutor designed a measurement and collected data from every student in the class. He compared Student C's measurement to the measurements of the other students. Probability did not have a role in this analysis. Probability statements require a random event and a look at long-run behavior of random events, so this is not an example of statistical inference.

### LEARN BY DOING

https://assessments.lumenlearning.co...sessments/3895

### LEARN BY DOING

https://assessments.lumenlearning.co...sessments/3958

The court case illustrates how we can view statistical inference as an argument based on probability. Here we briefly connect the probability argument with the vocabulary and ideas from the module *Probability and Probability Distribution*.

Recall the following important points about probability that we learned in that module:

- Probability is a measure of how likely an event is to occur.
- We can make probability statements only about random events. *Random* here means that the outcome is uncertain in the short run but has a predictable pattern in the long run.

*How does the logic of the probability argument in the court case relate to Probability and Probability Distribution?*

To understand the probability argument in the original court case, we used a computer simulation to analyze the long-run pattern that emerges if students randomly guess on multiple-choice questions. Our variable was *Proportion of matches with Student A's wrong answers on 16 questions*. We graphed the distribution of a large number of proportions from the random answer sets to see

the pattern. Using the vocabulary of *Probability and Probability Distribution*, the proportion of matches is a *random variable*. In the short run, we do not know the proportion of matches that will occur in a random answer set, but in the long run, we see a pattern in the distribution of these proportions. This distribution of proportions is treated as a probability model. From it we can see how much variability to expect in matches from random answer sets. We can also identify unusual values. From the pattern in this distribution, we can say it is very unlikely that Student C guessed.

*How does this logic relate to the Big Picture?*

This logic is similar to the logic of inference that we explore in future modules, where we randomly select samples from a population. We continue to use computer simulation throughout these modules to analyze the long-run pattern that emerges in measurements from random samples. We create probability models that tell us how much variability to expect in random samples. We also use these models to identify unusual measurements. From the patterns, we use data to make judgments about the population. In this way, our conclusions about a population are based on probability.

*Probability and Probability Distribution* included a long discussion of probability models that are normal curves. Under certain conditions, the long-run behavior of measurements from random samples can be modeled with a normal curve (or a similar curve), we see in this module and in *Inference for One Proportion*, *Inference for Two Proportions*, and *Inference for Means*. In each module, we ask the question, *When can we use a normal model?* Once we know these conditions, we can use what we learned in the previous module to make probability-based decisions about population values.

Here we add these ideas to the Big Picture to show how probability connects to inference.

Note that we highlighted two types of inference in the diagram:

- **Estimate a population value.**
- **Test a claim about the population value.**

We end our introduction to inference with a look at research questions that illustrate these two types of inference. We also connect these examples to the types of inference we learn about in upcoming modules.

### Research Questions That Involve Inference

| Type of Question | Examples | Variable Type | Unit |
|---|---|---|---|
| Make an estimate about the population | What proportion of all U.S. adults support the death penalty? | Categorical variable | Inference for One Proportion |
| | What is the average number of hours that community college students work each week? | Quantitative variable | Inference for Means |
| Test a claim about the population | Do the majority of community college students qualify for federal student loans? | Categorical variable | Inference for One Proportion |
| | Has the average birth weight in a town decreased from 3,500 grams? | Quantitative variable | Inference for Means |
| Compare two populations | Are teenage girls more likely to suffer from depression than teenage boys? | Categorical variable | Inference for Two Proportions |
| | In community colleges do female students have a higher average GPA than male students? | Quantitative variable | Inference for Means |

Note: Each research question relates to either a categorical variable or a quantitative variable. In this course, three criteria determine the inference procedure we use:

- The type of variable.
- The type of inference (estimate a population value or test a claim about a population value).
- The number of populations involved.

## LEARN BY DOING

https://assessments.lumenlearning.co...sessments/3959

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.1: Why It Matters- Linking Probability to Statistical Inference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.2: Parameters vs. Statistics

## Learning Objectives

- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.
- Distinguish between a sample statistic and a population parameter.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.2: Parameters vs. Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.3: Introduction- Statistical Inference

## What you'll learn to do: Find a confidence interval to estimate a population proportion and test a hypothesis about a population proportion using a simulated sampling distribution or a normal model of the sampling distribution.

### LEARNING OBJECTIVES

- Find a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.
- Test a hypothesis about a population proportion using a simulated sampling distribution or a normal model of the sampling distribution. State a conclusion in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.3: Introduction- Statistical Inference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.4: Distribution of Sample Proportions (1 of 6)

## Learning Objectives

- Introduction

  **After we develop a probability model of how sample proportions behave, we can answer questions like the following:**

  - The questions ask us to make an inference about a population. Our answers to these questions will be based on a sample. We will never be 100% sure of our answer, so we will make probability statements that describe the strength of the evidence and our certainty.

    **Brief Discussion of the Connection between These Questions and Probability**

    - What proportion of all college students in the United States are enrolled at a community college?
      - Note: Connected to each inference question about a population proportion, we see a probability question about the long-run behavior of sample proportions. We need to understand how proportions from random samples relate to the population proportion. We also need to understand how much variability we can expect in sample proportions. Therefore, in our early investigations, we will assume we know a population proportion and examine what happens when we select random samples from this population.

        Now we begin an investigation of the long-run behavior of sample proportions.

        ### Example

        ### Gender in the Population of Part-time College Students

        We don't have information about the population of part-time college students, so we select a random sample of 25 part-time college students and calculate the proportion of the sample that is female. We don't expect the sample proportion to be exactly 0.60. So, *how much could the sample proportion vary from 0.60 for us to feel confident in our prediction?*

        To answer this question, we need to understand how much sample proportions will vary if the parameter is 0.60.

        ### Learn By Doing

        Refer to the previous example for the following questions. These questions focus on how the proportion of females will vary in random samples if we assume that 0.60 of the population of part-time college students is female.

        https://assessments.lumenlearning.co...sessments/3574

        https://assessments.lumenlearning.co...sessments/3575

        Use the following simulation to select a random sample of 25 part-time college students. Repeat the selection many times to observe how the proportion of females in the samples vary. Then answer the following question.

        An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=270

        ### Learn By Doing

        https://assessments.lumenlearning.co...sessments/3585

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.4: Distribution of Sample Proportions (1 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.5: Distribution of Sample Proportions (2 of 6)

## Learning Objectives

- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.

Recall that our goal is to create a probability model that describes the long-run behavior of proportions from random samples. Previously, we used a simulation to collect a few random samples to get acquainted with making a distribution of sample proportions. We randomly selected 25 students at a time from a population of part-time college students that is 60% female. In the next example, we predict what happens in the long run when we select many, many random samples of 25 students at a time from this population. Then we watch a simulation to see if our predictions are correct.

## Example

## Predicting the Behavior of Sample Proportions

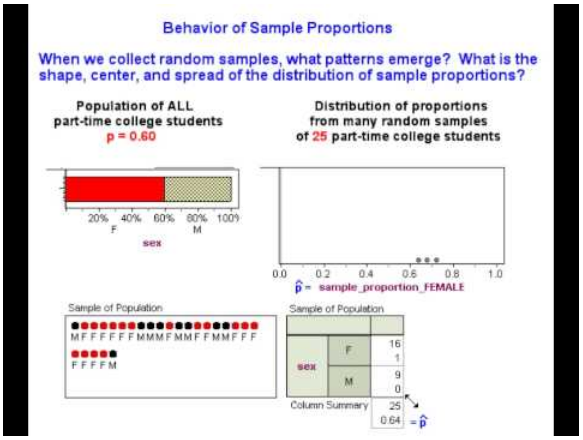
Based on our intuition and what we observed with the simulation, we might expect the following about the distribution of sample proportions that come from a population where $p = 0.60$:

**Center:** Some sample proportions will be on the low side – such as 0.52 or 0.56 – and others will be on the high side – such as 0.64 or 0.68. It is reasonable to expect all the sample proportions in repeated random samples to average out to the underlying population proportion, 0.6. In other words, the mean of the distribution of sample proportions should be $p$.

**Spread:** For samples of 25, we expect sample proportions of females not to stray too far from the population proportion 0.6. Sample proportions lower than 0.44 or higher than 0.72 will be unusual. Previously, we took smaller random samples of 8 and observed more variability in the sample proportions. We therefore think that sample size plays a role in the spread of the distribution of sample proportions. Smaller samples may be less accurate and more variable than larger samples.

**Shape:** Sample proportions closest to 0.6 will be most common, and sample proportions far from 0.6 in either direction will be progressively less likely. In other words, the shape of the distribution of sample proportions may be somewhat bell-shaped.

Now we use a simulation to collect numerous samples to see what happens in the long run. We use the simulation to check whether our intuition about the center, spread, and shape of the distribution of sample proportions is right.




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=272

## Learn By Doing

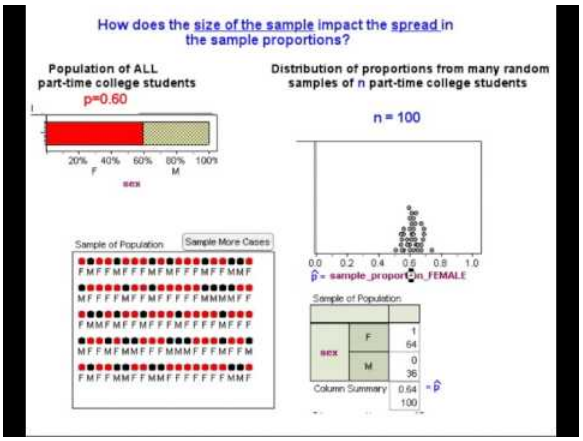
https://assessments.lumenlearning.co...sessments/3576

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3889

At this point, we have a good sense of what happens as we take random samples from a population. Our simulation suggests that our initial intuition about the shape and center of the distribution of sample proportions is correct.

Now we use another simulation to help us think more precisely about the variability we expect to see in the sample proportions. Our intuition tells us that larger samples will better approximate the population, so we might expect less variability in large samples. In the next walk-through, we use a simulation to investigate this idea. After that walk-through, we tie these ideas to more formal theory about the probability model for the long-run behavior of proportions from random samples.




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=272

We are now ready to use what we have observed to develop a formal probability model to describe the behavior of sample proportions. First, let's return to the original question that prompted our investigation of sample proportions.

We based our investigation on the prediction that 60% of part-time college students will be female. In our investigation, we asked, *How much could the sample proportion vary from a population proportion of 0.60 for us to feel confident in our prediction?*

We don't expect a sample proportion to be exactly equal to the population proportion. But how much error seems reasonable?

We now see that the answer to this question depends on the size of the sample.

- If we select a random sample of 25 students, the distribution of sample proportions has a standard deviation of about 0.1. We can see that most sample proportions fall within 2 standard deviations of 0.60. Therefore, we might decide that $2 \times 0.10 = 0.20$ is a reasonable margin of error, so a sample proportion between 0.40 and 0.80 is not surprising if 0.60 of all part-time college students are female.
- If we select a random sample of 100 students, the distribution of sample proportions has less variability. It has a standard deviation of about 0.05. Again we see that most sample proportions fall within 2 standard deviations of 0.60, so we might decide that $2 \times 0.05 = 0.10$ is a reasonable amount of error for these larger samples. For a sample of 100 students, then, a sample proportion between 0.50 and 0.70 is not surprising if 0.60 of all part-time college students are female.

We discuss this idea further in "Introduction to Statistical Inference" in this unit.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.5: Distribution of Sample Proportions (2 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.6: Distribution of Sample Proportions (3 of 6)

## Learning Objectives

- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.

Now we practice using a simulation to examine how sample proportions relate to a population proportion and to identify unusual sample values. The type of thinking we do here prepares us for the type of thinking we will do in statistical inference.

## Example

### Community College Enrollment

According to a report by the Pew Research Center, in October 2007 about 10% of the 3.1 million 18- to 24-year-olds in the United States were enrolled in a community college. Suppose in that year we randomly selected 100 young adults in this age group. Suppose 15% of the sample was enrolled in a community college. Is this surprising? Well, the sample proportion is off by only 5%. But how much error do we expect to see in random samples of this size? We do a simulation to find out.

**Simulation:**

First, we make an assumption about the population proportion. We set $p$ = 0.10 in the simulation. (If you would like to work through the example using the simulation, click here). We also set n=100 to represent the sample size of 100. When we hit the "Run simulation (5,000 samples)" button, the simulation simulates the random selection of 5,000 samples. Each sample has 100 young adults from this population. For each sample, the simulation plots the proportion who are enrolled at a community college. Here is a histogram of the results.

**Analysis:**

When p=0.10 and n=100, a sample proportion of 0.15 is too far away from 0.10 to be considered a typical sample result. It is not part of the central peak of the histogram of sample proportions, but it is also not in the small part of the histogram's tail. Therefore, this result is somewhat unusual, but not extremely unusual. In other words, the 5% error in this sample is larger than the error we see in most samples, but there are samples with larger amounts of error.

**Conclusion:**

For random samples of 100 young adults, a sample with 15% enrolled in a community college is unusual if only 10% of the population overall is enrolled.

## Example

### Another Look at Community College Enrollment

Here we think about a more precise way to analyze the results of our simulation. (If you would like to work through the example using the simulation, click here). We use the standard deviation of the distribution of sample proportions to describe the amount of error we expect to see in random samples. We use the simulation again and check "Show standard deviation bar."

The mean of the sample proportions is p=0.10. The standard deviation of the sample proportions is 0.03. The standard deviation describes the average amount of error in sample proportions that is due to chance. On average, sample proportions will have a 3% error. A sample proportion of 0.15 has a 5% error, so this is a larger error than we expect on average.

Here is another way to look at it. Typical samples have sample proportions within 1 standard deviation of 0.10, which is between 0.07 and 0.13 (just subtract 0.03 from 0.10 and then add 0.03 to 0.10). We can also see that most sample proportions fall within about 2 standard deviations of 0.10, which is between 0.04 and 0.16. So it is extremely unusual for sample proportions to have values outside of this range.

Therefore, a sample proportion of 0.15 is not typical, but it is also not extremely unusual, when sampling from a population with p=0.10.

Learn By Doing

https://assessments.lumenlearning.co...sessments/3890

Learn By Doing

https://assessments.lumenlearning.co...sessments/3577

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 7.6: Distribution of Sample Proportions (3 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.7: Distribution of Sample Proportions (4 of 6)

## Learning Objectives

- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.

The simulations on the previous page reinforce what we have observed about patterns in random sampling.

- Proportions from random samples approximate the population proportion, *p*, so sample proportions average out to the population proportion.
- Larger random samples better approximate the population proportion, so large samples have sample proportions closer to *p*. In other words, a sampling distribution for large samples has less variability.
- The distribution of sample proportions appears normal (at least for the examples we have investigated).

We can describe the sampling distribution with a mathematical model that has these same features.

## The Sampling Distribution of Sample Proportions

For a categorical variable, imagine a population with a proportion *p* of successes. (For example, for the variable gender, imagine a population of part-time college students with $p = 0.60$ female. Note that a *success* is the category of interest. It is what we are counting. Here a success is a female.) We create a mathematical model that describes the sample proportions from all possible random samples of size *n* from this population. The model has the following center, spread, and shape.

**Center:** Mean of the sample proportions is *p*, the population proportion.

**Spread:** Standard deviation of the sample proportions is $\sqrt{\frac{p(1-p)}{n}}$. The standard deviation of the sampling distribution is also called the **standard error**.

**Shape:** A normal model is a good fit if the expected number of successes and failures is at least 10. We can translate these conditions into formulas: $np \geq 10 \text{ }\mathrm{and}\text{ } n(1-p) \geq 10.$

## Comment

The distribution of sample proportions for ALL samples of the same size is called the **sampling distribution** of sample proportions.

In a simulation, we collect thousands of random samples to examine the distribution of sample proportions. But when we model this distribution, our model describes the sampling distribution that comes from ALL possible random samples of the same size.

## Example

## Applying the Model for the Sampling Distribution

Let's apply this model to our previous example about the population of part-time college students to see how it compares to our simulation. Recall that we assumed the population of part-time college students is 60% female. We selected samples of 25 part-time college students and calculated the proportion of females in each sample.

Compare the mean and standard deviation we observed in the simulation to the mathematical model. Notice that the conditions are met, so a normal model is a good fit. We see that the model is a good description of the center, spread, and shape we observed in the simulation.

## Learn By Doing

According to the National Postsecondary Student Aid Study conducted by the U.S. Department of Education in 2008, 62% of graduates from public universities had student loans.

https://assessments.lumenlearning.co...sessments/3578

https://assessments.lumenlearning.co...sessments/3891

https://assessments.lumenlearning.co...sessments/3579

Learn By Doing

https://assessments.lumenlearning.co...sessments/3580

https://assessments.lumenlearning.co...sessments/3586

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 7.7: Distribution of Sample Proportions (4 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.8: Distribution of Sample Proportions (5 of 6)

## Learning Objectives

- Use a z-score and the standard normal model to estimate probabilities of specified events.

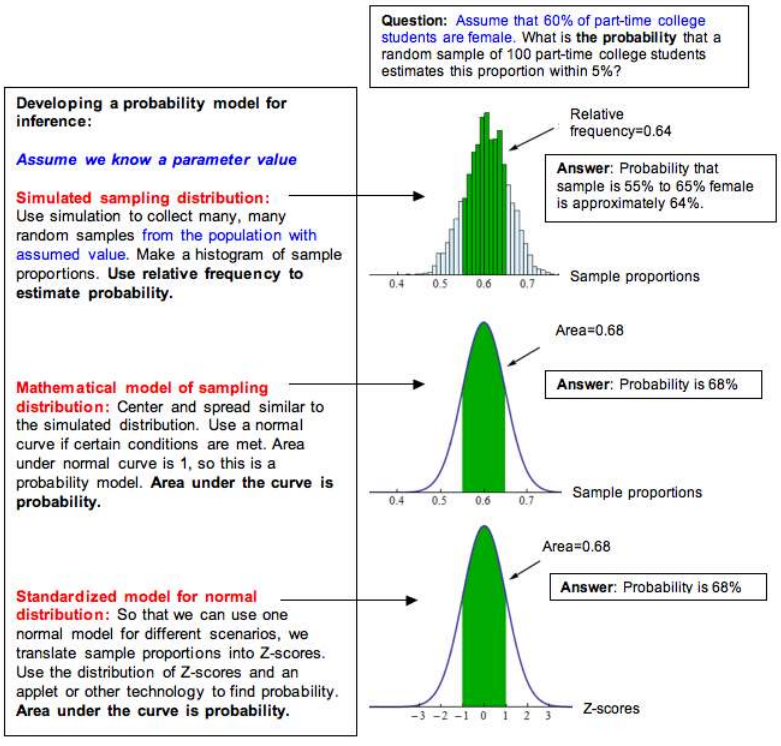
From our work on the previous page, we now have a mathematical model of the sampling distribution of sample proportions. This model describes how much variability we can expect in random samples from a population with a given parameter. If a normal model is a good fit for a sampling distribution, we can apply the empirical rule and use *z*-scores to determine probabilities. Here we link probability to the kind of thinking we do in inference.

## Making Connections to Probability Models in *Probability and Probability Distribution*

Probability describes the chance that a random event occurs. Recall the concept of a random variable from the module *Probability and Probability Distribution*. When a variable is random, it varies unpredictably in the short run but has a predictable pattern in the long run. Sample proportions from random samples are a random variable. We cannot predict the proportion for any one random sample; they vary. But we can predict the pattern that occurs when we select a great many random samples from a population. The sampling distribution describes this pattern. When a normal model is a good fit for the sampling distribution, we can use what we learned in the previous module to find probabilities.

Recall probability models we saw in *Probability and Probability Distribution*. We saw examples of models with skewed curves, but we focused on normal curves because we use normal probability models to describe sampling distributions in Modules 7 to 10 when we make inferences about a population. As we now know, we can use a normal model only when certain conditions are met. Whenever we want to use a normal model, we must check the conditions to make sure a normal model is a good fit.

Here we summarize our general process for developing a probability model for inference. This is essentially the same process we used in the previous module for developing normal probability models from relative frequencies.

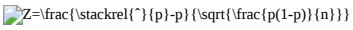



If a normal model is a good fit for the sampling distribution, we can standardize the values by calculating a *z*-score. Then we can use the standard normal model to find probabilities, as we did in *Probability and Probability Distribution*.

The *z*-score is the error in the statistic divided by the standard error. For sample proportions, we have the following formulas.

$$\begin{array}{l}\mathrm{standard}\text{ }\mathrm{error}=\sqrt{\frac{p(1-p)}{n}}\\ Z=\frac{\mathrm{statistic}-\mathrm{parameter}}{\mathrm{standard}\text{ }\mathrm{error}}=\frac{\stackrel{ˆ}{p}-p}{\mathrm{standard}\text{ }\mathrm{error}}\end{array}$$

We can also write this as one formula:



$$Z=\frac{\stackrel{ˆ}{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

### Comment

This *z*-score formula is similar to the *z*-score formula we used in *Probability and Probability Distribution*. We described the *z*-score as the number of standard deviations a data value is from the mean. Here we can describe the *z*-score as the number of standard errors a sample proportion is from the mean. Because the mean is the parameter value, we can say that the *z*-score is the number of standard errors a sample proportion is from the parameter.

A positive *z*-score indicates that the sample proportion is larger than the parameter. A negative *z*-score indicates that the sample proportion is smaller than the parameter.

### Example

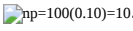
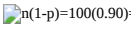
## Probability Calculations for Community College Enrollment

Let's return to the example of community college enrollment. Recall that a 2007 report by the Pew Research Center stated that about 10% of the 3.1 million 18- to 24-year-olds in the United States were enrolled in a community college. Let's again suppose we randomly selected 100 young adults in this age group and found that 15% of the sample was enrolled in a community college.

Previously, we determined that 15% is a surprising result. Now we want to be more precise. We ask this question: *What is the probability that a random sample of this size has 15% or more enrolled in a community college?*

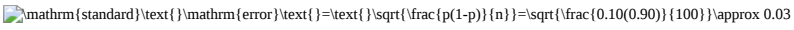
To answer this question, we first determine if a normal model is a good fit for the sampling distribution.

**Check normality conditions:**

Yes, the conditions are met. The number of expected successes and failures in a sample of 100 are at least 10. We expect 10% of the 100 to be enrolled in a community college,  . We expect 90% of the 100 to not be enrolled,  .

We therefore can use a normal model, which allows us to use a *z*-score to find the probability.

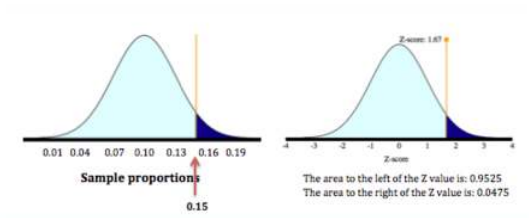
**Find the *z*-score:**



$$\mathrm{standard}\text{ }\mathrm{error}=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.10(0.90)}{100}}\approx 0.03$$

$$Z=\frac{\mathrm{statistic}-\mathrm{parameter}}{\mathrm{standard}\text{ }\mathrm{error}}=\frac{0.15-0.10}{0.03}\approx 1.67$$

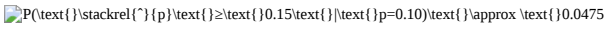
**Find the probability using the standard normal model:**




We want the probability that the sample proportion is 15% or more. So we want the probability that the *z*-score is greater than or equal to 1.67. The probability is about 0.0475.

**Conclusion:** If it is true that 10% of the population of 18- to 24-year-olds are enrolled at a community college, then it is unusual to see a random sample of 100 with 15% or more enrolled. The probability is about 0.0475.

Note: This probability is a conditional probability. Recall from *Relationships in Categorical Data with Intro to Probability* that we write a conditional probability *P*(*A* given *B*) as *P*(*A* | *B*). Here we write *P*(a sample proportion is 0.15 given that the population proportion is 0.10) as



$$P(\text{ }\stackrel{ˆ}{p}\text{ }\geq\text{ }0.15\text{ }|\text{ }p=0.10)\text{ }\approx \text{ }0.0475$$

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=278

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3892

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

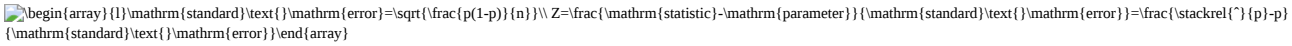
This page titled 7.8: Distribution of Sample Proportions (5 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

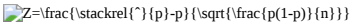
# 7.9: Distribution of Sample Proportions (6 of 6)

## Learning Objectives

- \begin{array}{l}\mathrm{standard}\text{}\mathrm{error}=\sqrt{\frac{p(1-p)}{n}}\\ Z=\frac{\mathrm{statistic}-\mathrm{parameter}}{\mathrm{standard}\text{}\mathrm{error}}=\frac{\stackrel{^}{p}-p}{\mathrm{standard}\text{}\mathrm{error}}\end{array}

Z=\frac{\stackrel{^}{p}-p}{\sqrt{\frac{p(1-p)}{n}}}





CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.9: Distribution of Sample Proportions (6 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.10: Introduction- Distribution of Sample Proportions

## What you'll learn to do: Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.

### LEARNING OBJECTIVES

- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results.
- Distinguish between a sample statistic and a population parameter.
- Use a z-score and the standard normal model to estimate probabilities of specified events.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.10: Introduction- Distribution of Sample Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.11: Statistical Inference (1 of 3)

## Learning Objectives

- Find a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.

From the Big Picture of Statistics, we know that our goal in statistical inference is to *infer* from the sample data some conclusion about the wider population the sample represents. In the first section, "Distribution of Sample Proportions," we investigated the obvious fact that random samples vary. Because different samples may lead to different conclusions, we cannot be certain that our conclusions are correct. Statistical inference uses the language of probability to say how trustworthy our conclusions are.

We learn two types of inference: **confidence intervals** and **hypothesis tests**. We construct a confidence interval when our goal is to estimate a population parameter (or a difference between population parameters). We conduct a hypothesis test when our goal is to test a claim about a population parameter (or a difference between population parameters). Both types of inference are based on the sampling distribution of sample statistics. For both, we report probabilities that state what would happen if we used the inference method repeatedly.

In this section, we build on the ideas in "Distribution of Sample Proportions" to reason as we do in inference, but we do not do formal inference procedures now. Instead, we focus on the logic of inference. We use categorical data and proportions to investigate the logic of inference. But all of the ideas we discuss here apply to quantitative variables and means.

## Confidence Intervals

When our goal is to estimate a population proportion, we select a random sample from the population and use the sample proportion as an estimate. Of course, random samples vary, so we want to include a statement about the amount of error that may be present. Because sample proportions vary in a predictable way, we can also make a probability statement about how confident we are in the process we used to estimate the population proportion.

We can find many examples of confidence intervals reported in the media. Here is an example.

### Example

### Do You Have Problems Sleeping?

The National Sleep Foundation sponsors an annual poll. In 2011, the poll found that "43% of Americans between the ages of 13 and 64 say they rarely or never get a good night's sleep on weeknights. More than half (60%) say that they experience a sleep problem every night or almost every night (i.e., snoring, waking in the night, waking up too early, or feeling unrefreshed when they get up in the morning" (as reported at www.sleepfoundation.org).

*Are these percentages sample statistics or population parameters?* These statistics describe the responses of a sample of Americans.

Let's focus on the 60% who say they experience a sleep problem every night or almost every night. Does this mean that 60% of all Americans have this same experience? Well, no. This is a sample statistic from a poll. But from this sample, we want to infer what percentage of the population does have sleep problems. Since the percentage with sleep problems will differ from one sample to the next, we need to make a statement about how much error we might expect between a sample percentage and the population percentage.

In the "Poll Methodology and Definitions" section of the article, we find more detailed information about the poll. According to the Sleep Foundation website, "The 2011 Sleep in America® annual poll was conducted for the National Sleep Foundation by WB&A Market Research, using a random sample of 1,508 adults between the ages of 13 and 64. The margin of error is 2.5 percentage points at the 95% confidence level."

There is a lot of important information here:

- The sample is random.
- The sample size is 1,508.
- The margin of error is 2.5%.

- The confidence level is 95%.

From this information, we can construct an interval that we are reasonably confident contains the population proportion.

- Sample statistic ± margin of error
- 60% ± 2.5%
- 57.5% to 62.5%

This interval is an example of a **confidence interval**. We interpret the interval this way: We are 95% confident that between 57.5% and 62.5% of *all* Americans experience a sleep problem every night or almost every night.

*How confident are we that this interval contains the population proportion?* In this case, we are 95% confident. This means that 95% of the time, a random sample of this size will have at most 2.5% error. So 95% of these intervals will contain the true population proportion. Another way to say this is that this method accurately estimates the population proportion 95% of the time.

*Note:* Notice that the sample is a *random* sample. We can construct a confidence interval only with a random sample.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3896

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3897

### Summary

A sample proportion from a random sample provides a reasonable estimate of the population proportion. We do not expect the sample proportion to be exactly equal to the population proportion, but we expect the population proportion to be somewhat close to the sample proportion. The purpose of confidence intervals is to use the sample proportion to construct an interval of values that we can be reasonably confident contains the true population proportion.

### What Is the Connection to the Sampling Distribution?

Sample proportions are estimates for the population proportion, so each sample proportion has error. For an individual sample, we will not know the exact amount of error, so we report a margin of error based on the standard error. Recall that the standard error is the standard deviation of sampling distribution. We can view the standard error as the typical or average error in the sample proportions. To see how this works, let's return to a familiar sampling distribution.

Recall our previous investigation of gender in the population of part-time college students. We investigated these questions: *What proportion of part-time college students are female? If we predict that the proportion is 0.60, how much error can we expect to be confident of in our prediction?*

We predicted the population proportion was 0.60 and ran a simulation to examine the variability in sample proportions for samples of 100 part-time college students. Here is the sampling distribution from the simulation.

We see that we can be very confident that most samples of this size will have proportions that differ from 0.60 by at most 2 standard errors. For this simulation, the standard error in sample proportions was about 0.049. About 95% of the samples have an error less than 2(0.049) = 0.098

If we use two standard errors as the margin of error, we can rewrite the confidence interval.

- *sample statistic ± margin of error*
- *sample proportion ± 2(standard errors)*
- *sample proportion ± 2(0.049)*
- *sample proportion ± 0.098*

Different sample proportions give different intervals. For example, if the sample proportion is 0.57, the confidence interval is 0.472 to 0.668. Here are our calculations.

- *sample proportion ± margin of error*
- *0.57 ± 2(0.049)*

- *0.57 ± 0.098*

The endpoints of the interval are 0.57 - 0.098 = 0.472 and 0.57 + 0.098 = 0.668. The confidence interval is 0.472 to 0.668.

Since about 95% of the samples have at most 9.8% error, we have a 95% confidence interval. Based on this sample, we say we are 95% confident that the percentage of part-time college students who are female is between 47.2% and 66.8%.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3898

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3899

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 7.11: Statistical Inference (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.12: Statistical Inference (2 of 3)

## Learning Objectives

- Find a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.
- Interpret the confidence level associated with a confidence interval.

## A Look at 95% Confidence Intervals on the Number Line

## When Does a 95% Confidence Interval Contain the True Population Proportion?

## How Confident Are We That a 95% Confidence Interval Contains the Population Proportion?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.12: Statistical Inference (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.13: Statistical Inference (3 of 3)

## Learning Objectives

- Test a hypothesis about a population proportion using a simulated sampling distribution or a normal model of the sampling distribution. State a conclusion in context.

Now we focus on the second type of inference: hypothesis testing and the logic behind it.

In hypothesis testing, we make a claim about a parameter and test it. On this page, we make a claim about a population proportion and use a sample proportion from data to test our claim. This is very similar to the thinking we did with simulations in the previous module.

## Example

## Test a Claim about Health Insurance Coverage

With data from the 2010 National Health Interview Survey, the Centers for Disease Control and Prevention (CDC) estimates that 22% of U.S. adults (age 18–64) did not have health insurance in 2010. Is the percentage higher this year? In a hypothesis test, we translate the research question into a claim about the population.

**Claim:** The percentage of U.S. adults (ages 18–64) who do not have health insurance is higher than 22% this year.

To test the claim, we assume that the percentage is 22% this year. Then we gather a random sample from the population to test the claim. Suppose 25% of a random sample of 600 U.S. adults (age 18–64) do not have health insurance this year. What can we conclude? Obviously, this sample has more than 22% uninsured adults. But does this data suggest the percentage *of the U.S. adult population* (age 18–24) who are uninsured is greater than 22%?

To test the claim, we begin with a population with and take random samples of 600 people at a time.

- If a sample proportion of 0.25 is *likely* to occur when sampling from a population with , then this sample could have come from a population with 22% uninsured. The evidence from the sample is not strong enough to conclude that the population percentage is greater than 22%.
- If a sample proportion is *unlikely* when sampling from a population with , then the sample provides evidence that the proportion of population who are uninsured is greater than 22%.

Likely or unlikely? It depends on how much the sample proportions vary. We need to use a simulation or a mathematical model to represent the sampling distribution of sample proportions.

**Simulation:** We used a simulation to select 2,000 random samples of 600 people, each from a population with . Judging from the simulation, a sample proportion of 0.25 is unlikely. Sample proportions of 0.25 or greater do not occur very often. In this simulation, only 90 out of the 2,000 random samples (4.5%) had proportions of 0.25 or greater.

**Normal Probability Model of the Sampling Distribution:** We can also apply what we know from our work with a normal model of the sampling distribution. Visually, the simulated sampling distribution looks like it has a normal shape. The formal conditions for use of a normal model require that the expected count of successes and failures are at least 10. Here we expect 22% (132 people) of the 600 people to be uninsured. We expect 78% (468 people) of the 600 to be insured. Since the sampling distribution meets the conditions for use of a normal model, we can calculate the *z*-score and find the probability that a random sample has a proportion of 0.25 or greater. The *z*-score is 1.76, and our simulation gives a probability of 0.039.

Both approaches suggest that a sample proportion of 0.25 is unlikely. We don't expect to see 25% or more uninsured in random samples of 600 very often. We estimate the chances as 4.5% with the simulation and only about 3.9% with the normal model. This is so unusual that we conclude the data from this year did not come from a population with only 22% uninsured. Our data provides strong evidence that more than 22% in the population are uninsured.

## Example

## Test the Claim about Health Insurance Again with Different Data

Now we retest the same claim using different data.

**Claim:** The percentage of U.S. adults (ages 18–64) who do not have health insurance is higher than 22% this year.

Suppose 26% of a random sample of 50 U.S. adults (age 18–64) do not have health insurance this year. What can we conclude?

**Simulation:** We again used a simulation to select 2,000 random samples from a population with $p=0.22$. This time there are 50 people in each sample. Judging from the simulation, a sample proportion of 0.26 is not unlikely. Sample proportions of 0.26 or greater occur frequently. In this simulation, 608 (30.4%) of the 2,000 random samples had proportions of 0.26 or greater.

**Normal Probability Model of the Sampling Distribution:** Visually, the simulated sampling distribution looks like it has a normal shape. The formal conditions for use of a normal model require that the expected count of successes and failures is at least 10. Here we expect 22% (11 people) of the 50 people to be uninsured. We expect 78% (39 people) of the 50 to be insured. Since the sampling distribution meets the conditions for use of a normal model, we can calculate the *z*-score and find the probability that a random sample has a proportion of 0.26 or greater. The *z*-score is 0.68, and our simulation gives a probability of 0.25.

Both approaches suggest that a sample proportion of 0.26 is not unlikely. It falls within a typical range of sample proportions that we expect to see from random samples of 50 people. The *z*-score is 0.68, meaning the sample proportion is less than 1 standard error from 0.22. We also see the probability that a random sample has 26% or more uninsured is high: about 30% according to the simulation and about 25% according to the normal model. This probability suggests the data from this year could have come from a population with only 22% uninsured. Even though 26% are uninsured in our sample, our data does not provide strong evidence that more than 22% of the population are uninsured this year.

### Comment

How can a sample proportion of 0.25 be unusual in the first example but a sample proportion of 0.26 *not* be unusual in the second example? These two examples highlight an important point. We have to judge a sample result by looking at it in relation to other samples of the same size. In the first example, the samples are large (600 adults in each sample), so the sample proportions do not vary much. In this sampling distribution, a sample result of 0.25 or greater is unlikely to occur. In the second example, the samples are smaller (only 50 adults in each sample), so the sample proportions vary more. In this sampling distribution, a sample result of 0.26 or greater is likely to occur.

### Learn By Doing

## Has the Asthma Rate in Children Decreased?

With data from the 2009 National Health Interview Survey, the Centers for Disease Control and Prevention estimated that 9.4% of U.S. children had asthma. Is the percentage lower this year?

Suppose we select a random sample of 50 children this year and find that 3 of the 50 have asthma.

The conditions are not met for use of a normal model because the expected number with asthma (0.094 of 50) is less than 10, so we ran a simulation with $p = 0.094$.

https://assessments.lumenlearning.co...sessments/3588

### Is the Asthma Rate Higher for Children Living in Poverty?

With data from the 2009 National Health Interview Survey, the Centers for Disease Control and Prevention estimated that 9.4% of U.S. children had asthma. Is the percentage higher for children living in poverty?

Suppose we select a random sample of 50 poor children and find that 9 of the 50 have asthma.

The conditions are not met for use of a normal model, so we ran a simulation with $p=0.094$.

### Learn By Doing

## Has the Percentage of Adults (18 and Older) Who Do Not Exercise Increased since 2007?

With data from the 2009 National Health Interview Survey, the Centers for Disease Control and Prevention estimated that 33% of U.S. adults (18 and older) do not exercise. Is the percentage higher this year?

Suppose we select a random sample of 100 adults (18 and older) this year. The conditions are met for use of a normal model, because we expect 33 (33% of 100) in the sample will not exercise and 67 (67% of 100) will. Both expected counts are greater than 10. We use a *z*-score and a standard normal curve to assess the evidence. (The standard error is 0.047.)

https://assessments.lumenlearning.co...sessments/3589

Click here to open the simulation in its own window.

### Learn By Doing

## Has the Percentage of U.S. Adults Who Smoke Decreased since 2007?

With data from the 2005–2007 National Health Interview Survey, the Centers for Disease Control and Prevention estimated that about 20% of U.S. adults (18 and older) smoke. Is the percentage lower this year?

Suppose we select a random sample of 100 adults (18 and older) this year. The conditions are met for use of a normal model, because we expect 20 smokers (20% of 100) in the sample and 80 nonsmokers. Both expected counts are greater than 10, so we use a *z*-score and a standard normal curve to assess the evidence. (The standard error is 0.04.)

https://assessments.lumenlearning.co...sessments/3590

Click here to open the normal distribution calculator in its own window.

### Comment

Are you wondering how extreme the results have to be before we conclude that the result is unusual? Well, it is a judgment call. No single cutoff point determines that a result is unusual. But in practice, we often agree on a cutoff point before we collect the data. In *Inference for One Proportion*, we discuss this idea further. However, if you are worried about this issue when taking a Checkpoint for this module, you can consider a result to be unusual if the probability is less than 5%.

### Let's Summarize

Introduction to Statistical Inference

The two types of inference procedures in this course are *confidence intervals* and *hypothesis tests*. The goal of a confidence interval is to estimate a parameter value. The goal of a hypothesis test is to test a claim about a parameter. Both types of inference are based on the sampling distribution of sample statistics. For both, we report probabilities that state what would happen if we used the inference method many times.

### Confidence Intervals

The purpose of a confidence interval is to estimate a population parameter on the basis of a sample statistic. Sample statistics vary, so there is always error in our estimate, but we will never know how much. We therefore use the standard error, which is the average error in our sample estimates, to create a margin of error. The margin of error is related to our confidence that the interval contains the population parameter.

We investigated the 95% confidence interval for a population proportion in depth. When a normal model is a good fit for the sampling distribution, the 95% confidence interval has a margin of error equal to 2 standard errors.

$$\begin{array}{l}\mathrm{sample}\text{ }\mathrm{statistic}\text{ }\pm\text{ }\mathrm{margin}\text{ }\mathrm{of}\text{ }\mathrm{error}\\ \mathrm{sample}\text{ }\mathrm{proportion}\text{ }\pm\text{ }2(\mathrm{standard}\text{ }\mathrm{error})\\ \stackrel{\frown}{p}\text{ }\pm\text{ }2\sqrt{\frac{p(1-p)}{n}}\end{array}$$

We say we are 95% confident that the calculated interval contains the population proportion, meaning that 95% of the time, these intervals will actually contain the population proportion, and we will be right. Five percent of the time, we will be wrong. We can never tell if a confidence interval does or does not contain the population proportion we are trying to estimate.

## Hypothesis Tests

The purpose of a hypothesis test is to use sample data to test a claim about a population parameter. We investigated testing a claim about a population proportion informally.

We make a claim about a population proportion. From the claim, we state an assumption about the value of the population proportion. Could the data have come from this population? Or is the sample proportion too far off? It depends on how much random samples from this population vary. We construct a simulation or a normal model to represent the sampling distribution that occurs when sampling from a population with this assumed value. We make a judgment about whether the data is likely or unlikely to occur in the sampling distribution. If the data supports our claim and is unlikely, then we doubt our assumption about the population proportion.

For example, if last year 20% of the U.S adult population smoked, we might claim that the percentage of smokers in the United States this year is greater. So in the simulation we set $p=0.20$ and see if the data causes us to question this claim.

- If a sample proportion is *likely* to occur in the sampling distribution, then this sample result could have come from a population with the assumed value. In this situation, the data do not lead us to doubt our assumption about the value of the parameter. We therefore conclude that the evidence from the sample is not strong enough to support our original claim
- In our example, a sample proportion that is likely to occur means we do not question the assumption that we made when we set $p=0.20$. We cannot conclude that the percentage of smokers in the United States is greater than 20% this year.
- If a sample proportion supports our claim and is *unlikely* to occur in the sampling distribution, then it is unlikely that this sample result came from a population with the assumed value. In this situation, the data lead us to doubt our assumption about the value of the parameter. We conclude that the evidence from the sample is strong enough to support the claim.
- In our example, a sample proportion that is unusually large means that the data makes us doubt the assumption we made when we set $p=0.20$. We therefore is probably greater than 20% this year.

Likely or unlikely? It depends on how much the sample proportions vary. If the normal model is a good fit for the sampling distribution, we can calculate a *z*-score and use a simulation to associate a probability with our “likely” or “unlikely” statement. Recall what we learned in “Distribution of Sample Proportions” to calculate the *z*-score.

$$\begin{array}{l}\mathrm{standard}\text{ }\mathrm{error}=\sqrt{\frac{p(1-p)}{n}}\\ Z=\frac{\mathrm{statistic}-\mathrm{parameter}}{\mathrm{standard}\text{ }\mathrm{error}}=\frac{\stackrel{\wedge}{p}-p}{\mathrm{standard}\text{ }\mathrm{error}}\end{array}$$

We can also write this as one formula:

$$Z=\frac{\stackrel{\wedge}{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.13: Statistical Inference (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 7.14: Putting It Together- Linking Probability to Statistical Inference

## Let's Summarize

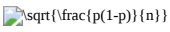
Overview of Statistical Inference

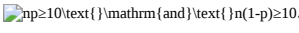
- Inference is based on probability.
- A parameter is a number that describes a population. A statistic is a number that describes a sample. In inference, we use a statistic to draw a conclusion about a parameter. These conclusions include a probability statement that describes the strength of the evidence or our certainty.
- For a categorical variable, the parameter and statistics are proportions. For a quantitative variable, the parameter and statistics are means.
- For a given situation, we assume the parameter is fixed. It does not change. In contrast, statistics always vary. When we take random samples, the fluctuation in statistics is due to chance. We create simulations and mathematical models to describe the variability we expect to see in sample statistics.

## Sampling Distribution for a Sample Proportion

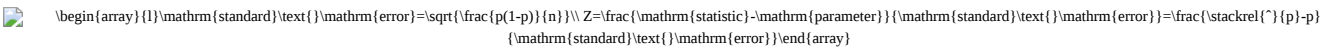
- Larger samples have less variability.
- For a categorical variable we assume that population has a proportion *p* of successes. When we select random samples from this population, the sample proportions have a pattern in the long run. We can describe this pattern with a mathematical model of the sampling distribution. The model has the following center, spread, and shape.

**Center:** Mean of the sample proportions is *p*, the population proportion.

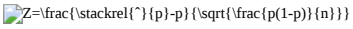
**Spread:** Standard deviation of the sample proportions is 

**Shape:** A normal model is a good fit if the expected number of successes and failures is at least 10. We can translate these conditions into formulas:  .

- When a normal is a good fit for the sampling distribution, we can calculate a *z*-score, which allows us to use the standard normal model to find probabilities associated with the sampling distribution.



We can also write this as one formula:



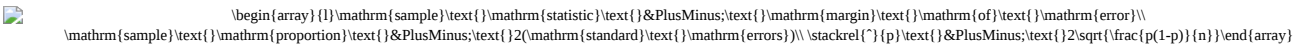
## Introduction to Statistical Inference

This course presents two types of inference procedures: confidence intervals and hypothesis tests. The goal of a confidence interval is to estimate a parameter value. The goal of a hypothesis test is to test a claim about a parameter. Both types of inference are based on the sampling distribution of sample statistics. For both, we report probabilities that state what would happen if we used the inference method many times.

## Confidence Intervals

The purpose of a confidence interval is to estimate a population parameter on the basis of a sample statistic. Sample statistics vary, so there is always error in our estimate, but we never know how much. We therefore use the standard error, which is the average error in our sample estimates, to create a margin of error. The margin of error is related to our confidence that the interval contains the population parameter.

We investigated the 95% confidence interval for a population proportion in depth. When a normal model is a good fit for the sampling distribution, the 95% confidence interval has a margin of error equal to 2 standard errors.



We say we are 95% confident that the calculated interval contains the population proportion. This means that 95% of the time, these intervals will actually contain the population proportion, and we will be right. Five percent of the time, we will be wrong. We

can never tell if a confidence interval does or does not contain the population proportion we are trying to estimate.

### Hypothesis Tests

The purpose of a hypothesis test is to use sample data to test a claim about a population parameter. We investigated testing a claim about a population proportion informally.

We make a claim about a population proportion. From the claim, we state an assumption about the value of the population proportion. Could the data have come from this population? Or is the sample proportion too far off? It depends on how much random samples from this population vary. We construct a simulation or a normal model to represent the sampling distribution that occurs when sampling from a population with this assumed value. We make a judgment about whether the sample proportion is likely or unlikely to occur in the sampling distribution. If the data supports our claim and is unlikely, then we doubt our assumption about the population proportion.

Likely or unlikely? It depends on how much the sample proportions vary. If the normal model is a good fit for the sampling distribution, we can find a *z*-score and use a simulation to associate a probability with our "likely" or "unlikely" statement.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 7.14: Putting It Together- Linking Probability to Statistical Inference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 8: Inference for One Proportion


8.1: Why It Matters- Inference for One Proportion
8.2: Introduction- Estimating a Population Proportion
8.3: Estimating a Population Proportion (1 of 3)
8.4: Estimating a Population Proportion (2 of 3)
8.5: Estimating a Population Proportion (3 of 3)
8.6: Introduction- Hypothesis Testing
8.7: Hypothesis Testing (1 of 5)
8.8: Hypothesis Testing (2 of 5)
8.9: Hypothesis Testing (4 of 5)
8.10: Hypothesis Testing (3 of 5)
8.11: Hypothesis Testing (5 of 5)
8.12: Introduction- Hypothesis Test for a Population Proportion
8.13: Hypothesis Test for a Population Proportion (1 of 3)
8.14: Hypothesis Test for a Population Proportion (2 of 3)
8.15: Hypothesis Test for a Population Proportion (3 of 3)
8.16: Putting It Together- Inference for One Proportion
8.17: StatTutor- Cell Phones
8.18: Assignment- Hypothesis Testing for the Population Proportion p


---


This page titled 8: Inference for One Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.1: Why It Matters- Inference for One Proportion

## Learning Objectives

- Recognize situations that call for testing a claim about a population proportion or estimating a population proportion.

In *Inference for One Proportion*, we focus on making inferences about population proportions. The types of research questions we focus on in this module are bolded in the text below. Notice that we are working with categorical variables again.

| Type of Question | Examples | Variable Type | Unit |
|---|---|---|---|
| Make an estimate about the population | **What proportion of all U.S. adults support the death penalty?** | **Categorical variable** | **Inference for One Proportion** |
| | What is the average number of hours that community college students work each week? | Quantitative variable | Inference for Means |
| Test a claim about the population | **Do the majority of community college students qualify for federal student loans?** | **Categorical variable** | **Inference for One Proportion** |
| | Has the average birth weight in a town decreased from 3,500 grams? | Quantitative variable | Inference for Means |
| Compare two populations | Are teenage girls more likely to suffer from depression than teenage boys? | Categorical variable | Inference for Two Proportions |
| | In community colleges do female students have a higher average GPA than male students? | Quantitative variable | Inference for Means |

We will build on what we learned in in the previous module with two additions.

- We use more formal vocabulary and notation for hypothesis testing.
- We will not know the population proportion, so we make some minor adjustments to the model of the sampling distribution that we developed in *Linking Probability to Statistical Inference*. The adjustments affect how we calculate the standard error.

Here again is the Big Picture. We highlighted ideas new to *Inference for One Proportion* in purple.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3591

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.1: Why It Matters- Inference for One Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.2: Introduction- Estimating a Population Proportion

## What you'll learn to do: Construct a confidence interval to estimate a population proportion.

### LEARNING OBJECTIVES

- Recognize situations that call for testing a claim about a population proportion or estimating a population proportion.
- Construct a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.
- For a confidence interval, interpret the meaning of a confidence level and relate it to the margin of error.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.2: Introduction- Estimating a Population Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.3: Estimating a Population Proportion (1 of 3)

## Learning Objectives

- Construct a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.

## Introduction

In “Estimating a Population Proportion,” we continue our discussion of estimating a population proportion with a confidence interval. Recall that the purpose of a confidence interval is to use a sample proportion to construct an interval of values that we can be reasonably confident contains the true population proportion.

The basic idea is summarized here:

- When we select a random sample from the population of interest, we expect the sample proportion to be a good estimate of the population proportion. But we also know that sample proportions vary, so we expect some error. (Remember that the error here is due to chance. It is not due to a mistake that anyone made.)
- For a given sample proportion, we will not know the amount of error, so we use the standard error as an estimate for the average amount of error we expect in sample proportions. (Recall that the standard error is the expected standard deviation of sample proportions when we take many, many random samples.)
- If a normal model is a good fit for the sampling distribution, then about 95% of sample proportions estimate the population proportion within 2 standard errors. We say that we are 95% confident that the following interval contains the population proportion.



$$\begin{array}{l}p\text{}\±\text{}\mathrm{margin}\text{}\mathrm{of}\text{}\mathrm{error}\\ p\text{}\±\text{}2(\mathrm{standard}\text{}\mathrm{error})\\ p\text{}\±\text{}2\sqrt{\frac{p(1-p)}{n}}\end{array}$$

You may realize that this formula for the confidence interval is a bit odd, since our goal in calculating the confidence interval is to estimate the population proportion $p$. Yet the formula requires that we know $p$. In the section “Introduction to Statistical Inference,” we used an estimate for $p$ from a previous study when calculating the confidence interval. This is not the usual way statisticians estimate the standard error, but it captured the main idea and allowed us to practice finding and interpreting confidence intervals. Now, we develop a different way to estimate standard error that is commonly used in statistical practice.

## Example

### Community College Students and Gender

According to a 2010 report from the American Council on Education, females make up 57% of the college population in the United States. Students in a statistics class at Tallahassee Community College want to determine the proportion of female students at TCC. They select a random sample of 135 TCC students and find that 72 are female, which is a sample proportion of 72 / 135 ≈ 0.533. So 53.3% of the students in the sample are female.

*What can they conclude about the proportion of females at the college? How confident can they be in their estimate?*

To answer these questions, we need to find a confidence interval.

**Checking conditions:**

We learned in *Linking Probability to Statistical Inference* that a confidence interval comes from a normal model of the sampling distribution. Let’s first make sure that a normal model is appropriate here. Recall the two conditions for using a normal model for sample proportions:

- The sample must be random.
- The expected number of successes in the sample, $np$, and the expected number of failures, $n(1 – p)$, are both greater than or equal to 10. In symbols, this is $np \geq 10$ and $n(1 - p) \geq 10$. Recall that *success* doesn’t mean good and *failure* doesn’t mean bad. A success is just what we are counting.

When we try to check these conditions, we have a problem. We do not know $p$, the population proportion. In fact, $p$ is what we are trying to estimate! So we cannot determine the expected number of successes and failures. Our solution to this problem is to adjust

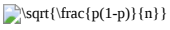
these conditions. Advanced theory tells us that if the *actual* number of successes and failures in the sample are greater than or equal to 10, then a normal model is still a good fit.

This sample contains 72 successes (female students) and 63 failures (male students). Both are greater than 10. We therefore use the normal model for the sampling distribution.

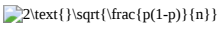
**Finding the margin of error:**

We know that a sample proportion is only an estimate for the population proportion. We do not expect the sample proportion to equal the population proportion, so there is some error due to random chance. We use the standard deviation of the sample proportions to describe the amount of error we can expect in random samples. We call this the standard error.

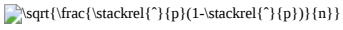
In *Linking Probability to Statistical Inference*, we learned that the standard error of the sample proportion depends on the population proportion and sample size. Here is the formula for the standard error:



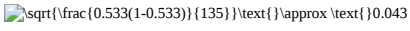
When we use a normal model for the sampling distribution, 95% of sample proportions estimate the population proportion within approximately 2 standard errors. So the *margin of error* is the following:



Now let's calculate the margin of error for the TCC estimate of 53.3%. Notice that we have the same problem we had earlier. We don't know *p*, the population proportion. So we can't calculate the margin of error! Our solution to this problem is to estimate the standard error using the sample proportion in place of *p*. We call this the *estimated standard error*, and the formula is:



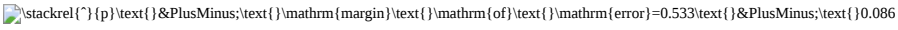
For this example, the estimated standard error is



So the margin of error for the 95% confidence interval is:

**Finding the confidence interval:**

We can interpret the margin of error by saying we are 95% confident that the proportion of all students at TCC who are female is within 0.086 of our sample proportion of 0.533. We can then write the interval in the following form:



When we add and subtract the margin of error from the sample proportion, the confidence interval is 0.447 to 0.619.

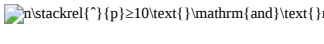
**Conclusion:**

We are 95% confident that the proportion of all TCC students who are female is between 0.447 and 0.619. We can also make this statement using percentages. We are 95% confident that the percentage of all TCC students who are female is between 44.7% and 61.9%.

Recall from *Linking Probability to Statistical Inference* that *95% confidence* means this method captures the population proportion about 95% of the time.

## Summary

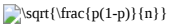
**Conditions for using the normal model of the sampling distribution:**

In *Linking Probability to Statistical Inference*, we saw that a normal model describes the behavior of sample proportions if $np \geq 10$ and $n(1-p) \geq 10$. These formulas say that the *expected* number of successes and failures in the sample must be 10 or greater. In *Inference for One Proportion*, we will never know the value of the population proportion *p*, so we estimate *p* with a sample proportion. Now we will assume that we can use a normal model if  .

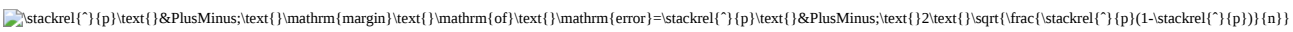
These formulas say that the *actual* number of successes and failures in the sample are 10 or greater.

**The 95% confidence interval for estimating population proportion *p* :**

In *Linking Probability to Statistical Inference*, we learned that the error in an estimate is related to the spread in the sampling distribution. We saw that the standard error of the sampling distribution of sample proportions is given by this formula:



In *Inference for One Proportion*, we are estimating the population proportion $p$. So we estimate the standard error by replacing $p$ with the sample proportion, which affects the margin of error in the confidence interval. We have the following adjustment to the confidence interval formula:



## Learn By Doing

Foothill College's athletic department wants to calculate the proportion of students who have attended a women's basketball game at the college. They use student email addresses, randomly choose 220 students, and email them. Of the 145 who responded, 22 had attended a women's basketball game.

https://assessments.lumenlearning.co...sessments/3592

https://assessments.lumenlearning.co...sessments/3593

https://assessments.lumenlearning.co...sessments/3903

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3904

https://assessments.lumenlearning.co...sessments/3905

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.3: Estimating a Population Proportion (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.4: Estimating a Population Proportion (2 of 3)

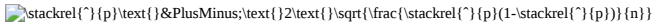
## Learning Objectives

- Construct a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.
- For a confidence interval, interpret the meaning of a confidence level and relate it to the margin of error.

## Introduction

On the previous page, we estimated a population proportion by calculating the approximate 95% confidence interval.

We used the following formula:



This formula is valid only if we can use a normal distribution to model the sampling distribution for the sample proportions. We can use the normal model if we have at least 10 successes and at least 10 failures in the sample.

Recall that we used 2 estimated standard errors because of the empirical rule. The empirical rule says that approximately 95% of all sample proportions will fall within 2 standard errors of the population proportion. So 95% of the sample proportions have an error that is less than 2 standard errors. On the previous page, we made a slight modification using the estimated standard error where we replaced  with .

We often use the 95% confidence level, but in practice you may also see 90% and 99% confidence levels. On this page, we begin to investigate the impact of changing the confidence level on the confidence interval.

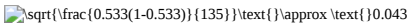
## Example

## Community College Students and Gender

Recall from the previous page that students in a statistics class at Tallahassee Community College wanted to determine the proportion of female students at TCC. They selected a random sample of 135 students and found that 72 were female. Previously, we calculated an approximate 95% confidence interval. We estimated that the proportion of all TCC students who are female is between 0.447 and 0.619.

Now we calculate the 90% confidence interval for the proportion of all TCC students who are female. Because the results from the sample are the same, we do not need to check the conditions for a normal model for the sampling distribution. We already verified that these conditions are met.

Because the sample proportion is the same, the estimated standard error will also be the same:



But the margin of error will change. We estimated the margin of error for the 95% confidence interval by multiplying the estimated standard error by 2. Now we need to determine the *z*-scores that will give us the middle 90% of the normal distribution.

Technology is used to determine the *z*-scores that mark off the middle 90% of the sampling distribution. The *z*-scores are ±1.65. Using this value in place of 2 in the margin of error gives us a 90% confidence interval:

**95% confidence interval:** 0.533 ± **2**(0.043) ≈ 0.533 ± **0.086** = (0.447, 0.619)

**90% confidence interval:** 0.533 ± **1.65**(0.043) ≈ 0.533 ± **0.07** = (0.463, 0.603)

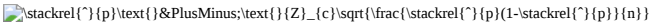
Note: Frequently, you will see the *z*-scores that mark off the middle 90% of the sample proportions represented more precisely as ±1.645.

*What is the impact of decreasing the confidence level to 90%?*

The 90% interval allows a smaller margin for error than the 95% interval. The 90% confidence interval is narrower than the 95% confidence interval. It may seem like an advantage, but there is a trade-off because we now have less confidence that the interval contains the population proportion. This is an important point. Lower confidence means smaller margin of error. We investigate this idea in more depth later.

## Confidence Interval Formula

Since we are no longer restricting our confidence level to 95%, we can generalize the formula for a confidence interval:




We use a little subscript $c$ on the $z$-score, $Z_c$ , to emphasize that the $z$-score is connected to the confidence level. When giving the value of $Z_c$, we always use the positive $z$-score.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3906

## Comment

Technology often uses 3 decimal places for $Z_c$.

For our most common confidence levels, the values of $Z_c$ are:

90% confidence interval: $Z_c \approx 1.645$
95% confidence interval: $Z_c \approx 1.960$ (2 is a rough approximation; 1.960 is more accurate)
99% confidence interval: $Z_c \approx 2.576$

So when you calculate the confidence interval, rounding will slightly affect the values in your interval.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.4: Estimating a Population Proportion (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

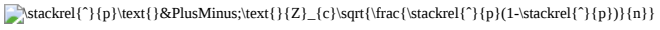
# 8.5: Estimating a Population Proportion (3 of 3)

## Learning Objectives

- Construct a confidence interval to estimate a population proportion when conditions are met. Interpret the confidence interval in context.
- For a confidence interval, interpret the meaning of a confidence level and relate it to the margin of error.

## Introduction

On the previous page, we learned the general formula for a confidence interval for a population proportion:



Recall that, for our most common confidence levels, the values of $Z_c$ are:

90% confidence interval: $Z_c \approx 1.645$
95% confidence interval: $Z_c \approx 1.960$
99% confidence interval: $Z_c \approx 2.576$

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3594

## Confidence Interval Width

The goal of the confidence interval is to estimate the population proportion. If the interval contains the population proportion, a small amount of error means we have a more precise estimate. Narrower confidence intervals give more precise interval estimates for the population proportion, but this is true only if the intervals contain the population proportion.

We saw in the previous activity that a lower confidence level corresponds to a smaller margin of error. In general, 90% confidence intervals are narrower than 95% confidence intervals because there is a smaller margin of error. But we are less confident that 90% confidence intervals contain the population proportion. Recall that in the long run, 10% of these intervals will *not* contain the population proportion at all! We therefore have to choose between precision and confidence.

Of course, ideally, we would like to have a narrow interval *and* a high level of confidence. We can achieve this by increasing the size of the sample.

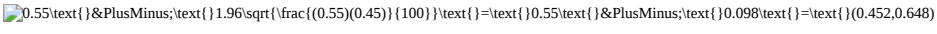
## Example

## College Students and Marijuana Legalization

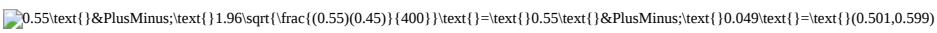
National surveys show that about 43% of American adults support the legalization of marijuana. What proportion of students at Capital Community College support the legalization of marijuana? Suppose students conduct two surveys. For one survey, they randomly select a sample of 100 students. For the other survey, they randomly select a sample of 400 students. Surprisingly, in both surveys, the proportion in favor of legalization is 55%. The students calculate the 95% confidence interval for both surveys.

*What is the impact of the size of the sample on the confidence interval?*

For the sample of size 100, the confidence interval is



For the sample of size 400, the confidence interval is



Notice that the larger sample gives a smaller margin of error. The margin of error for the sample of 400 is half that of the sample of 100.

This makes sense. Our intuition tells us that larger samples should give more precise estimates of the population proportion. We also saw this in Module 7 where we investigated the impact of sample size on the variability in the sample proportions. We saw that

proportions from larger samples vary less. If there is less variability in the sampling distribution, the standard error is smaller. Since we use the standard error to find the margin of error, larger samples will produce a smaller margin of error.

More specifically, we can see that a sample four times larger gives a margin of error half as large because we divide by , the square root of the sample size, in the formula. Similarly, a sample nine times larger gives a margin of error one-third as large.

In general, to decrease the margin of error, we can increase the sample size or decrease the confidence level. We always prefer to increase the sample size because it allows us to keep a higher level of confidence. We want a higher level of confidence because the confidence level is the proportion of intervals that actually contains the population proportion.

### Learn By Doing

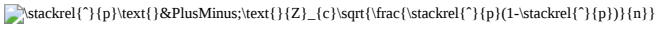
https://assessments.lumenlearning.co...sessments/3595

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3907

## Let's Summarize

- The sample proportion is a point estimate for the population proportion, but it is almost always wrong. We therefore use an interval estimate, called a *confidence interval*, to give us a range of values for the population proportion.
- We can calculate a confidence interval for a population proportion when we can use a normal distribution to model the long-run behavior of sample proportions. We can use a normal distribution model when there are at least 10 observed successes and 10 observed failures.
- The interpretation of a confidence interval depends on the confidence level. For example, using a 95% confidence level, we are 95% confident that the population proportion falls within the interval.
- A confidence interval is a sample proportion plus or minus a margin of error. The margin of error is related to the confidence level. For a 95% confidence level, the margin of error is approximately two standard errors. The formula is

  
- Lower confidence levels and higher sample sizes lead to narrower confidence intervals. A narrower confidence interval has a smaller error. Since we want to be confident that an interval accurately estimates the population proportion, high levels of confidence are desirable. So larger sample sizes are the preferred way to decrease the error and create narrower confidence intervals.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.5: Estimating a Population Proportion (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.6: Introduction- Hypothesis Testing

## What you'll learn to do: Given a claim about a population, construct an appropriate set of hypotheses to test and properly interpret p values and Type I / II errors.

### LEARNING OBJECTIVES

- Given a claim about a population, determine null and alternative hypotheses.
- Recognize the logic behind a hypothesis test and how it relates to the P-value.
- Recognize type I and type II errors.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.6: Introduction- Hypothesis Testing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.7: Hypothesis Testing (1 of 5)

## Learning Objectives

- When testing a claim, distinguish among situations involving one population mean, one population proportion, two population means, or two population proportions.
- Given a claim about a population, determine null and alternative hypotheses.

## Introduction

In inference, we use a sample to draw a conclusion about a population. Two types of inference are the focus of our work in this course:

- Estimate a population parameter with a confidence interval.
- Test a claim about a population parameter with a hypothesis test.

We can also use samples from two populations to compare those populations. In this situation, the two types of inference focus on differences in the parameters.

- Estimate a difference in population parameters with a confidence interval.
- Test a claim about a difference in population parameters with a hypothesis test.

In "Estimating a Population Proportion," we learned to estimate a population proportion using a confidence interval. For example, we estimated the proportion of all Tallahassee Community College students who are female and the proportion of all American adults who used the Internet to obtain medical information in the previous month. We will revisit confidence intervals in future modules.

Now we look more carefully at how to test a claim with a hypothesis test. Statistical investigations begin with research questions. We begin our discussion of hypothesis tests with research questions that require us to test a claim. Later we look at how a claim becomes a hypothesis.

## Example

### Research Questions about Testing Claims

Let's revisit some of the research questions from examples in the module *Types of Statistical Studies and Producing Data* that involve testing a claim.

*Is the average course load for community college students less than 12 semester hours?* This question contains a claim about a population mean. The question contains information about the population, the variable, and the parameter. The population is all community college students. The variable is *course load in semester hours*. It is quantitative, so the parameter is a mean. The claim is, "The mean course load for all community college students is less than 12 semester hours."

*Do the majority of community college students qualify for federal student loans?* This question contains a claim about a population proportion and information about the population, the variable, and the parameter. The population is all community college students. The variable is *Qualify for federal student loan* (yes or no). It is categorical, so the parameter is a proportion. The claim is, "The proportion of community college students who qualify is greater than 0.5" (a majority means more than half, or 0.5).

*In community colleges, do female students and male students have different mean GPAs?* This question contains a claim that compares two population means. Again, we see information about the populations, the variable, and the parameters. The two populations are female community college students and male community college students. The variable is *GPA*. It is quantitative, so the parameters are means. The claim is, "The mean GPA for female community college students is different from the mean GPA for male community college students." Notice that the claim compares the two population means, but there is no claim about the numeric value of either mean.

*Are college athletes more likely than nonathletes to receive academic advising?* This question contains a claim that compares two population proportions: college athletes and college students who are not athletes. The variable is *Receive academic advising* (yes or no). The variable is categorical, so the parameters are proportions. The claim is, "The proportion of all college athletes who

receive academic advising is greater than the proportion of all nonathletes in college who receive academic advising." Notice that the claim compares two population proportions, but there is no claim about the numeric value of either proportion.

In the case of testing a claim about a single population parameter, we compare it to a numeric value. In the case of testing a claim about two population parameters, we compare them to each other.

### Learn By Doing

**Identify the type of claim in each research question below.**

https://assessments.lumenlearning.co...sessments/3960

https://assessments.lumenlearning.co...sessments/3961

https://assessments.lumenlearning.co...sessments/3962

https://assessments.lumenlearning.co...sessments/3963

### Next Steps: Forming Hypotheses

We already know that in inference we use a sample to draw a conclusion about a population. If the research question contains a claim about the population, we translate the claim into two related hypotheses.

The **null hypothesis** is a hypothesis about the value of the parameter. The null hypothesis relates to our work in *Linking Probability to Statistical Inference* where we drew a conclusion about a population parameter on the basis of the sampling distribution. We started with an assumption about the value of the parameter, then used a simulation to simulate the selection of random samples from a population with this parameter value. Or we used the parameter value in a mathematical model to describe the center and spread of the sampling distribution. *The null hypothesis gives the value of the parameter that we will use to create the sampling distribution.* In this way, the null hypothesis states what we assume to be true about the population.

The **alternative hypothesis** usually reflects the claim in the research question about the value of the parameter. The alternative hypothesis says the parameter is "greater than" or "less than" or "not equal to" the value we assume to true in the null hypothesis.

### Example

## Stating Hypotheses

Here are the hypotheses for the research questions from the previous example. The null hypothesis is abbreviated $H_0$. The alternative hypothesis is abbreviated $H_a$.

*Is the average course load for community college students less than 12 semester hours?*

$H_0$: The mean course load for community college students is equal to 12 semester hours.
$H_a$: The mean course load for community college students is less than 12 semester hours.

*Do the majority of community college students qualify for federal student loans?*

$H_0$: The proportion of community college students who qualify for federal student loans is 0.5.
$H_a$: The proportion of community college students who qualify for federal student loans is greater than 0.5.

When the research question contains a claim that compares two populations, the null hypothesis states that the parameters are equal. We will see in Modules 9 and 10 that we translate the null hypothesis into a statement about "no difference" in parameter values. We revisit this idea in more depth later.

*In community colleges, do female students and male students have different mean GPAs?*

$H_0$: In community colleges, female and male students have the same mean GPAs.
$H_a$: In community colleges, female and male students have different mean GPAs.

*Are college athletes more likely than nonathletes to receive academic advising?*

$H_0$: In colleges, the proportion of athletes who receive academic advising is equal to the proportion of nonathletes who receive academic advising.

$H_a$: In colleges, the proportion of athletes who receive academic advising is greater than the proportion of nonathletes who receive academic advising.

### Comment

Here are some general observations about null and alternative hypotheses.

- The hypotheses are competing claims about the parameter or about the comparison of parameters.
- Both hypotheses are statements about the same population parameter or same two population parameters.
- The null hypothesis contains an equal sign.
- The alternative hypothesis is always an inequality statement. It contains a “less than” or a “greater than” or a “not equal to” symbol.
- In a statistical investigation, we determine the research question, and thus the hypotheses, before we collect data.

The process of forming hypotheses, collecting data, and using the data to draw a conclusion about the hypotheses is called **hypothesis testing**.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3596

https://assessments.lumenlearning.co...sessments/3597

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3598

https://assessments.lumenlearning.co...sessments/3599

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.7: Hypothesis Testing (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.8: Hypothesis Testing (2 of 5)

## Learning Objectives

- Recognize the logic behind a hypothesis test and how it relates to the P-value.

In this section, our focus is hypothesis testing, which is part of inference. On the previous page, we practiced stating null and alternative hypotheses from a research question. Forming the hypotheses is the first step in a hypothesis test. Here are the general steps in the process of hypothesis testing. We will see that hypothesis testing is related to the thinking we did in *Linking Probability to Statistical Inference*.

**Step 1: Determine the hypotheses.**

The hypotheses come from the research question.

**Step 2: Collect the data.**

Ideally, we select a random sample from the population. The data comes from this sample. We calculate a statistic (a mean or a proportion) to summarize the data.

**Step 3: Assess the evidence.**

Assume that the null hypothesis is true. Could the data come from the population described by the null hypothesis? Use simulation or a mathematical model to examine the results from random samples selected from the population described by the null hypothesis. Figure out if results similar to the data are likely or unlikely. Note that the wording "likely or unlikely" implies that this step requires some kind of probability calculation.

**Step 4: State a conclusion.**

We use what we find in the previous step to make a decision. This step requires us to think in the following way. *Remember that we assume that the null hypothesis is true.* Then one of two outcomes can occur:

- One possibility is that results similar to the actual sample are extremely unlikely. This means that the data do not fit in with results from random samples selected from the population described by the null hypothesis. In this case, it is unlikely that the data came from this population, so we view this as strong evidence against the null hypothesis. We reject the null hypothesis in favor of the alternative hypothesis.
- The other possibility is that results similar to the actual sample are fairly likely (not unusual). This means that the data fit in with typical results from random samples selected from the population described by the null hypothesis. In this case, we do not have evidence against the null hypothesis, so we cannot reject it in favor of the alternative hypothesis.

## Example

## Data Use on Smart Phones

According to an article by Andrew Berg ("Report: Teens Texting More, Using More Data," *Wireless Week*, October 15, 2010), Nielsen Company analyzed cell phone usage for different age groups using cell phone bills and surveys. Nielsen found significant growth in data usage, particularly among teens, stating that "94 percent of teen subscribers self-identify as advanced data users, turning to their cellphones for messaging, Internet, multimedia, gaming, and other activities like downloads." The study found that the mean cell phone data usage was 62 MB among teens ages 13 to 17. A researcher is curious whether cell phone data usage has increased for this age group since the original study was conducted. She plans to conduct a hypothesis test.

**Step 1: Determine the hypotheses.**

The null hypothesis is often a statement of "no change," so the null hypothesis will state that there is no change in the mean cell phone data usage for this age group since the original study. In this case, the alternative hypothesis is that the mean has increased from 62 MB.

$H_0$: The mean data usage for teens with smart phones is still 62 MB.
$H_a$: The mean data usage for teens with smart phones is greater than 62 MB.

**Step 2: Collect the data.**

The next step is to obtain a sample and collect data that will allow the researcher to test the hypotheses. The sample must be representative of the population and, ideally, should be a random sample. In this case, the researcher must randomly sample teens who use smart phones.

For the purposes of this example, imagine that the researcher randomly samples 50 teens who use smart phones. She finds that the mean data usage for these teens was 75 MB with a standard deviation of 45 MB. Since it is greater than 62 MB, this sample mean provides some evidence in favor of the alternative hypothesis. But the researcher anticipates that samples will vary when the null hypothesis is true. So how much of a difference will make her doubt the null hypothesis? Does she have evidence strong enough to reject the null hypothesis?

**Step 3: Assess the evidence.**

To assess the evidence, the researcher needs to know how much variability to expect in random samples when the null hypothesis is true. She begins with the assumption that $H_0$ is true – in this case, that the mean data usage for teens is still 62 MB. She then determines how unusual the results of the sample are: *If the mean for all teens with smart phones actually is 62 MB, what is the chance that a random sample of 50 teens will have a sample mean of 75 MB or higher?* Obviously, this probability depends on how much variability there is in random samples of this size from this population.

The probability of observing a sample mean at least this high if the population mean is 62 MB is approximately 0.023 (later topics explain how to calculate this probability). The probability is quite small. It tells the researcher that if the population mean is actually 62 MB, a sample mean of 75 MB or higher will occur only about 2.3% of the time. This probability is called the **P-value**.

Note: The P-value is a conditional probability, discussed in the module *Relationships in Categorical Data with Intro to Probability*. The condition is the assumption that the null hypothesis is true.

**Step 4: Conclusion.**

The small P-value indicates that it is unlikely for a sample mean to be 75 MB or higher if the population has a mean of 62 MB. It is therefore unlikely that the data from these 50 teens came from a population with a mean of 62 MB. The evidence is strong enough to make the researcher doubt the null hypothesis, so she rejects the null hypothesis in favor of the alternative hypothesis. *The researcher concludes that the mean data usage for teens with smart phones has increased since the original study. It is now greater than 62 MB. (P = 0.023)*

### Comment

Notice that the P-value is included in the preceding conclusion, which is a common practice. It allows the reader to see the strength of the evidence used to draw the conclusion.

### How Small Does the P-Value Have to Be to Reject the Null Hypothesis?

A small P-value indicates that it is unlikely that the actual sample data came from the population described by the null hypothesis. More specifically, a small P-value says that there is only a small chance that we will randomly select a sample with results at least as extreme as the data if $H_0$ is true. The smaller the P-value, the stronger the evidence against $H_0$.

*But how small does the P-value have to be in order to reject $H_0$?*

In practice, we often compare the P-value to 0.05. We reject the null hypothesis in favor of the alternative if the P-value is less than (or equal to) 0.05.

Note: This means that sampling variability will produce results at least as extreme as the data 5% of the time. In other words, in the long run, 1 in 20 random samples will have results that suggest we should reject $H_0$ even when $H_0$ is true. This variability is just due to chance, but it is unusual enough that we are willing to say that results this rare suggest that $H_0$ is not true.

### Statistical Significance: Another Way to Describe Unlikely Results

When the P-value is less than (or equal to) 0.05, we also say that the difference between the actual sample statistic and the assumed parameter value is **statistically significant**. In the previous example, the P-value is less than 0.05, so we say the difference between the sample mean (75 MB) and the assumed mean from the null hypothesis (62 MB) is statistically significant. You will also see this described as a **significant difference**. A significant difference is an observed difference that is too large to attribute to chance. In

other words, it is a difference that is unlikely when we consider sampling variability alone. If the difference is statistically significant, we reject $H_0$.

### Other Observations about Stating Conclusions in a Hypothesis Test

In the example, the sample mean was greater than 62 MB. This fact alone does not suggest that the data supports the alternative hypothesis. We have to determine that the data is not only larger than 62 MB but larger than we would expect to see in a random sampling if the population mean is 62 MB. We therefore need to determine the P-value. If the sample mean was less than or equal to 62 MB, it would not support the alternative hypothesis. We don't need to find a P-value in this case. The conclusion is clear without it.

We have to be very careful in how we state the conclusion. There are only two possibilities.

- We have enough evidence to reject the null hypothesis and support the alternative hypothesis.
- We do not have enough evidence to reject the null hypothesis, so there is not enough evidence to support the alternative hypothesis.

If the P-value in the previous example was greater than 0.05, then we would not have enough evidence to reject $H_0$ and accept $H_a$. In this case our conclusion would be that "there is not enough evidence to show that the mean amount of data used by teens with smart phones has increased." Notice that this conclusion answers the original research question. It focuses on the alternative hypothesis. It does *not* say "the null hypothesis is true." We never accept the null hypothesis or state that it is true. When there is not enough evidence to reject $H_0$, the conclusion will say, in essence, that "there is not enough evidence to support $H_a$." But of course we will state the conclusion in the specific context of the situation we are investigating.

We compared the P-value to 0.05 in the previous example. The number 0.05 is called the **significance level** for the test, because a P-value less than or equal to 0.05 is statistically significant (unlikely to have occurred solely by chance). The symbol we use for the significance level is α (the lowercase Greek letter alpha). We sometimes refer to the significance level as the α-level. We call this value the significance level because if the P-value is less than the significance level, we say the results of the test showed a significance difference.

**If the P-value ≤ α, we reject the null hypothesis in favor of the alternative hypothesis.**

**If the P-value > α, we fail to reject the null hypothesis.**

In practice, it is common to see 0.05 for the significance level. Occasionally, researchers use other significance levels. In particular, if rejecting $H_0$ will be controversial or expensive, we may require stronger evidence. In this case, a smaller significance level, such as 0.01, is used. As with the hypotheses, we should choose the significance level before collecting data. It is treated as an agreed-upon benchmark prior to conducting the hypothesis test. In this way, we can avoid arguments about the strength of the data.

We look more at how to choose the significance level later. On this page we continue to use a significance level of 0.05.

First let's look at some exercises that focus on the P-value and its meaning. Then we'll try some that cover the conclusion.

### Learn By Doing

For many years, working full-time has meant working 40 hours per week. Nowadays, it seems that corporate employers expect their employees to work more than this amount. A researcher decides to investigate this hypothesis.

$H_0$: The average time full-time corporate employees work per week is 40 hours.
$H_a$: The average time full-time corporate employees work per week is more than 40 hours.

To substantiate his claim, the researcher randomly selects 250 corporate employees and finds that they work an average of 47 hours per week with a standard deviation of 3.2 hours.

https://assessments.lumenlearning.co...sessments/3600

According to the Centers for Disease Control (CDC), roughly 21.5% of all high school seniors in the United States have used marijuana. (The data were collected in 2002. The figure represents those who smoked during the month prior to the survey, so the actual figure might be higher.) A sociologist suspects that the rate among African American high school seniors is lower. In this case, then,

$H_0$: The rate of African American high-school seniors who have used marijuana is 21.5% (same as the overall rate of seniors).
$H_a$: The rate of African American high-school seniors who have used marijuana is lower than 21.5%.

To check his claim, the sociologist chooses a random sample of 375 African American high school seniors,and finds that 16.5% of them have used marijuana.

https://assessments.lumenlearning.co...sessments/3601
CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.8: Hypothesis Testing (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.9: Hypothesis Testing (4 of 5)

## Learning Objectives

- Recognize the logic behind a hypothesis test and how it relates to the P-value.

Hypothesis testing appears in all upcoming modules. The process and the logic of the hypothesis test will always be the same, but the details will differ somewhat.

Every hypothesis test will use a P-value to make a decision about the population(s). The P-value is the connection between probability and decision-making in inference. Now we discuss the P-value in more depth and relate it to our work in *Linking Probability to Statistical Inference*. Later we use both simulations and statistical software to find the P-value.

To develop a better understanding of the P-value, we need to return to the idea of a sampling distribution and a normal probability model. These are ideas from *Linking Probability to Statistical Inference*.

## Example

## What Is a P-value?

Let's return to the familiar example of the 2008 presidential election. In that election, newspapers reported that Obama received 40% of the white male vote. We wonder if a smaller percentage of white males will support Obama in the 2012 election. We define the following hypotheses and conduct a hypothesis test.

$H_0$: The proportion of white males voting for Obama in 2012 is 0.40.
$H_a$: The proportion of white males voting for Obama in 2012 is less than 0.40.

We select a random sample of 200 white male voters and find that 35% plan to vote for Obama in 2012. Clearly 35% is less than 40%. But is the difference statistically significant or due to chance? If the population proportion is 0.40, we expect to see sample proportions vary from this. But will sample proportions as small as or smaller than 0.35 occur very often? What's the probability?

The probability (P-value) is about 0.078.

The P-value is the chance that a random sample of 200 white males will have, at most, 35% supporting Obama if 40% of this population supports Obama. This is quite a mouthful. We find that visualizing the sampling distribution helps us understand the P-value. Here is a diagram that may be helpful in interpreting the P-value.

In general, the P-value is the probability that sample results are as extreme as or more extreme than the result observed in the data if the null hypothesis is true. The phrase "as extreme as or more extreme than" means further from the center of the sampling distribution in the direction of the alternative hypothesis.

**Note:** You may recall the concept of a conditional probability from *Relationships in Categorical Data with Intro to Probability*. The P-value is a conditional probability. The condition is "the null hypothesis is true."

**Note:** We can also look at the P-value in terms of error in the sample proportion. If 40% of this population support Obama, then our sample with 35% supporting Obama has a 5% error. From this perspective, the P-value is the chance that sample proportions *supporting the alternative hypothesis* have as much as or more error than the data. For this example, the P-value describes sample proportions less than 0.40 that deviate 0.05 or more from 0.40. This describes sample proportions at or below 0.35.

## Comment

Recall in *Linking Probability to Statistical Inference* when we investigated the conditions that make a normal model a good fit for the sampling distribution. When a normal model is a good fit, we use it to find probabilities. In the Obama example, we can see that a normal model is a good fit for the sampling distribution, so we can find the P-value by calculating the *z*-score and using a simulation. Below we created a diagram to remind you how the sampling distribution relates to the standard normal model of *z*-scores. We will find P-values in this way in "Hypothesis Test for a Population Proportion," but for now, we focus on what the P-value is and how to use it to make decisions.

Notice that for our example , the P-value from the standard normal curve (0.074) is not exactly equal to the relative frequency in the simulated sampling distribution (0.078), but it is close. Both of these values represent estimates of the probability we want.

In hypothesis testing you use the standard normal curve (or a similar model) to find P-values. For this reason, you will frequently see the P-value defined in terms of the "test statistic," which is the *z*-score in our example. Here is a common definition: The P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3606

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3908

### Comment

The purpose of the hypothesis test is to describe the degree of evidence that the sample provides against the null hypothesis. The P-value does this job. In the next two activities, we do not give a P-value. We want you to practice visualizing the sampling distribution to identify the most convincing evidence against the null hypothesis. This type of visualization is important to understanding the ideas we discuss on this page.

### Example

## More on Using the P-Value to Make a Decision

Let's finish the hypothesis test about white male support for Obama.

$H_0$: The proportion of white males voting for Obama in 2012 is 0.40.
$H_a$: The proportion of white males voting for Obama in 2012 is less than 0.40.

Recall our random sample of 200 white male voters with 35% planning to vote for Obama in 2012. Clearly 35% is less than 40%, but is the difference statistically significant or due to chance?

The probability (P-value) is about 0.078. What can we conclude? A small P-value indicates that the data are unlikely to occur in random sampling from a population in which the null hypothesis is true. So the smaller the P-value, the stronger the evidence is against the null hypothesis.

Do you view something that happens 7.8% of the time as "unlikely"? This is a judgment call. There is no right or wrong answer. Because this is a judgment call, we will often agree to a definition of "unlikely" before we run the test. This is the significance level, α. The significance level is a benchmark for how small the P-value must be in order for us to say results are statistically significant. It gives us a cutoff point for rejecting the null hypothesis. Here are the decision-making rules that we gave earlier.

If the P-value ≤ α, we reject the null hypothesis in favor of the alternative hypothesis.

If the P-value > α, we fail to reject the null hypothesis.

Why do these "rules" make sense? Again, we think that visualizing a simulation of the sampling distribution is helpful.

Keep in mind that all of the sample proportions in the simulation did actually occur when we selected random samples from a population with $P = 0.40$. Results less than 0.40 support the alternative hypothesis but to varying degrees. To provide enough evidence to reject the null hypothesis and accept the alternative hypothesis, results have to be smaller than 0.40 and "rare." When we set the significance level at 5% ($\alpha = 0.05$), we agree to view results that occur less than 5% of the time as "rare enough" to question whether the sample came from the population described by the null. So we reject the null hypothesis and accept the alternative. If we set the significance level at 10% ($\alpha = 0.10$), we have changed the definition of rare. As you can see, different significance levels can lead to different conclusions.

Here are our conclusions for the two different levels of significance.

At the 5% level, our poll results are not statistically significant (P-value = 0.078). We conclude that white male support for Obama will not be less than 40% in 2012. (Note: This statement says we do not have enough evidence to accept $H_a$. Because $H_a$ is related to the claim or hunch that motivated our investigation, we state our conclusion in terms of $H_a$.)

At the 10% level, our poll results are statistically significant (P-value = 0.078). We conclude that white male support for Obama will be less than 40% in 2012. (Note: This statement says we have enough evidence to accept $H_a$. Again, the conclusion is stated in terms of $H_a$.)

### How Do You Choose a Level of Significance?

Remember that the purpose of the hypothesis test is to describe the degree of evidence that the sample provides against the null hypothesis. The P-value does this. How small a P-value is convincing evidence? It depends on the situation and the opinions of the people who use the hypothesis test to make a decision. If rejecting the null hypothesis will be controversial or expensive, then the users of the test results may want to use a smaller level of significance than we did. For this reason, we always report the P-value with our conclusions so that the people who use the results of the test can determine if the P-value is strong enough evidence for their purpose.

Later we discuss other considerations for choosing a level of significance.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3609

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.9: Hypothesis Testing (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.10: Hypothesis Testing (3 of 5)

## Learning Objectives

- Recognize the logic behind a hypothesis test and how it relates to the P-value.

## Example

## Community College Students and Federal Student Loans

According to the Project on Student Debt, "at least one million community college students, one in 10 nationally, do not have access to federal student loans – the safest, most affordable way to borrow for college. A new issue brief from the Project on Student Debt finds that almost a quarter of all community colleges do not participate in federal loan programs, thereby forcing needy students to resort to riskier, more expensive options such as private student loans and credit cards" *(Source: Project on Student Debt, Press release, April 17, 2008).*

*Is the proportion of community colleges that do not participate in federal loan programs less than 25%, as reported?* Let's conduct a hypothesis test to find out.

**Step 1: Determine the Hypotheses.**

$H_0$: The proportion of community colleges that do not participate in federal loan programs is 0.25.
$H_a$: The proportion of community colleges that do not participate in federal loan programs is less than 0.25.

**Step 2: Collect the data.**

For the purposes of this example, imagine that we select a random sample of 80 community colleges from the over 1,100 community colleges in the United States. Of the 80, suppose that 16 do not participate in federal loan programs, so the sample proportion is 0.20.

Because this sample proportion is less than 0.25, it provides evidence in favor of the alternative hypothesis. But we anticipate that samples will vary when the null hypothesis is true. How much of a difference will make us doubt the null hypothesis? Do we have evidence strong enough to reject the null hypothesis and accept the alternative hypothesis?

**Step 3: Assess the evidence.**

To assess the evidence, we need to know how much variability to expect in random samples when the null hypothesis is true. We begin with the assumption that $H_0$ is true. In this case, we assume that 25% of community colleges do not participate in the federal loan programs. We then determine how unusual the results of the sample are. We ask, *If the proportion of all community colleges without federal loan programs is 0.25, what is the chance that the proportion in a random sample of 80 community colleges is 0.20 or less?* Obviously, this probability depends on how much variability exists in random samples of this size from this population.

The probability of observing a sample proportion at least this small if the population proportion is 0.25 is approximately 0.15 (upcoming topics explain how to calculate this probability). This is the P-value. It tells us that if the population proportion is actually 0.25, we will see a sample proportion of 0.20 or less about 15% of the time in random sampling.

Note: The P-value is a conditional probability. The condition is the assumption that the null hypothesis is true – in this case, that the population proportion is 0.25.

**Step 4: Conclusion.**

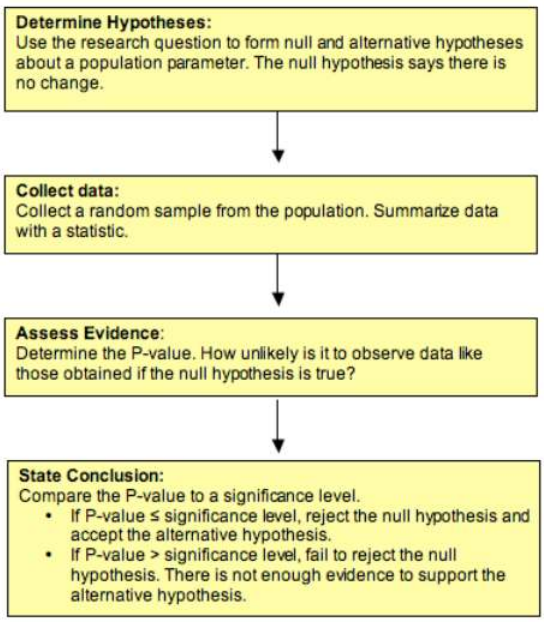
Note that the P-value is fairly large, so it is not surprising to see a sample proportion of 0.20 or lower if the population proportion is 0.25. If we use a significance level of 0.05, the P-value is larger than 0.05, so the difference we observe between the sample proportion and the assumed population proportion is not statistically significant. Differences this large can be explained by chance. We fail to reject the null hypothesis. Here is our conclusion.

*The data do not provide significant evidence that the proportion of community colleges without federal loan programs is less than 25%.*

Note: The conclusion answers our original research question. It focuses on the claim that is the alternative hypothesis. It does *not* say "the null hypothesis is true." We never accept the null hypothesis or state that it is true. When there is not enough evidence to reject $H_0$, the conclusion will say, in essence, "that there is not enough evidence to support $H_a$."

## Summary

Now that we have seen two hypothesis tests, let's summarize the steps:




## Learn By Doing

The following two hypotheses are tested:

$H_0$: The proportion of U.S. adults who support gay marriage is roughly 50%.
$H_a$: The proportion of U.S. adults who support gay marriage is above 50% (i.e., the majority support).

Suppose a survey was conducted in which a random sample of 1,100 U.S. adults were asked about their opinions on gay marriage, and based on the data, the P-value was found to be 0.002.

Comment: Throughout this activity, use a 0.05 (5%) significance level (cutoff).

https://assessments.lumenlearning.co...sessments/3602

https://assessments.lumenlearning.co...sessments/3603

https://assessments.lumenlearning.co...sessments/3604

https://assessments.lumenlearning.co...sessments/3605

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.10: Hypothesis Testing (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.11: Hypothesis Testing (5 of 5)

## Learning Objectives

- Recognize type I and type II errors.

## What Can Go Wrong: Two Types of Errors

Statistical investigations involve making decisions in the face of uncertainty, so there is always some chance of making a wrong decision. In hypothesis testing, two types of wrong decisions can occur.

If the null hypothesis is true, but we reject it, the error is a **type I** error.

If the null hypothesis is false, but we fail to reject it, the error is a **type II** error.

The following table summarizes type I and II errors.

## Comment

Type I and type II errors are not caused by mistakes. These errors are the result of random chance. The data provide evidence for a conclusion that is false. It's no one's fault!

## Example

### Data Use on Smart Phones

In a previous example, we looked at a hypothesis test about data usage on smart phones. The researcher investigated the claim that the mean data usage for all teens is greater than 62 MBs. The sample mean was 75 MBs. The P-value was approximately 0.023. In this situation, the P-value is the probability that we will get a sample mean of 75 MBs or higher if the true mean is 62 MBs.

Notice that the result (75 MBs) isn't impossible, only very unusual. The result is rare enough that we question whether the null hypothesis is true. This is why we reject the null hypothesis. But it is possible that the null hypothesis hypothesis is true and the researcher happened to get a very unusual sample mean. In this case, the result is just due to chance, and the data have led to a type I error: rejecting the null hypothesis when it is actually true.

## Example

### White Male Support for Obama in 2012

In a previous example, we conducted a hypothesis test using poll results to determine if white male support for Obama in 2012 will be less than 40%. Our poll of white males showed 35% planning to vote for Obama in 2012. Based on the sampling distribution, we estimated the P-value as 0.078. In this situation, the P-value is the probability that we will get a sample proportion of 0.35 or less if 0.40 of the population of white males support Obama.

At the 5% level, the poll did not give strong enough evidence for us to conclude that less than 40% of white males will vote for Obama in 2012.

Which type of error is possible in this situation? If, in fact, it is true that less than 40% of this population support Obama, then the data led to a type II error: failing to reject a null hypothesis that is false. In other words, we failed to accept an alternative hypothesis that is true.

We definitely did not make a type I error here because a type I error requires that we reject the null hypothesis.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3610

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3909

### What Is the Probability That We Will Make a Type I Error?

If the significance level is 5% ($\alpha = 0.05$), then 5% of the time we will reject the null hypothesis (when it is true!). Of course we will not know if the null is true. But if it is, the natural variability that we expect in random samples will produce rare results 5% of the time. This makes sense because we assume the null hypothesis is true when we create the sampling distribution. We look at the variability in random samples selected from the population described by the null hypothesis.

Similarly, if the significance level is 1%, then 1% of the time sample results will be rare enough for us to reject the null hypothesis hypothesis. So if the null hypothesis is actually true, then by chance alone, 1% of the time we will reject a true null hypothesis. The probability of a type I error is therefore 1%.

**In general, the probability of a type I error is $\alpha$.**

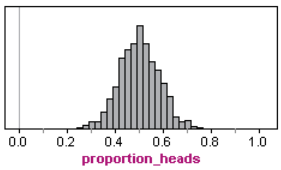
### What Is the Probability That We Will Make a Type II Error?

The probability of a type I error, if the null hypothesis is true, is equal to the significance level. The probability of a type II error is much more complicated to calculate. We can reduce the risk of a type I error by using a lower significance level. The best way to reduce the risk of a type II error is by increasing the sample size. In theory, we could also increase the significance level, but doing so would increase the likelihood of a type I error at the same time. We discuss these ideas further in a later module.

### Learn By Doing

## A Fair Coin

In the long run, a fair coin lands heads up half of the time. (For this reason, a weighted coin is not fair.) We conducted a simulation in which each sample consists of 40 flips of a fair coin. Here is a simulated sampling distribution for the proportion of heads in 2,000 samples. Results ranged from 0.25 to 0.75.




https://assessments.lumenlearning.co...sessments/3910

https://assessments.lumenlearning.co...sessments/3911

https://assessments.lumenlearning.co...sessments/3912

### Comment

In general, if the null hypothesis is true, the significance level gives the probability of making a type I error. If we conduct a large number of hypothesis tests using the same null hypothesis, then, a type I error will occur in a predictable percentage ($\alpha$) of the hypothesis tests. This is a problem! If we run one hypothesis test and the data is significant at the 5% level, we have reasonably good evidence that the alternative hypothesis is true. If we run 20 hypothesis tests and the data in one of the tests is significant at the 5% level, it doesn't tell us anything! We expect 5% of the tests (1 in 20) to show significant results just due to chance.

### Example

## Cell Phones and Brain Cancer

The following is an excerpt from a 1999 *New York Times* article titled "Cell phones: questions but no answers," as referenced by David S. Moore in *Basic Practice of Statistics* (4th ed., New York: W. H. Freeman, 2007):

> *A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant association between cell phone use and a group of brain cancers known as gliomas. But when 20 types of glioma were considered separately, an association was found between cell phone use and one rare form. Puzzlingly, however, this risk appeared to decrease rather than increase with greater mobile phone use.*

This is an example of a probable type I error. Suppose we conducted 20 hypotheses tests with the null hypothesis "Cell phone use is not associated with cancer" at the 5% level. We expect 1 in 20 (5%) to give significant results by chance alone when there is no

association between cell phone use and cancer. So the conclusion that this one type of cancer is related to cell phone use is probably just a result of random chance and not an indication of an association.

Click here to see a fun cartoon that illustrates this same idea.

### Learn By Doing

## How Many People Are Telepathic?

Telepathy is the ability to read minds. Researchers used Zener cards in the early 1900s for experimental research into telepathy.

In a telepathy experiment, the "sender" looks at 1 of 5 Zener cards while the "receiver" guesses the symbol. This is repeated 40 times, and the proportion of correct responses is recorded. Because there are 5 cards, we expect random guesses to be right 20% of the time (1 out of 5) in the long run. So in 40 tries, 8 correct guesses, a proportion of 0.20, is common. But of course there will be variability even when someone is just guessing. Thirteen or more correct in 40 tries, a proportion of 0.325, is statistically significant at the 5% level. When people perform this well on the telepathy test, we conclude their performance is not due to chance and take it as an indication of the ability to read minds.

https://assessments.lumenlearning.co...sessments/3611

In the next section, "Hypothesis Test for a Population Proportion," we learn the details of hypothesis testing for claims about a population proportion. Before we get into the details, we want to step back and think more generally about hypothesis testing. We close our introduction to hypothesis testing with a helpful analogy.

### A Courtroom Analogy for Hypothesis Tests

When a defendant stands trial for a crime, he or she is innocent until proven guilty. It is the job of the prosecution to present evidence showing that the defendant is guilty *beyond a reasonable doubt*. It is the job of the defense to challenge this evidence to establish a reasonable doubt. The jury weighs the evidence and makes a decision.

When a jury makes a decision, it has only two possible verdicts:

- **Guilty:** The jury concludes that there is enough evidence to convict the defendant. The evidence is so strong that there is not a reasonable doubt that the defendant is guilty.
- **Not Guilty:** The jury concludes that there is not enough evidence to conclude beyond a reasonable doubt that the person is guilty. Notice that they do not conclude that the person is innocent. This verdict says only that there is not enough evidence to return a guilty verdict.

*How is this example like a hypothesis test?*

The null hypothesis is "The person is innocent." The alternative hypothesis is "The person is guilty." The evidence is the data. In a courtroom, the person is assumed innocent until proven guilty. In a hypothesis test, we assume the null hypothesis is true until the data proves otherwise.

The two possible verdicts are similar to the two conclusions that are possible in a hypothesis test.

**Reject the null hypothesis:** When we reject a null hypothesis, we accept the alternative hypothesis. This is like a guilty verdict. The evidence is strong enough for the jury to reject the assumption of innocence. In a hypothesis test, the data is strong enough for us to reject the assumption that the null hypothesis is true.

**Fail to reject the null hypothesis:** When we fail to reject the null hypothesis, we are delivering a "not guilty" verdict. The jury concludes that the evidence is not strong enough to reject the assumption of innocence, so the evidence is too weak to support a guilty verdict. We conclude the data is not strong enough to reject the null hypothesis, so the data is too weak to accept the alternative hypothesis.

*How does the courtroom analogy relate to type I and type II errors?*

**Type I error:** The jury convicts an innocent person. By analogy, we reject a true null hypothesis and accept a false alternative hypothesis.

**Type II error:** The jury says a person is not guilty when he or she really is. By analogy, we fail to reject a null hypothesis that is false. In other words, we do not accept an alternative hypothesis when it is really true.

## Let's Summarize

In this section, we introduced the four-step process of hypothesis testing:

**Step 1: Determine the hypotheses.**

- The hypotheses are claims about the population(s).
- The null hypothesis is a hypothesis that the parameter equals a specific value.
- The alternative hypothesis is the competing claim that the parameter is less than, greater than, or not equal to the parameter value in the null. The claim that drives the statistical investigation is usually found in the alternative hypothesis.

**Step 2: Collect the data.**

Because the hypothesis test is based on probability, random selection or assignment is essential in data production.

**Step 3: Assess the evidence.**

- Use the data to find a P-value.
- The P-value is a probability statement about how unlikely the data is if the null hypothesis is true.
- More specifically, the P-value gives the probability of sample results at least as extreme as the data if the null hypothesis is true.

**Step 4: Give the conclusion.**

- A small P-value says the data is unlikely to occur if the null hypothesis is true. We therefore conclude that the null hypothesis is probably not true and that the alternative hypothesis is true instead.
- We often choose a significance level as a benchmark for judging if the P-value is small enough. If the P-value is less than or equal to the significance level, we reject the null hypothesis and accept the alternative hypothesis instead.
- If the P-value is greater than the significance level, we say we "fail to reject" the null hypothesis. We never say that we "accept" the null hypothesis. We just say that we don't have enough evidence to reject it. This is equivalent to saying we don't have enough evidence to support the alternative hypothesis.
- Our conclusion will respond to the research question, so we often state the conclusion in terms of the alternative hypothesis.

Inference is based on probability, so there is always uncertainty. Although we may have strong evidence against it, the null hypothesis may still be true. If this is the case, we have a **type I** error. Similarly, even if we fail to reject the null hypothesis, it does not mean the alternative hypothesis is false. In this case, we have a **type II** error. These errors are not the result of a mistake in conducting the hypothesis test. They occur because of random chance.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.11: Hypothesis Testing (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.12: Introduction- Hypothesis Test for a Population Proportion

## What you'll learn to do: Conduct a hypothesis test for a population proportion.

### LEARNING OBJECTIVES

- Recognize when a situation calls for testing a hypothesis about a population proportion.
- Conduct a hypothesis test for a population proportion. State a conclusion in context.
- Interpret the P-value as a conditional probability in the context of a hypothesis test about a population proportion.
- Distinguish statistical significance from practical importance.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.12: Introduction- Hypothesis Test for a Population Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.13: Hypothesis Test for a Population Proportion (1 of 3)

## Learning Objectives

- Recognize when a situation calls for testing a hypothesis about a population proportion.
- Conduct a hypothesis test for a population proportion. State a conclusion in context.

## Introduction

In the previous section, we introduced the concept of hypothesis testing. In a hypothesis test, we test competing claims about a population parameter or the difference between two population parameters.

We looked at four hypothesis testing situations:

- Testing a claim about a single population proportion.
- Testing a claim about a single population mean.
- Testing a claim about the difference between two population proportions.
- Testing a claim about the difference between two population means.

Although we follow the four steps we examined in the previous section, "Hypothesis Testing," for each of these situations, the specifics for each test are different. In this section, we look at the hypothesis test for a single population proportion. When we conduct a test about a population proportion, we are working with a categorical variable. Later in the course, after we have learned a variety of hypothesis tests, we will need to be able to identify which test is appropriate for which situation. Identifying the variable as categorical or quantitative is an important component of choosing an appropriate hypothesis test. We also have to distinguish between testing a claim about a population proportion and estimating a population proportion.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3612

Once we know that we are dealing with a single population proportion, we can conduct the hypothesis test. Recall that the first step of a hypothesis test is to determine the hypotheses. In the previous section, our hypotheses were in words. In this section, we use symbols. Recall that the symbol for the population proportion is $p$.

## Example

### Health Insurance Coverage

According to the Government Accountability Office, 80% of all college students ages 18 to 23 had health insurance coverage in 2006. The Patient Protection and Affordable Care Act passed in 2010 allowed young people under age 26 to stay on their parents' health insurance policy. Has the proportion of college students ages 18 to 23 who have health insurance increased since 2006? A survey of 800 randomly selected college students ages 18 to 23 indicated that 83% of them had health insurance coverage.

$H_0$: $p = 0.80$ (No change; the proportion of college students ages 18 to 23 who have health insurance is still 80%.)
$H_a$: $p > 0.80$ (The proportion of college students ages 18 to 23 who have health insurance is now greater than 80%.)

The results of the survey do not affect our hypotheses. We use the results to determine whether to reject the null hypothesis in favor of the alternative hypothesis.

## Example

### Internet Access

According to the Kaiser Family Foundation, 84% of U.S. children ages 8 to 18 had Internet access at home as of August 2009. Researchers wonder if this percentage has changed since then. They survey 500 randomly selected children ages 8 to 18 and find that 430 of them have Internet access at home. The research question helps us form our hypotheses:

$H_0$: $p = 0.84$ (No change; the proportion of children with Internet access at home is the same.)
$H_a$: $p \neq 0.84$ (The proportion of children with Internet access at home has changed since 2009.)

Again, the results of the survey do not affect our hypotheses.

Example

## Jury Selection

Jefferson Parish is a suburb of New Orleans, Louisiana. Its population is about 23% African American. Is there evidence that African Americans are underrepresented on juries in murder trials in Jefferson Parish? According to a *New York Times* article (June 4, 2007), there were 18 murder trials in Jefferson Parish between 1986 and 2007 in which the ethnicity of the jurors was known. Ten of the juries had no black jurors, 7 juries had 1 black juror, and 1 jury had 2 black jurors. The research question helps us to form our hypotheses:

$H_0$: $p = 0.23$ (No difference; the proportion of African Americans on juries in murder trials is the same as the proportion of African Americans in the population.)
$H_a$: $p < 0.23$ (The proportion of African Americans on juries in murder trials is less than the proportion of African Americans in the population.)

### Summary of Hypotheses

As a reminder, the null hypothesis is always a statement of equality. The alternative hypothesis is always a statement of inequality, using <, >, or ≠. So hypotheses take the form:

$H_0$: $p = p_0$
$H_a$: $p < p_0$ or $p > p_0$ or $p \neq p_0$

We use $p_0$ to represent the proportion from the null hypothesis.

### College Students and Federal Grants

According to the American Association of Community Colleges, 23% of community college students receive federal grants. The California Community College Chancellor's Office anticipates that the percentage is smaller for California community college students. They collect a sample of 1,000 community college students in California and find that 210 received federal grants.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.13: Hypothesis Test for a Population Proportion (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.14: Hypothesis Test for a Population Proportion (2 of 3)

## Learning Objectives

- Conduct a hypothesis test for a population proportion. State a conclusion in context.

On the previous page, we looked at determining hypotheses for testing a claim about a population proportion. On this page, we look at how to determine P-values.

As we learned earlier, the P-value for a hypothesis test for a population proportion comes from a normal model for the sampling distribution of sample proportions. The normal distribution is an appropriate model for this sampling distribution if the expected number of success and failures are both at least 10. Using the symbols for the population proportion and sample size, a normal curve is a reasonable model if the following conditions are met: $np \geq 10$ and $n(1 - p) \geq 10$.

## Example

## Health Insurance Coverage

Recall this example from the previous page. According to the Government Accountability Office, 80% of all college students (ages 18 to 23) had health insurance in 2006. The Patient Protection and Affordable Care Act of 2010 allowed young people under age 26 to stay on their parents' health insurance policy. Has the proportion of college students (ages 18 to 23) who have health insurance increased since 2006? A survey of 800 randomly selected college students (ages 18 to 23) indicated that 83% of them had health insurance. Use a 0.05 level of significance.

**Step 1: Determine the hypotheses.**

We did this on the previous page. The hypotheses are:

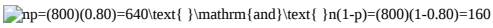
$H_0$: $p = 0.80$
$H_a$: $p > 0.80$

where $p$ is the proportion of college students ages 18 to 23 who have health insurance now.

**Step 2: Collect the data.**

In this random sample of 800 college students, 83% have health insurance. If 80% of all college students have health insurance, is this 3% difference statistically significant or due to chance? We need to find a P-value to answer this question. We must determine if we can use this data in a hypothesis test.

First note that the data are from a random sample. That is essential. Now we need to determine if a normal model is a good fit for the sampling distribution. Since we assume that the null hypothesis is true, we build the sampling distribution with the assumption that 0.80 is the population proportion. We check the following conditions, using 0.80 for *p:*

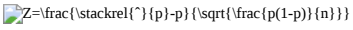



Because these are both more than 10, we can use the normal model to find the P-value.

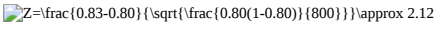
**Step 3: Assess the evidence.**

Now that we know that the normal distribution is an appropriate model for the sampling distribution, our next goal is to determine the P-value. The first step is to determine the *z*-score for the observed sample proportion (the data).

The sample proportion is 0.83. Recall from *Linking Probability to Statistical Inference* that the formula for the *z*-score of a sample proportion is as follows:




For this example, we calculate:




This *z*-score is called the **test statistic**. It tells us the sample proportion of 0.83 is about 2.12 standard errors above the population proportion given in the null hypothesis. We use this statistic to find the P-value. The P-value describes the strength of the evidence

against the null hypothesis.

We use the simulation that we first saw in *Probability and Probability Distributions* to determine the P-value. The P-value is a probability that describes the likelihood of the data if the null hypothesis is true. More specifically, the P-value is the probability that sample results are as extreme as or more extreme than the data if the null hypothesis is true. The phrase "as extreme as or more extreme than" means farther from the center of the sampling distribution in the direction of the alternative hypothesis.

In this situation, we want the area to the right of 0.83 because the alternative hypothesis is a "greater-than" statement. The P-value, in this case, is the probability of getting a sample proportion equal to or greater than 0.83. Since we are using the standard normal curve to find probabilities, the P-value is the area to the right of the $Z = 2.12$.

We can find this area with a simulation or other technology.

The P-value is approximately 0.0170. Thus, the probability that a random sample proportion is at least as large as 0.83 is about 0.017 (if the population proportion is actually 0.80). If the null hypothesis is true, we observe sample proportions this high or higher only about 1.7% of the time.

The P-value is our evidence of statistical significance. It is a measure of whether random chance can explain the deviation of the data from the null hypothesis.

**Step 4: State a conclusion.**

To determine our conclusion, we compare the P-value to the level of significance, $\alpha = 0.05$. If our data are predicted to occur by chance less than 5% of the time, we have reason to reject the null hypothesis and accept the alternative. Since our P-value of 0.017 is less than 0.05, we reject the null hypothesis. We state our conclusion in terms of the alternative hypothesis. We also state it in context.

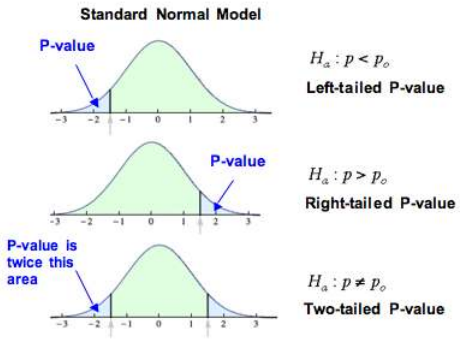
The data from this study provides strong evidence that the proportion of all college students who have health insurance is now greater than 0.80 (P-value = 0.017). The 0.03 increase in the proportion who have health insurance since 2008 is statistically significant at the 0.05 level.

Alternatively, we can give the conclusion using the percentage rather than the decimal:

The data from this study provides strong evidence that the percentage of all college students who have health insurance is now greater than 80% (P-value = 0.017). The 3% increase in the percentage who have health insurance since 2008 is statistically significant at the 5% level.

## Important Note

A hypothesis test can be **one-tailed** or **two-tailed**. The previous example was a one-tailed hypothesis test. The P-value was the area of the right tail. If the inequality in the alternative hypothesis is < or >, the test is one-tailed. If the inequality is ≠, the test is two-tailed.




p0, we have a right-tailed P-value so the P-value is an area on the right tail of the curve. When Ha has p ≠ p0, we have a two-tailed P-value. So, the P-value is twice the area of one tail." width="401″ height="296″>

## Internet Access

Recall the following example from the previous page. According to the Kaiser Family Foundation, 84% of U.S. children iages 8 to 18 had Internet access at home as of August 2009. Researchers wonder if this percentage has changed since then. They survey 500 randomly selected children (ages 8 to 18) and find that 430 of them have Internet access at home.

Use a level of significance of $\alpha = 0.05$ for this hypothesis test.

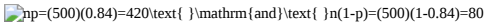
**Step 1: Determine the hypotheses.**

$H_0$: $p = 0.84$
$H_a$: $p \neq 0.84$

where $p$ is the proportion of children ages 8 to 18 with Internet access at home now.

**Step 2: Collect the data.**

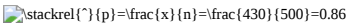
Our sample is random, so there is no problem there. Again, we want to determine whether the normal model is a good fit for the sampling distribution of sample proportions. Based on the null hypothesis, we will use 0.84 as our population proportion to check the conditions.



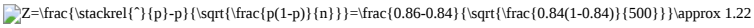

Because these are both more than 10, we can use the normal model to find the P-value.

**Step 3: Assess the evidence.**

Since we can use the normal model, we need to calculate the $z$-test statistic for the sample proportion. We first calculate the sample proportion.




Next, we calculate our $Z$-score, the test statistic:




The sample proportion of 0.86 is about 1.22 standard errors above the population proportion given in the null hypothesis. Now we calculate the P-value. This is where the two-tailed nature of the test is important. The P-value is the probability of seeing a sample proportion at least as extreme as the one observed from the data if the null hypothesis is true.

In the previous example, only sample proportions higher than the null proportion were evidence in favor of the alternative hypothesis. In this example, any sample proportion that differs from 0.84 is evidence in favor of the alternative. Statistically significant differences are at least as extreme as the difference we see in the data. We want to determine the probability that the difference in either direction (above or below 0.84) is at least as large as the difference seen in the data, so we include sample proportions at or above 0.86 *and* sample proportions at or below 0.82. For this reason, we look at the area in both tails. Our simulation shows one tail, so we have to double this area.

The area above the test statistic of 1.22 is about 0.11. We double this area to include the area in the left tail, below $Z = -1.22$. This gives us a P-value of approximately 0.22.

Our sample proportion was 0.02 above the population proportion from the null hypothesis. In a sample of size 500, we would observe a sample proportion 0.02 or more away from 0.84 about 22% of the time by chance alone.

**Step 4: State a conclusion.**

Again we compare the P-value to the level of significance, $\alpha = 0.05$. In this case, the P-value of 0.22 is greater than 0.05, which means we do not have enough evidence to reject the null hypothesis. A sample result that could occur 22% of the time by chance alone is not statistically significant. Now we can state the conclusion in terms of the alternative hypothesis.

The data from this study does not provide evidence that is strong enough to conclude that the proportion of all children ages 8 to 18 who have Internet access at home has changed since 2009 (P-value = 0.22). The 2% change observed in the data is not statistically significant. These results can be explained by predictable variation in random samples.

### A Note about the Conclusion

In the conclusion above, we did not have enough evidence to reject the null hypothesis. As we noted in "Hypothesis Testing," failing to reject the null hypothesis does not mean the null hypothesis is true.

In the case of the previous example, it is possible that the proportion of children who have Internet access at home has changed. But the data we gathered did not provide the evidence to detect that the proportion had changed significantly.

Researchers often note improvements that could be made in their research and suggest follow-up research that might be done. In our example, a second sample with a larger sample size might provide the evidence needed to reject the null hypothesis.

The important thing to keep in mind is that at the end of a hypothesis test, we *never* say that the null hypothesis is true.

### Learn By Doing

## California College Students Who Drink

According to the Centers for Disease Control and Prevention, 60% of all American adults ages 18 to 24 currently drink alcohol. Is the proportion of California college students who currently drink alcohol different from the proportion nationwide? A survey of 450 California college students indicates that 66% currently drink alcohol. The hypotheses were:

$H_0$: $p = 0.60$
$H_a$: $p \neq 0.60$

https://assessments.lumenlearning.co...sessments/3613

Click here to open the simulation

https://assessments.lumenlearning.co...sessments/3614

https://assessments.lumenlearning.co...sessments/3913

### Learn By Doing

## Coin Flips

Recall the scenario from the previous page. A psychic claims to be able to predict the outcome of coin flips before they happen. Someone who guesses randomly will predict about half of coin flips correctly. In 100 flips, the psychic correctly predicts 57 flips. Do the results of this test indicate that the psychic does better than random guessing? The hypotheses are

$H_0$: $p = 0.50$
$H_a$: $p > 0.50$

where $p$ is the proportion of correct coin flip predictions by the psychic.

https://assessments.lumenlearning.co...sessments/3615

Click here to open the simulation

https://assessments.lumenlearning.co...sessments/3616

https://assessments.lumenlearning.co...sessments/3914

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3617

https://assessments.lumenlearning.co...sessments/3618

https://assessments.lumenlearning.co...sessments/3619

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.14: Hypothesis Test for a Population Proportion (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.15: Hypothesis Test for a Population Proportion (3 of 3)

## Learning Objectives

- Conduct a hypothesis test for a population proportion. State a conclusion in context.
- Interpret the P-value as a conditional probability in the context of a hypothesis test about a population proportion.
- Distinguish statistical significance from practical importance.
- From a description of a study, evaluate whether the conclusion of a hypothesis test is reasonable.

## More about the P-Value

The P-value is a probability that describes the likelihood of the data if the null hypothesis is true. More specifically, the P-value is the probability that sample results are as extreme as or more extreme than the data if the null hypothesis is true. The phrase "as extreme as or more extreme than" means farther from the center of the sampling distribution in the direction of the alternative hypothesis.

More generally, we view the P-value a description of the strength of the evidence against the null hypothesis and in support of the alternative hypothesis. But the P-value is a probability about sample results, not about the null or alternative hypothesis.

## One More Note about P-Values and the Significance Level

You may wonder why 5% is often selected as the significance level in hypothesis testing and why 1% is also a commonly used level. It is largely due to just convenience and tradition. When Ronald Fisher (one of the founders of modern statistics) published one of his tables, he used a mathematically convenient scale that included 5% and 1%. Later, these same 5% and 1% levels were used by other people, in part just because Fisher was so highly esteemed. But mostly, these are arbitrary levels.

The idea of selecting some sort of relatively small cutoff was historically important in the development of statistics. But it's important to remember that there is really a continuous range of increasing confidence toward the alternative hypothesis, not a single all-or-nothing value. There isn't much meaningful difference, for instance, between the P-values 0.049 and 0.051, and it would be foolish to declare one case definitely a "real" effect and the other case definitely a "random" effect. In either case, the study results are roughly 5% likely by chance if there's no actual effect.

Whether such a P-value is sufficient for us to reject a particular null hypothesis ultimately depends on the risk of making the wrong decision and the extent to which the hypothesized effect might contradict our prior experience or previous studies.

## Example

## Sample Size and Hypothesis Testing

Consider our earlier example about teenagers and Internet access. According to the Kaiser Family Foundation, 84% of U.S. children ages 8 to 18 had Internet access at home as of August 2009. Researchers wonder if this number has changed since then. The hypotheses we tested were:

$H_0$: $p = 0.84$
$H_a$: $p \neq 0.84$

The original sample consisted of 500 children, and 86% of them had Internet access at home. The P-value was about 0.22, which was not strong enough to reject the null hypothesis. There was not enough evidence to show that the proportion of all U.S. children ages 8 to 18 have Internet access at home.

Suppose we sampled 2,000 children and the sample proportion was still 86%. Our test statistic would be $Z \approx 2.44$, and our P-value would be about 0.015. The larger sample size would allow us to reject the null hypothesis even though the sample proportion was the same.

Why does this happen? Larger samples vary less, so a sample proportion of 0.86 is more unusual with larger samples than with smaller samples if the population proportion is really 0.84. This means that if the alternative hypothesis is true, a larger sample size will make it more likely that we reject the null. Therefore, we generally prefer a larger sample as we have seen previously.

## Drawing Conclusions from Hypothesis Tests

It is tempting to get involved in the details of a hypothesis test without thinking about how the data was collected. Whether we are calculating a confidence interval or performing a hypothesis test, the results are meaningless without a properly designed study. Consider the following exercises about how data collection can affect the results of a study.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3964

## Let's Summarize

In this section, we looked at the four steps of a hypothesis test as they relate to a claim about a population proportion.

**Step 1: Determine the hypotheses.**

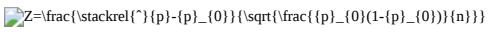
- The hypotheses are claims about the population proportion, $p$.
- The null hypothesis is a hypothesis that the proportion equals a specific value, $p_0$.
- The alternative hypothesis is the competing claim that the parameter is less than, greater than, or not equal to $p_0$.

**Step 2: Collect the data.**

Since the hypothesis test is based on probability, random selection or assignment is essential in data production. Additionally, we need to check whether the sample proportion can be $np \geq 10$ and $n(1 - p) \geq 10$.

**Step 3: Assess the evidence.**

- Determine the test statistic which is the $z$-score for the sample proportion. The formula is:

  
  
- Use the test statistic, together with the alternative hypothesis to determine the P-value. You can use a standard normal table (or $Z$-table) or technology (such as the simulations on the second page of this topic) to find the P-value.
- If the alternative hypothesis is greater than, the P-value is the area to the right of the test statistic. If the alternative hypothesis is less than, the P-value is the area to the left of the test statistic. If the alternative hypothesis is not equal to, the P-value is equal to double the tail area beyond the test statistic.

**Step 4: Give the conclusion.**

- A small P-value says the data is unlikely to occur if the null is true. If the P-value is less than or equal to the significance level, we reject the null hypothesis and accept the alternative hypothesis instead.
- If the P-value is greater than the significance level, we say we "fail to reject" the null hypothesis. We never say that we "accept" the null hypothesis. We just say that we don't have enough evidence to reject it. This is equivalent to saying we don't have enough evidence to support the alternative hypothesis.
- We write the conclusion in the context of the research question. Our conclusion is usually a statement about the alternative hypothesis (we accept $H_a$ or fail to accept $H_a$) and should include the P-value.

## Other Hypothesis Testing Notes

Remember that the P-value is the probability of seeing a sample proportion as extreme as the one observed from the data if the null hypothesis is true. The probability is about the random sample, not about the null or alternative hypothesis.

A larger sample size makes it more likely that we will reject the null hypothesis if the alternative is true. Another way of thinking about this is that increasing the sample size will decrease the likelihood of a type II error. Recall that a type II error is failing to reject the null hypothesis when the alternative is true.

Increasing the sample size can have the unintended effect of making the test sensitive to differences so small they don't matter. A statistically significant difference is one large enough that it is unlikely to be due to sampling variability alone. Even a difference so small that it is not important can be statistically significant if the sample size is big enough.

Finally, remember the phrase "garbage in, garbage out." If the data collection methods are poor, then the results of a hypothesis test are meaningless. No statistical methods can create useful information if our data comes from convenience or voluntary response samples. Additionally, the results of a hypothesis test apply only to the population from whom the sample was chosen.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 8.15: Hypothesis Test for a Population Proportion (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.16: Putting It Together- Inference for One Proportion

## Let's Summarize

In *Inference for One Proportion*, we learned two inference procedures to draw conclusions about a population proportion:

- A confidence interval when our goal is to estimate a population proportion.
- A hypothesis test when our goal is to test a claim about a population proportion.

## Confidence Interval for Estimating a Population Proportion

- A confidence interval estimates the population proportion with a range of possible values. The interval is based on a sample proportion and a margin of error.
- Every confidence interval has a confidence level associated with it. The confidence level is a probability statement. It tells us the chance that a confidence interval, with a specific margin of error, contains the population proportion. But we can never determine if a specific interval does or does not contain the population proportion. We also cannot determine the probability that the population proportion lies in a specific interval. We can only say that in the long run the confidence level describes the percentage of the confidence intervals that will estimate the population proportion within a specific margin of error.
- We can calculate a confidence interval for a population proportion when we can use a normal distribution to model the long-run behavior of sample proportions. We can use a normal distribution model when there are at least 10 observed successes and 10 observed failures.
- We calculate the confidence interval for a population proportion using this formula:

$$\begin{array}{l}\mathrm{Sample}\ \mathrm{proportion}\ \pm\ \mathrm{margin}\ \mathrm{of}\ \mathrm{error}\\ \hat{p}\ \pm\ Z_c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\end{array}$$

where $Z_c$ depends on the confidence level. The part of the formula after the ± is the margin of error. The most common confidence levels are 90%, 95%, and 99%. The critical *z*-scores are 1.65, 1.96, and 2.576.

- The margin of error comes from the standard error in the sampling distribution. Sample proportions from larger sample sizes have less variability, so the standard error is smaller. Therefore, confidence intervals based on larger sample sizes will have a smaller margin of error. This fits our intuition that larger samples will give more accurate estimates of the population proportion.
- A higher level of confidence makes us more confident that the interval contains the population proportion because the interval is wider. This also means that the margin of error is larger.

## Hypothesis Tests in General:

Hypothesis tests consist of four steps, which apply to all the hypothesis tests we will do in this course.

**Step 1: Determine the hypotheses.**

The hypotheses are statements about the parameter(s) in question. The null hypothesis, $H_0$, is always a statement of equality and usually means no change or difference. The alternative hypothesis, $H_a$, is always an inequality, either <, >, or ≠, and is based on the research question.

**Step 2: Collect the data.**

The data must come from a random sample that is representative of the population in question.

**Step 3: Assess the evidence.**

The P-value is the evidence. The P-value is the probability that we would get sample results at least as extreme as those observed if the null hypothesis is true. If the P-value is smaller than the significance level, the results are unusual enough for us to reject the null hypothesis. Otherwise, we "fail to reject" the null hypothesis.

**Step 4: Give the conclusion.**

Our conclusion is stated in terms of the alternative hypothesis. Either there is or there is not enough evidence to say that the alternative hypothesis is true. We always use the context of the problem in the conclusion and always include the P-value. Finally,

we never say that the null hypothesis is true, only that we reject or fail to reject it.

### Hypothesis Test for a Population Proportion:

For the four steps, the following are specific to hypothesis testing for a population proportion.

**Step 1: Determine the hypotheses.**

The hypotheses for a test about a population proportion are stated in terms of the $p$. Here $p_0$ is a number to which we compare the population proportion.

$H_0$: $p = p_0$
$H_a$: $p < p_0$ or $p > p_0$ or $p \neq p_0$

**Step 2: Collect the data.**

We also check at this point that $np \geq 10$ and $n(1 - p) \geq 10$, where $p$ is the value from the null hypothesis, $p_0$. If these conditions are true, a normal model is a good fit for the sampling distribution of sample proportions. We need this model to do the remaining steps in the hypothesis test.

**Step 3: Assess the evidence.**

We calculate the test statistic (the $z$-score) for our sample proportion. We use the test statistic to determine the P-value, using a standard normal curve. We can do this using a $Z$-table or technology. We used simulations or statistical software in our work. As always, if the P-value is smaller than the significance level, the results are unusual enough for us to reject the null hypothesis. Otherwise, we "fail to reject" the null hypothesis.

**Step 4: Give the conclusion.**

See the information about stating conclusions for the general hypothesis test. There is nothing to add to this when we test a hypothesis about a population proportion.

**Other important notes:**

- In a hypothesis test, we make a decision based on probability, so there is uncertainty. A type I error occurs when we reject the null hypothesis even though it is true. A type II error occurs when we fail to reject the null hypothesis even though the alternative hypothesis is true. These errors are due to chance: the data from a random sample has led us to a wrong conclusion without our knowledge, which can happen even if we do all the steps correctly.
- A difference may be statistically significant but not practically important for decision making. Examining the hypotheses and the sample results can help us realize when this happens.
- For both confidence intervals and hypothesis tests about a population proportion, we must make sure that our sample is representative of the population. Using bad data to calculate a confidence interval or conduct a hypothesis test will give us worthless results.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.16: Putting It Together- Inference for One Proportion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.17: StatTutor- Cell Phones

You are now ready to practice what you learned in this module by doing a StatTutor exercise. We design StatTutor exercises to help you apply what you have learned to a real life data analysis question.

**Instructions:** One of the first few screens in StatTutor has a link to download the dataset for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click "Save," which will allow you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you're using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you're using the TI Calculator.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=326

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.17: StatTutor- Cell Phones is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 8.18: Assignment- Hypothesis Testing for the Population Proportion p

The objectives of this activity are:

1. To give you guided practice in carrying out a hypothesis test about a population proportion. (Note: This hypothesis test is also called a z-test for the population proportion.)

2. To learn how to use statistical software to help you carry out the test.

**Background:** This activity is based on the results of a recent study on the safety of airplane drinking water that was conducted by the U.S. Environmental Protection Agency (EPA). A study found that out of a random sample of 316 airplanes tested, 40 had coliform bacteria in the drinking water drawn from restrooms and kitchens. As a benchmark comparison, in 2003 the EPA found that about 3.5% of the U.S. population have coliform bacteria-infected drinking water. The question of interest is whether, based on the results of this study, we can conclude that drinking water on airplanes is more contaminated than drinking water in general

### Question 1:

Let p be the proportion of contaminated drinking water in airplanes. Write down the appropriate null and alternative hypotheses.

### Question 2:

Based on the collected data, is it safe to use the z-test for p in this scenario? Explain.

Use the following instructions to conduct the z-test for the population proportion:

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 3:

Now that we have established that it is safe to use the Z-test for p for our problem, go ahead and carry out the test. Paste the output below.

### Question 4:

Note that, according to the output, the test statistic for this test is 8.86. Make sure you understand how this was calculated, and give an interpretation of its value.

### Question 5:

We calculated a P-value of 0 in this test. Interpret what that means, and draw your conclusions.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 8.18: Assignment- Hypothesis Testing for the Population Proportion p is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 9: Inference for Two Proportions


9.1: Why It Matters- Inference for Two Proportions
9.2: Introduction- Distribution of Differences in Sample Proportions
9.3: Distribution of Differences in Sample Proportions (1 of 5)
9.4: Distribution of Differences in Sample Proportions (2 of 5)
9.5: Distribution of Differences in Sample Proportions (3 of 5)
9.6: Distribution of Differences in Sample Proportions (4 of 5)
9.7: Distribution of Differences in Sample Proportions (5 of 5)
9.8: Introduction- Estimate the Difference Between Population Proportions
9.9: Estimate the Difference between Population Proportions (1 of 3)
9.10: Estimate the Difference between Population Proportions (2 of 3)
9.11: Estimate the Difference between Population Proportions (3 of 3)
9.12: Introduction- Hypothesis Test for Difference in Two Population Proportions
9.13: Hypothesis Test for Difference in Two Population Proportions (1 of 6)
9.14: Hypothesis Test for Difference in Two Population Proportions (2 of 6)
9.15: Hypothesis Test for Difference in Two Population Proportions (3 of 6)
9.16: Hypothesis Test for Difference in Two Population Proportions (4 of 6)
9.17: Hypothesis Test for Difference in Two Population Proportions (5 of 6)
9.18: Hypothesis Test for Difference in Two Population Proportions (6 of 6)
9.19: Putting It Together- Inference for Two Proportions
9.20: Assignment- A Statistical Investigation using Software



This page titled 9: Inference for Two Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.1: Why It Matters- Inference for Two Proportions

## Learning Objectives

- Recognize when to use a hypothesis test or a confidence interval to compare two population proportions or to investigate a treatment effect for a categorical variable.
- Determine if a study involving two proportions is an experiment or an observational study.

In previous modules, we learned to make inferences about a population proportion. In particular, we learned the following:

- Random samples vary. When we use a sample proportion to make an inference about a population proportion, there is uncertainty. For this reason, inference involves probability.
- Under certain conditions, we can model the variability in sample proportions with a normal curve. We use the normal curve to make probability-based decisions about population values.
- We can estimate a population proportion with a confidence interval. The confidence interval is an actual sample proportion with a margin of error. We state our confidence in the accuracy of these intervals using probability.
- We can test a hypothesis about a population proportion using an actual sample proportion. Again, we base our conclusion on probability using a P-value. The P-value describes the strength of our evidence in rejecting a hypothesis about the population.

In *Inference for Two Proportions*, we continue to work with categorical data, so we continue to use proportions. *But now we make inferences that compare two populations (or two treatments).*

As an overview, consider again the Big Picture of Statistics.

Here we discuss the four steps in a statistical investigation for situations from Module 9.

1. **Produce Data:***Determine what to measure, then collect the data*. In this module, we collect categorical data from two samples. In an observational study, we begin with two populations and randomly select a sample from each population. In an experiment, we randomly assign individuals to two treatments. The use of random selection or random assignment allows us to view the samples as independent. This means we assume that the variable values from one sample do not influence the values for the other sample.
2. **Exploratory Data Analysis:***Analyze and summarize the data*. We are working with categorical data, so from each sample, we compute a sample proportion. To compare the two samples, we subtract the proportions. When we conduct inference in the next step, our goal is to to determine if the actual difference in the sample proportions is significantly different from what we expect in random sampling.
3. **Draw a Conclusion:***Use data, probability, and statistical inference to draw a conclusion about the populations*.Our approach to inference repeats the reasoning we did in *Inference for One Proportion*.
    - We use simulation to observe the behavior of *the differences in sample proportions* when we randomly select many, many samples. We create the simulation to reflect a claim about the populations. Then we develop a probability model to describe the shape, center, and spread of the sampling distribution. Of course, we are interested in the conditions that allow us to use a normal curve.
    - We use this model to determine when a given difference is unusual in a formal hypothesis test.
    - We also construct confidence intervals to estimate the difference between two *population* proportions. As before, we make a probability statement about our confidence in the accuracy of these intervals.

## Example

## The Abecedarian Early Intervention Project

In the 1970s, Abecedarian Early Intervention Project studied the long-term effects of early childhood education for poor children.

**Research question:** *Does early childhood education increase the likelihood of college attendance for poor children?*

1. **Produce Data:***Determine what to measure, then collect the data.*In this experiment, researchers selected 111 high-risk infants on the basis of the mothers' education, family income, and other factors. They randomly assigned 57 infants to receive 5 years of high-quality preschool. The remaining 54 infants were a control group. All children received nutritional supplements, social services, and health care to control the effects of these confounding factors on the outcomes of the experiment.

2. **Exploratory Data Analysis:***Analyze and summarize the data*.By the age of 21 a much higher percentage of the treatment group enrolled in college, 42% vs. 20%.
3. **Draw a Conclusion:***Use data, probability, and statistical inference to draw a conclusion about the populations*.Is this difference statistically significant? In other words, is this difference due to the pre-school experience or due to chance? We will test the claim that a larger proportion of children who attend pre-school will attend college.

The following figure summarizes this investigation in the Big Picture.

## Learn By Doing

**Health Care for Non-Union and Union Workers**

In a recent study the AFL/CIO selected random samples of non-union and union employees. They compared the proportion of workers in each sample who had health insurance. They found that the proportion of non-union workers with health insurance was significantly lower than the proportion of union workers with health insurance.

https://assessments.lumenlearning.co...sessments/3620

https://assessments.lumenlearning.co...sessments/3621

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.1: Why It Matters- Inference for Two Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.2: Introduction- Distribution of Differences in Sample Proportions

## What you'll learn to do: Recognize when to use a two population proportion hypothesis test to compare two populations/treatment groups.

### LEARNING OBJECTIVES

- Recognize when to use a hypothesis test or a confidence interval to compare two population proportions or to investigate a treatment effect for a categorical variable.
- Determine if a study involving two proportions is an experiment or an observational study.
- Describe the sampling distribution of the difference between two proportions.
- Draw conclusions about a difference in population proportions from a simulation.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.2: Introduction- Distribution of Differences in Sample Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.3: Distribution of Differences in Sample Proportions (1 of 5)

## Learning Objectives

- Describe the sampling distribution of the difference between two proportions.
- Draw conclusions about a difference in population proportions from a simulation.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.3: Distribution of Differences in Sample Proportions (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.4: Distribution of Differences in Sample Proportions (2 of 5)

## Learning Objectives

- Draw conclusions about a difference in population proportions from a simulation.

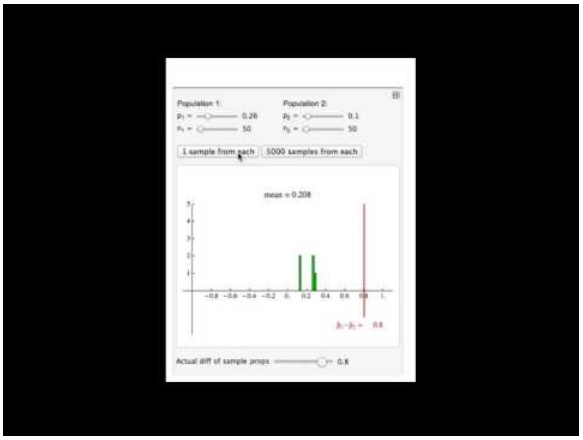
**Introduction**

Recall that we are in the middle of an investigation about the difference in female and male teen depression rates. In our investigation, we are assuming that 26% of female teens and 10% of male teens are depressed. That is, we assume a 16% = 0.16 difference favoring girls.

- *We saw a 0.06 gender difference in teen depression rates from the National Survey of Adolescents. Again, girls had a higher rate of depression. Does this study suggest that our assumption about a 0.16 difference in the populations is wrong?*
- *Or could the results have come from populations with a 0.16 difference in depression rates?*

At this point, we may have a sense of the answers to these questions for samples of 64 females and 100 males. But we need to look at the long-run behavior of the differences in sample proportions. We also need to investigate the effect of sample size on our conclusion. The samples in the National Survey of Adolescents are very large.

So we continue this investigation in a Simulation WalkThrough.




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=338

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3920

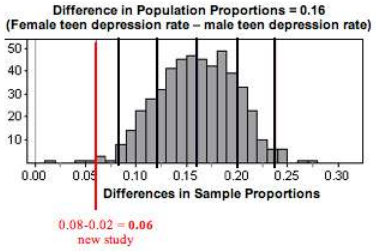
https://assessments.lumenlearning.co...sessments/3921

On the next page, we use the simulation shown in the WalkThrough to make inferences about a difference in population proportions. As we did in *Linking Probability to Statistical Inference*, we use a simulation to make observations about the sampling distribution before we develop the mathematical model that we will use in inference. The logic we use to make inferences with simulated sampling distributions is the same logic we use with mathematical models. Let's practice that way of thinking now.

## Learn By Doing

Suppose in a study of 540 female and 475 male U.S. teens, we find that 8% of the females and 2% of the males are depressed. What does this study suggest about our assumption that the depression rate of female teens is 16% higher than that of male teens in the United States?

Here is a simulated distribution of differences for a large number of independent random samples for these sample sizes. Note that we have rescaled the axis, so the distribution may look wider than the distributions in the WalkThrough, but it actually has less variability.




https://assessments.lumenlearning.co...sessments/3632

https://assessments.lumenlearning.co...sessments/3633

https://assessments.lumenlearning.co...sessments/3634

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.4: Distribution of Differences in Sample Proportions (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.5: Distribution of Differences in Sample Proportions (3 of 5)

## Learning Objectives

- Draw conclusions about a difference in population proportions from a simulation.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.5: Distribution of Differences in Sample Proportions (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.6: Distribution of Differences in Sample Proportions (4 of 5)

## Learning Objectives

- Draw conclusions about a difference in population proportions from a simulation.

## The Sampling Distribution of Differences in Sample Proportions

Let's summarize what we have observed about the sampling distribution of the differences in sample proportions. We want to create a mathematical model of the sampling distribution, so we need to understand when we can use a normal curve. We also need to understand how the center and spread of the sampling distribution relates to the population proportions.

**Shape:**

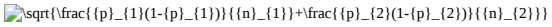
In each situation we have encountered so far, the distribution of differences between sample proportions appears somewhat normal, but that is not always true. We discuss conditions for use of a normal model later.

**Center:**

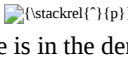
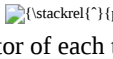
Regardless of shape, the mean of the distribution of sample differences is the difference between the population proportions, $p_1 - p_2$. This is always true if we look at the long-run behavior of the differences in sample proportions.

**Spread:**

We have observed that larger samples have less variability. Advanced theory gives us this formula for the standard error in the distribution of differences between sample proportions:



Notice the following:

- The terms under the square root are familiar. These terms are used to compute the standard errors for the individual sampling distributions of  and  .
- The sample size is in the denominator of each term. As we learned earlier this means that increases in sample size result in a smaller standard error.

## Comment

Let's look at the relationship between the sampling distribution of differences between sample proportions and the sampling distributions for the individual sample proportions we studied in *Linking Probability to Statistical Inference*. We compare these distributions in the following table.

| Sampling Distribution | Sample Proportions from Population 1 | Sample Proportions from Population 2 | All Differences in Sample Proportions from the two Populations |
|---|---|---|---|
| Mean | $p_1$ | $p_2$ | $p_1 - p_2$ |
| Standard Error | $\sqrt{\frac{p_1(1-p_1)}{n_1}}$ | $\sqrt{\frac{p_2(1-p_2)}{n_2}}$ | $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ |

Notice the relationship between the means:

- The mean of the differences is the difference of the means. This makes sense. The mean of each sampling distribution of individual proportions is the population proportion, so the mean of the sampling distribution of differences is the difference in population proportions.

Notice the relationship between standard errors:

- The standard error of differences relates to the standard errors of the sampling distributions for individual proportions. Look at the terms under the square roots. Since we add these terms, the standard error of differences is always larger than the standard error in the sampling distributions of individual proportions. In other words, there is more variability in the differences.

## Variability and Variance

In this module, we sample from two populations of categorical data, and compute sample proportions from each.

We have seen that the means of the sampling distributions of sample proportions are $\text{p}_{1}\text{}\mathrm{and}\text{}\text{p}_{2}$ and the standard errors are $\sqrt{\frac{\text{p}_{1}(1-\text{p}_{1})}{\text{n}_{1}}}\text{}\mathrm{and}\text{}\sqrt{\frac{\text{p}_{2}(1-\text{p}_{2})}{\text{n}_{2}}}$ .

Statisticians often refer to the square of a standard deviation or standard error as a *variance*. The variances of the sampling distributions of sample proportion are

$$\frac{\text{p}_{1}(1-\text{p}_{1})}{\text{n}_{1}}\text{}\mathrm{and}\text{}\frac{\text{p}_{2}(1-\text{p}_{2})}{\text{n}_{2}}$$

If we add these variances we get the variance of the differences between sample proportions.

$$\frac{\text{p}_{1}(1-\text{p}_{1})}{\text{n}_{1}}\text{+}\frac{\text{p}_{2}(1-\text{p}_{2})}{\text{n}_{2}}$$

For the sampling distribution of all differences, $\stackrel{\text{^}}{\text{p}}_{1}-\stackrel{\text{^}}{\text{p}}_{2}$ the mean, $\mathrm{\mu}\stackrel{\text{^}}{\text{p}}_{1}-\stackrel{\text{^}}{\text{p}}_{2}$ , of all differences is the difference of the means $\text{p}_{1}-\text{p}_{2}$ . The variance of all differences, $\mathrm{\sigma}^{2}\stackrel{\text{^}}{\text{p}}_{1}-\stackrel{\text{^}}{\text{p}}_{2}$ , is the sum of the variances, $\frac{\text{p}_{1}(1-\text{p}_{1})}{\text{n}_{1}}\text{+}\frac{\text{p}_{2}(1-\text{p}_{2})}{\text{n}_{2}}$ .

We will now do some problems similar to problems we did earlier. Only now, we do not use a simulation to make observations about the variability in the differences of sample proportions. Instead, we use the mean and standard error of the sampling distribution. But our reasoning is the same.

### Example

## Controversy about HPV Vaccine

During a debate between Republican presidential candidates in 2011, Michele Bachmann, one of the candidates, implied that the vaccine for HPV is unsafe for children and can cause mental retardation. Scientists and other healthcare professionals immediately produced evidence to refute this claim. A *USA Today* article, “No Evidence HPV Vaccines Are Dangerous” (September 19, 2011), described two studies by the Centers for Disease Control and Prevention (CDC) that track the safety of the vaccine. Here is an excerpt from the article:

> *First, the CDC monitors reports to the Vaccine Adverse Event Reporting System, a database to which anyone can report a suspected side effect. CDC officials then investigate to see whether reported problems could possibly be caused by vaccines or are simply a coincidence. Second, the CDC has been following girls who receive the vaccine over time, comparing them with a control group of unvaccinated girls….Again, the HPV vaccine has been found to be safe.*

According to an article by Elizabeth Rosenthal, “Drug Makers’ Push Leads to Cancer Vaccines’ Rise” (*New York Times*, August 19, 2008), the FDA and CDC said that “with millions of vaccinations, by chance alone some serious adverse effects and deaths will occur in the time period following vaccination, but have nothing to do with the vaccine.” The article stated that the FDA and CDC monitor data to determine if more serious effects occur than would be expected from chance alone.

According to another source, the CDC data suggests that serious health problems after vaccination occur at a rate of about 3 in 100,000. This is a proportion of 0.00003. But are these health problems due to the vaccine? Is the rate of similar health problems any different for those who don’t receive the vaccine? Let’s assume that there are no differences in the rate of serious health problems between the treatment and control groups. That is, let’s assume that the proportion of serious health problems in both groups is 0.00003.

Suppose the CDC follows a random sample of 100,000 girls who had the vaccine and a random sample of 200,000 girls who did not have the vaccine. Over time, they calculate the proportion in each group who have serious health problems.

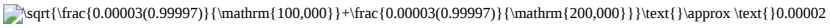
**Question:** *How much of a difference in these sample proportions is unusual if the vaccine has no effect on the occurrence of serious health problems?*

To answer this question, we need to see how much variation we can expect in random samples if there is no difference in the rate that serious health problems occur, so we use the sampling distribution of differences in sample proportions.

- Center: Mean of the differences in sample proportions is

$$\text{p}_{1}-\text{p}_{2}=0.00003-0.00003=0$$

- Spread: The large samples will produce a standard error that is very small. The standard error of the differences in sample proportions is



**Answer:** We can view random samples that vary more than 2 standard errors from the mean as unusual. If there is no difference in the rate that serious health problems occur, the mean is 0. So differences in rates larger than 0 + 2(0.00002) = 0.00004 are unusual. This is equivalent to about 4 more cases of serious health problems in 100,000. With such large samples, we see that a small number of additional cases of serious health problems in the vaccine group will appear unusual. But are 4 cases in 100,000 of practical significance given the potential benefits of the vaccine? This is an important question for the CDC to address.

### Learn By Doing

According to a 2008 study published by the AFL-CIO, 78% of union workers had jobs with employer health coverage compared to 51% of nonunion workers. In 2009, the Employee Benefit Research Institute cited data from large samples that suggested that 80% of union workers had health coverage compared to 56% of nonunion workers. Let's suppose the 2009 data came from random samples of 3,000 union workers and 5,000 nonunion workers.

https://assessments.lumenlearning.co...sessments/3924

https://assessments.lumenlearning.co...sessments/3636

### Learn By Doing

The following is an excerpt from a press release on the AFL-CIO website published in October of 2003.

> *Wal-Mart exemplifies the harmful trend among America's large employers to shirk health insurance responsibilities at the cost of their workers and community…. With reduced coverage and increased workers' premium fees, Wal-Mart – the largest private employer in the U.S. – sets a troubling standard. Fewer than half of Wal-Mart workers are insured under the company plan – just 46 percent. This rate is dramatically lower than the 66 percent of workers at large private firms who are insured under their companies' plans, according to a new Commonwealth Fund study released today which documents the growing trend among large employers to drop health insurance for their workers.*

Suppose we want to see if this difference reflects insurance coverage for workers in our community. We select a random sample of 50 Wal-Mart employees and 50 employees from other large private firms in our community. Suppose that 20 of the Wal-Mart employees and 35 of the other employees have insurance through their employer.

https://assessments.lumenlearning.co...sessments/3925

https://assessments.lumenlearning.co...sessments/3637

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.6: Distribution of Differences in Sample Proportions (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.7: Distribution of Differences in Sample Proportions (5 of 5)

## Learning Objectives

- Estimate the probability of an event using a normal model of the sampling distribution.

## Why Do We Care about a Normal Model?

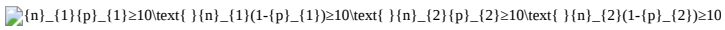
Now we focus on the conditions for use of a normal model for the sampling distribution of differences in sample proportions.

We use a normal model for inference because we want to make probability statements without running a simulation. If we are conducting a hypothesis test, we need a P-value. If we are estimating a parameter with a confidence interval, we want to state a level of confidence. These procedures require that conditions for normality are met.

**Note:** If the normal model is not a good fit for the sampling distribution, we can still reason from the standard error to identify unusual values. We did this previously. For example, we said that it is unusual to see a difference of more than 4 cases of serious health problems in 100,000 if a vaccine does not affect how frequently these health problems occur. But without a normal model, we can't say *how* unusual it is or state the probability of this difference occurring.

## When Is a Normal Model a Good Fit for the Sampling Distribution of Differences in Proportions?

A normal model is a good fit for the sampling distribution of differences if a normal model is a good fit for both of the individual sampling distributions. More specifically, we use a normal model for the sampling distribution of differences in proportions if the following conditions are met.



These conditions translate into the following statement:

*The number of expected successes and failures in both samples must be at least 10.* (Recall here that *success* doesn't mean good and *failure* doesn't mean bad. A success is just what we are counting.)

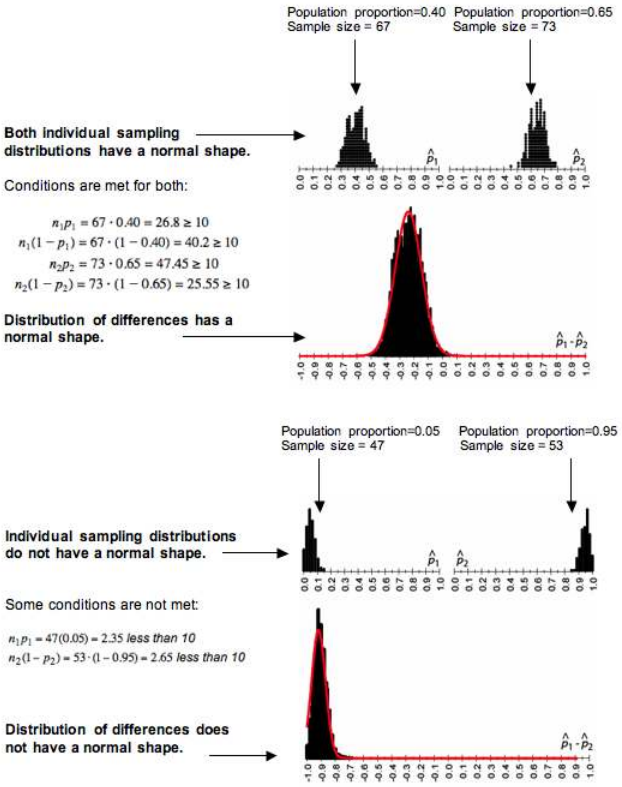
Here we complete the table to compare the *individual* sampling distributions for sample proportions to the sampling distribution of differences in sample proportions.

| Sampling Distribution | Sample Proportions from Population 1 | Sample Proportions from Population 2 | All Differences in Sample Proportions from the two Populations |
|---|---|---|---|
| Mean | $p_1$ | $p_2$ | $p_1 - p_2$ |
| Standard Error | $\sqrt{\frac{p_1(1-p_1)}{n_1}}$ | $\sqrt{\frac{p_2(1-p_2)}{n_2}}$ | $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ |
| Conditions for Use of a Normal Model | $n_1p_1 \geq 10$ *and* $n_1(1-p_1) \geq 10$ In sample from Population 1, expected successes and failures at least 10 | $n_2p_2 \geq 10$ *and* $n_2(1-p_2) \geq 10$ In sample from Population 2, expected successes and failures at least 10 | $n_1p_1 \geq 10$ *and* $n_1(1-p_1) \geq 10$ $n_2p_2 \geq 10$ *and* $n_2(1-p_2) \geq 10$ expected successes and failures in BOTH samples at least 10 |

## Example

## More on Conditions for Use of a Normal Model

All of the conditions must be met before we use a normal model. If one or more conditions is not met, do not use a normal model. Here we illustrate how the shape of the individual sampling distributions is inherited by the sampling distribution of differences.




## Learn By Doing

Recall the AFL-CIO press release from a previous activity. “Fewer than half of Wal-Mart workers are insured under the company plan – just 46 percent. This rate is dramatically lower than the 66 percent of workers at large private firms who are insured under their companies’ plans, according to a new Commonwealth Fund study released today, which documents the growing trend among large employers to drop health insurance for their workers.”

https://assessments.lumenlearning.co...sessments/3628

https://assessments.lumenlearning.co...sessments/3629

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3926

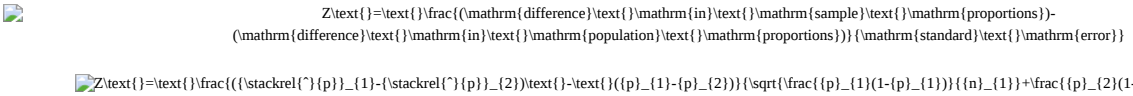
## Using the Normal Model in Inference

When conditions allow the use of a normal model, we use the normal distribution to determine P-values when testing claims and to construct confidence intervals for a difference between two population proportions.

We can standardize the difference between sample proportions using a *z*-score. We calculate a *z*-score as we have done before.




For a difference in sample proportions, the *z*-score formula is shown below.







## Example

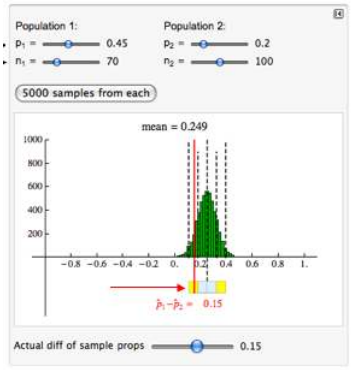
## Abecedarian Early Intervention Project

Recall the Abecedarian Early Intervention Project. For this example, we assume that 45% of infants with a treatment similar to the Abecedarian project will enroll in college compared to 20% in the control group. That is, we assume that a high-quality prechool experience will produce a 25% increase in college enrollment. We call this the *treatment effect*.

Let's suppose a daycare center replicates the Abecedarian project with 70 infants in the treatment group and 100 in the control group. After 21 years, the daycare center finds a 15% increase in college enrollment for the treatment group. This is still an impressive difference, but it is 10% less than the effect they had hoped to see.

*What can the daycare center conclude about the assumption that the Abecedarian treatment produces a 25% increase?*

Previously, we answered this question using a simulation.




This difference in sample proportions of 0.15 is less than 2 standard errors from the mean. This result is not surprising if the treatment effect is really 25%. We cannot conclude that the Abecedarian treatment produces less than a 25% treatment effect.

Now we ask a different question: *What is the probability that a daycare center with these sample sizes sees less than a 15% treatment effect with the Abecedarian treatment?*

We use a normal model to estimate this probability. The simulation shows that a normal model is appropriate. We can verify it by checking the conditions. All expected counts of successes and failures are greater than 10.

For the treatment group: $\begin{array}{l}70(0.45)=31.5\\ 70(0.55)=38.5\end{array}$

For the control group: $\begin{array}{l}100(0.20)=20\\ 100(0.80)=80\end{array}$

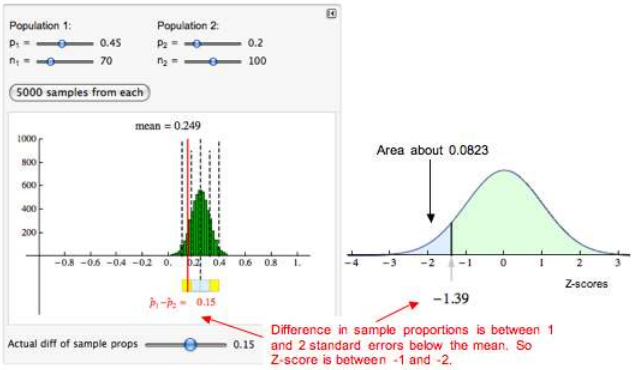
In the simulated sampling distribution, we can see that the difference in sample proportions is between 1 and 2 standard errors below the mean. So the $z$-score is between −1 and −2. When we calculate the $z$-score, we get approximately −1.39.

$$Z=\frac{(\mathrm{difference}\text{ }\mathrm{in}\text{ }\mathrm{sample}\text{ }\mathrm{proportions})\text{ }-\text{ }(\mathrm{difference}\text{ }\mathrm{in}\text{ }\mathrm{population}\text{ }\mathrm{proportions})}{\mathrm{standard}\text{ }\mathrm{error}}$$

$$\mathrm{standard}\text{ }\mathrm{error}\text{ }=\text{ }\sqrt{\frac{0.45(0.55)}{70}+\frac{0.20(0.80)}{100}}\text{ }\approx \text{ }0.072$$

$$Z\text{ }=\text{ }\frac{0.15-0.25}{0.072}\text{ }\approx \text{ }-1.39$$

We use a simulation of the standard normal curve to find the probability. We get about 0.0823.




**Conclusion:** If there is a 25% treatment effect with the Abecedarian treatment, then about 8% of the time we will see a treatment effect of less than 15%. This probability is based on random samples of 70 in the treatment group and 100 in the control group.

## Learn By Doing

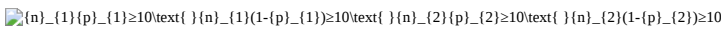
https://assessments.lumenlearning.co...sessments/3965

## Let's Summarize

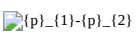
In "Distributions of Differences in Sample Proportions," we compared two population proportions by subtracting. When we select independent random samples from the two populations, the sampling distribution of the difference between two sample proportions has the following shape, center, and spread.

***Shape:***

A normal model is a good fit for the sampling distribution if the number of expected successes and failures in each sample are all at least 10. Written as formulas, the conditions are as follows.



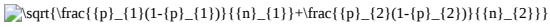

***Center:***

Regardless of shape, the mean of the distribution of sample differences is the difference between the population proportions, 

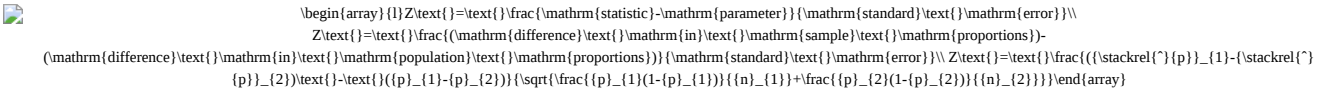
. This is always true if we look at the long-run behavior of the differences in sample proportions.

***Spread:***

As we know, larger samples have less variability. The formula for the standard error is related to the formula for standard errors of the individual sampling distributions that we studied in *Linking Probability to Statistical Inference*.




If a normal model is a good fit, we can calculate *z*-scores and find probabilities as we did in Modules 6, 7, and 8. The formula for the *z*-score is similar to the formulas for *z*-scores we learned previously.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.7: Distribution of Differences in Sample Proportions (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.8: Introduction- Estimate the Difference Between Population Proportions

## What you'll learn to do: Construct and interpret confidence intervals to compare two population/treatment group proportions.

### LEARNING OBJECTIVES

- Construct a confidence interval to estimate the difference between two population proportions (or the size of a treatment effect) when conditions are met. Interpret the confidence interval in context.
- Interpret the meaning of a confidence level associated with a confidence interval and describe how the confidence level affects the margin of error.
- Given the description of a statistical study, evaluate whether conclusions are reasonable.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.8: Introduction- Estimate the Difference Between Population Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.9: Estimate the Difference between Population Proportions (1 of 3)

## Learning Objectives

- Recognize when to use a hypothesis test or a confidence interval to compare two population proportions or to investigate a treatment effect for a categorical variable.
- Construct a confidence interval to estimate the difference between two population proportions (or the size of a treatment effect) when conditions are met. Interpret the confidence interval in context.

In "Distributions of Differences in Sample Proportions," we used simulation to observe the behavior of *the differences in sample proportions* when we randomly select many, many samples. From the simulation, we developed a normal probability model to describe the sampling distribution of sample differences. With this model, we are now ready to do inference about a difference in population proportions (or about a treatment effect.)

When our goal is to estimate a difference between two population proportions (or the size of a treatment effect), we select two independent random samples and use the difference in sample proportions as an estimate. Of course, random samples vary, so we want to include a statement about the amount of error that may be present. Because the differences in sample proportions vary in a predictable way, we can also make a probability statement about how confident we are in the process that we used to estimate the difference between the population proportions. You may recognize that what we are describing is a confidence interval.

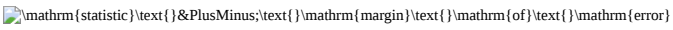
In *Inference for One Proportion*, we calculated confidence intervals to estimate a single population proportion. In this section, "Estimate the Difference between Population Proportions," we learn to calculate a confidence interval to estimate the difference between two population proportions. If the data comes from an experiment, we estimate the size of the treatment effect.

## Learn By Doing

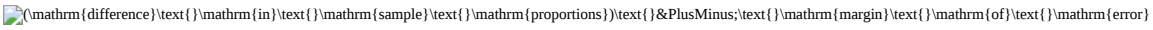
https://assessments.lumenlearning.co...sessments/3927

## Confidence Interval for a Difference in Two Population Proportions: the Basics

Every confidence interval has this form:

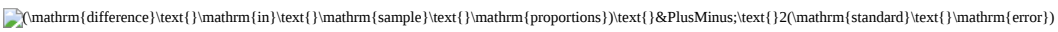


To estimate a difference in population proportions (or a treatment effect), the statistic is a difference in sample proportions, so the confidence interval is



When we select two random samples and calculate the difference in the sample proportions, we do not know the exact amount of error for this particular pair of samples. We therefore use the standard error as a typical amount of error and calculate the margin of error from the standard error, as we did in *Inference for One Proportion*.

If a normal model is a good fit for the sampling distribution, we can use it to make probability statements that describe our confidence in the interval. More specifically, we use the normal model to describe our confidence that the difference in population proportions lies within a given margin of error of the difference in sample proportions. For example, we can state that we are 95% confident that the difference in population proportions is contained in the following interval:



## Learn By Doing

## Nuclear Power

The following problem is based on a report, "Opposition to Nuclear Power Rises amid Japanese Crisis," by the Pew Research Center (March 21, 2011).

After the nuclear reactor accidents in Japan during the spring of 2011, there was a shift in public support for expanded use of nuclear power in the United States. A few months *before* the accident, 47% of a random sample of 1,004 U.S. adults supported expanded use of nuclear power. *After* the nuclear accident in Japan, 39% of a different random sample of 1,004 U.S. adults favored expanded use.

https://assessments.lumenlearning.co...sessments/3928

https://assessments.lumenlearning.co...sessments/3929

### What Does *95% Confident* Really Mean?

95% confident comes from a normal model of the sampling distribution.

To review this idea in the context of differences in sample proportions, let's start with a picture to help us visualize a confidence interval and its relationship to the sampling distribution. The following normal model represents the sampling distribution.

In the diagram, notice the sample difference. In the sampling distribution, we can see that the error in this sample difference is less than the margin of error. We know this because the distance between the sample difference and the population difference is shorter than the length of the margin of error (abbreviated MOE in the figure). When we create a confidence interval with this sample difference, we mark a distance equal to a margin of error on either side of the sample difference. Notice that this interval contains the population difference, which makes sense because the distance between the population difference and the sample difference has not changed.

*So where does the "95%" come from?* If the normal model is a good fit for the sampling distribution, then the empirical rule applies. The empirical rule says that 95% of the values in a normal distribution fall within 2 standard deviations of the mean. So 95% of the sample differences are within 2 standard errors of the mean difference. Remember that the mean difference is the difference in population proportions. Now consider the confidence interval centered at a sample difference. From the empirical rule, it follows that 95% of the confidence intervals, with a margin of error equal to 2 standard errors, will contain the population difference.

Following is another illustration of 95% confidence, a concept that is often misinterpreted. This diagram helps us remember the correct interpretation. If we construct confidence intervals with a margin of error equal to 2 standard errors, then 95% confidence means that in the long run, 95% of these confidence intervals will contain the population difference, and 5% of the time, the interval we calculate will not contain it. We show one of these less common intervals with a red dot at the sample difference.

Of course, in reality, we don't know the difference in population proportions. (This is the why we want to estimate it with a confidence interval!) So, in reality, we will not be able to determine if a specific confidence interval does or does not contain the true difference in population proportions. This is why we state a level of confidence. For a specific interval, we say we are 95% confident that the interval contains the true difference in population proportions.

### Example

## Correct and Incorrect Interpretations of 95% Confidence

Recall our earlier example about the change in public opinion after the nuclear accident in Japan. We concluded, "We are 95% confident that there was a 4% to 12% drop in support for the expanded use of nuclear power in the United States after the nuclear accident in Japan."

Here are some accurate ways to describe the phrase "95% confident" for this confidence interval:

- There is a 95% chance that poll results from two random samples will give a confidence interval that contains the true change in public support for the expanded use of nuclear power in the United States after the nuclear accident in Japan
- 95% of the time, this method produces an interval that covers the true difference in the proportions of the U.S. adult population supporting expanded use of nuclear power in the United States before and after the nuclear accident in Japan.

Here are some incorrect interpretations:

- There is a 95% chance that the true difference in public opinion (before and after the nuclear accident in Japan) is between 4% and 12%.
- There is a 95% chance that there was a 4% to 12% drop in support for the expanded use of nuclear power in the United States after the nuclear accident in Japan.

What do you notice? In the correct interpretations, the 95% is a probability statement about the random event of sampling. So statements like "95% chance that two random samples give a confidence interval" and "95% of the time, this method produces an

interval" describe the chance that confidence intervals, in the long run, contain the difference in population proportions.

In the wrong interpretations, the phrase "95% chance" is a probability statement about the specific interval 4% to 12%. Since we already found the confidence interval, there is no random event in this description, so we cannot make a probability statement about a single specific interval. For this reason, we use the phrase "95% confident" when we are describing a single interval from a specific study.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3930

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.9: Estimate the Difference between Population Proportions (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

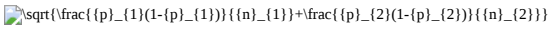
# 9.10: Estimate the Difference between Population Proportions (2 of 3)

## Learning Objectives

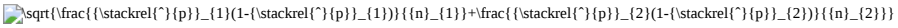
- Construct a confidence interval to estimate the difference between two population proportions (or the size of a treatment effect) when conditions are met. Interpret the confidence interval in context.
- Interpret the meaning of a confidence level associated with a confidence interval and describe how the confidence level affects the margin of error.
- Given the description of a statistical study, evaluate whether conclusions are reasonable.

## Confidence Interval for a Difference in Two Population Proportions: Beyond the Basics

For all confidence intervals, the margin of error is based on the standard error. We know from "Distributions of Differences in Sample Proportions" that the standard error for the sampling distribution of differences in sample proportions is:

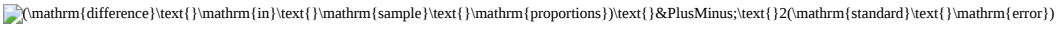


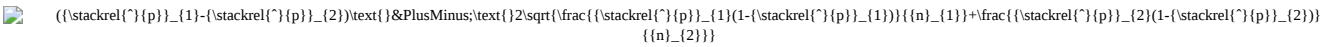
Obviously, if we are trying to estimate the difference in population proportions, we will not know or . So we estimate these population proportions with our sample proportions. This is the same approach we used when had to estimate the standard error for the distribution of sample proportions in *Inference for One Proportion*. The estimated standard error becomes



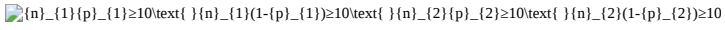
This formula estimates the average error between a difference in sample proportions and the true difference in population proportions.

So a 95% confidence interval has the following formula:





We can use this formula only if a normal model is a good fit for the sampling distribution. Recall that this is true only if the expected number of successes and failures in each sample is at least 10. For those who like formulas, these conditions translate into the following inequalities.



We have to adjust these conditions because we do not know the population proportions and . We make the same adjustment we made in *Inference for One Proportion*. We require that the *actual* number of successes and failures in each sample is at least 10. For those who like formulas, these conditions translate into replacing and with the corresponding sample proportions. Luckily, this tweak works and the normal distribution still gives fairly accurate confidence levels for different critical *z*-scores.

## Learn By Doing

## Nicotine Replacement Therapy

The Centre for Addiction and Mental Health in Canada posted the following description of a clinical trial on clinicaltrials.gov in September 2011.

> *This study will examine the efficacy of mailed distribution of free Nicotine Replacement Therapy to smokers. Telephone numbers will be randomly selected from across Canada in order to recruit adult smokers interested in completing a smoking survey and willing to be interviewed again in 8 weeks and 6 months times. Study participants will be asked about their smoking history and a hypothetical question: would they be interested in receiving the nicotine patch if this were to be provided to them free of charge? Participants expressing interest will be randomly assigned to one of two groups. One group will be offered the opportunity to actually receive a program of 5 weeks of nicotine patch for free right away and the other group will not be offered the free nicotine patches. The proportions of smokers in the two groups who quit smoking by the 6-month interview will be compared.*

https://assessments.lumenlearning.co...sessments/3640

https://assessments.lumenlearning.co...sessments/3931

https://assessments.lumenlearning.co...sessments/3932

https://assessments.lumenlearning.co...sessments/3933

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3639

### Other Levels of Confidence

In *Inference for One Proportion*, we saw that we can create confidence intervals for other levels of confidence. Changing the level of confidence changes the critical *z*-score. The following image shows the three most commonly used confidence levels and their critical *z*-scores.

The following table summarizes the critical values for the most commonly used confidence levels.

Note: A more exact value for the margin of error of a 95% confidence interval uses $Z_c = 1.96$ instead of 2 standard errors.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3934

### Learn By Doing

## What Is the Effect of Increasing the Confidence Level on the Margin of Error?

In an article titled "The Patriotism God Gap: Is the U.S. the Greatest Country in the World?" (*Christianity Today*, August 5, 2011), Tobin Grant cites data from the Pew Research Center. Here is an excerpt from the article:

> *About 40 percent of other Christians [non-evangelicals] said the U.S. stands alone as the greatest country. Those with no religion stand out as being much less likely to see the U.S. as the greatest country. Only 20 percent said the U.S. was the best country in the world.*

The article does not give the sample sizes for these two groups. For this activity, let's suppose the data describes random samples of 500 from the populations of "other Christians" and those with "no religion." With samples this large, we can safely model the sampling distribution of sample differences with a normal curve.

Use the simulation to find the margin of error for the 90%, 95%, and 99% confidence intervals. (Note: Conditions for use of the confidence interval formula are met because the sample size is large.)

https://assessments.lumenlearning.co...sessments/3935

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.10: Estimate the Difference between Population Proportions (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.11: Estimate the Difference between Population Proportions (3 of 3)

## Learning Objectives

- Construct a confidence interval to estimate the difference between two population proportions (or the size of a treatment effect) when conditions are met. Interpret the confidence interval in context.
- Given the description of a statistical study, evaluate whether conclusions are reasonable.

## Drawing Conclusions from Confidence Intervals

It is tempting to get involved in the details of calculating and interpreting a confidence interval without thinking about how the data was collected. Whether we are calculating a confidence interval or performing a hypothesis test, the results are meaningless without a properly designed study.

Here is a quick review of what we already know about the connection between study design, use of inference procedures, and valid conclusions.

- The goal of statistical inference is to use sample statistics to estimate population parameters. Therefore, the data must be a representative sample of the population of interest. This also applies to inference that compares two population parameters.
- In general, we can use statistical inference procedures if the data come from randomly selected or randomly assigned individuals.
- Cause-and-effect conclusions are possible when we randomly assign individuals to treatment groups in a well-designed experiment.
- Since inference procedures are based on probability models, the data must also meet the specific conditions for the procedure we have chosen.

In the next activities, we apply these ideas to the use of confidence intervals for estimating a difference between two population proportions (or estimating a treatment effect.)

## Learn By Doing

## Does Involving a Statistician Improve the Chance That a Medical Research Paper Will Be Published?

The following excerpt from "How Statistical Expertise Is Used in Medical Research" *(Altman, D. G., S. N. Goodman, and S. Schroter, Journal of the American Medical Association 287(21):2817–20, 2002)* describes the data collection method for this study.

> *Authors of original research articles who submitted to* BMJ [British Medical Journal] *and* Annals of Internal Medicine *from May through August 2001 were sent a short questionnaire.…Authors were asked if they received assistance from a person with statistical expertise.*

Of the 190 who did not work with a statistician, 134 had papers rejected without peer review. Of the 514 who did work with a statistician, 293 had papers rejected without peer review.

https://assessments.lumenlearning.co...sessments/3936

https://assessments.lumenlearning.co...sessments/3937

https://assessments.lumenlearning.co...sessments/3938

## Comment

Even when inference is inappropriate, exploratory data analysis can give us important information. The authors of the previous study list two other reasons that their data "make inference difficult." But they end their paper with the following statement. "Nevertheless, this study provides a picture of the norms and practices of this aspect of the medical research enterprise in 2001 and identifies several areas for possible exploration and improvement in the future."

## Learn By Doing

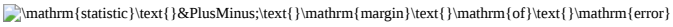
https://assessments.lumenlearning.co...sessments/3939

## Community College Student Athletes and Steroid Use

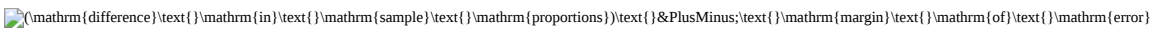
Robert D. Kersey published an article titled “Anabolic-Androgenic Steroid Use among California Community College Student Athletes” *(Journal of Athletic Trainers 31(3):237 – 41, 1996)* comparing various aspects of users and nonusers. The study used an advanced random sampling technique to select 10 representative community colleges in California and then to select a random sample of student athletes from the 10 colleges. The group of 1,185 male and female student-athletes completed an anonymous questionnaire. Of the sample, 4.2% of the males and 1.2% of the females admitted to using steroids.

### Let’s Summarize

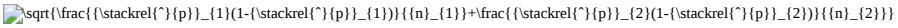
Every confidence interval has the following form:



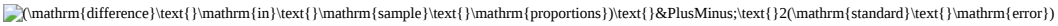
To estimate a difference in population proportions (or a treatment effect), the statistic is a difference in sample proportions, so the confidence interval is

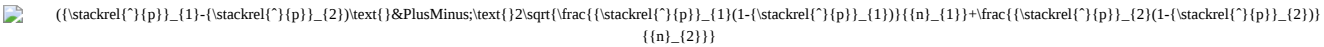


The margin of error is based on the estimated standard error in the sampling distribution:



If a normal model is a good fit for the sampling distribution, then the 95% confidence interval is





Use this formula only if a normal model is a good fit for the sampling distribution. A normal model is a good fit when the counts of successes and failures in both samples are at least 10.

When the conditions for normality are met, the confidence level is related to margin of error. To find a confidence interval for a different level of confidence, replace 2 with the appropriate *z*-score.

There are many differences between sample proportions, . Each of these differences generates its own confidence interval. The proportion of confidence intervals that contains the difference between the population proportions, , is equal to the level of confidence.

As always, “garbage in, garbage out.” The results of a confidence interval are meaningless without a properly designed study.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.11: Estimate the Difference between Population Proportions (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.12: Introduction- Hypothesis Test for Difference in Two Population Proportions

## What you'll learn to do: Construct and interpret an appropriate hypothesis test to compare two population/treatment group proportions.

### LEARNING OBJECTIVES

- Under appropriate conditions, conduct a hypothesis test for comparing two population proportions or two treatments. State a conclusion in context.
- Interpret the P-value as a conditional probability.
- Identify type I and type II errors and select an appropriate significance level based on an analysis of the consequences of each type of error.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.12: Introduction- Hypothesis Test for Difference in Two Population Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.13: Hypothesis Test for Difference in Two Population Proportions (1 of 6)

## Learning Objectives

- Recognize when to use a hypothesis test or a confidence interval to compare two population proportions or to investigate a treatment effect for a categorical variable.
- Under appropriate conditions, conduct a hypothesis test for comparing two population proportions or two treatments. State a conclusion in context.

## Introduction

In *Inference for Two Proportions*, our focus is on inference that compares two populations or two treatments with a categorical response variable. The parameters and statistics are proportions. In the section "Estimate the Difference between Population Proportions," we learned how to use a difference in sample proportions to calculate a confidence interval. The confidence interval estimates a treatment effect or the difference between two population proportions. In this section, "Hypothesis Test for a Difference in Population Proportions," we learn to use a difference in sample proportions to test a hypothesis about a treatment effect or a hypothesis that compares two population proportions.

We did hypothesis tests in *Inference for One Proportion*. Each claim involved a single population proportion. Now we will test claims about a treatment effect or about a difference in population proportions, and we'll see that the steps and the logic of the hypothesis test are the same. Before we get into the details, let's practice identifying research questions and studies that involve two populations or two treatments with a categorical response variable. Here are some examples.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3622

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3623

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3624

## Stating Hypotheses about Two Population Proportions

Whenever we test a hypothesis, we begin by stating null and alternative hypotheses.

The null hypothesis is a statement of "no effect" or "no difference," so the null hypothesis for all hypothesis tests about two population proportions is $H_0$: $p_1 - p_2 = 0$. When we say there is no difference in the population proportions (or no treatment effect), it is equivalent to saying that the population proportions are equal: $p_1 = p_2$.

The alternative hypothesis is one of the following:

$H_a$: $p_1 - p_2 > 0$ (or $p_1 > p_2$)
$H_a$: $p_1 - p_2 < 0$ (or $p_1 < p_2$)
$H_a$: $p_1 - p_2 \neq 0$ (or $p_1 \neq p_2$)

## Example

## The Abecedarian Project

*Will early childhood education improve the likelihood of college attendance for poor children?* Recall the experiment conducted by the Abecedarian (*A-B-C-Darian*) project in the 1970s. The study randomly assigned children to a control group (with no preschool) or a treatment group (with high-quality preschool).

To test the claim that the treatment increases the proportion of children who eventually attend college, we define a null and an alternative hypothesis.

Define $p_1$ to be the proportion of children who attend a quality preschool and eventually go to college. Define $p_2$ to be the proportion of children who did not attend preschool but eventually go to college.

The null hypothesis is always a statement of “no effect” or “no difference,” so we assume that these proportions are equal: $p_1 = p_2$. Their difference is therefore zero:

$H_0: p_1 - p_2 = 0$

In this example, the null hypothesis says that the preschool treatment has no effect on the proportion of children who eventually go to college.

The alternative hypothesis reflects our claim of a treatment effect. We chose to make $p_1$ connected to the treatment, so our claim says that $p_1$ is greater than $p_2$ ($p_1 > p_2$). This translates into a difference that is greater than zero. It is positive:

$H_a: p_1 - p_2 > 0$

Establishing the null and alternative hypotheses in a comparison of two proportions is an important part of the hypothesis testing process. The next few activities provide an opportunity to practice this skill.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3644

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.13: Hypothesis Test for Difference in Two Population Proportions (1 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.14: Hypothesis Test for Difference in Two Population Proportions (2 of 6)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test for comparing two population proportions or two treatments. State a conclusion in context.
- Interpret the P-value as a conditional probability.

Before we get into the details of the hypothesis test for a difference in two population proportions, let's review the general steps in hypotheses testing that we learned in *Inference for One Proportion*. The steps and the logic of the hypothesis test are the same as in that module. We also practiced this type of thinking more informally in the section "Distribution of Differences in Sample Proportions."

**Step 1: Determine the hypotheses.**

The hypothesis comes from the research question. The null hypothesis is a statement of "no effect" or "no difference." The alternative hypothesis reflects our claim.

**Step 2: Collect the data.**

Ideally, we select two independent random samples from two populations, or we randomly assign subjects to two treatments in an experiment.

**Step 3: Assess the evidence.**

We assume that the null hypothesis is true. This means we assume the population proportions are the same. Then we ask, *Could the data come from populations with the same proportions?* Now imagine taking random samples from these populations. Even if there is no difference between the population proportions, we expect variability in the differences between sample proportions. These differences are due to chance. We use simulation or a mathematical model to see how much the differences in sample proportions vary. Then we figure out if the difference we see in the data is likely or unlikely. Note that the wording "likely or unlikely" implies that this step requires some kind of probability calculation. We will again find a P-value. As before, the P-value is a probability related to the sampling distribution. It describes the chance that random samples will have a difference in sample proportions that is at least as extreme as we see in the data if *the null hypothesis is true*. "At least as extreme as" means as far from the center of the sampling distribution or further.

**Step 4: State a conclusion.**

We use the P-value to make a decision. The P-value helps us determine if the difference in proportions seen in the data is statistically significant or due to chance. One of two outcomes can occur.

- One possibility: The difference in sample proportions from the data is extremely unlikely. In this case, there is only a small chance that proportions from random samples differ more than what we observed in the data. So the probability (the P-value) is small, suggesting that the data did not come from populations with the same proportions. We view this as strong evidence against the null hypothesis. We reject the null hypothesis in favor of the alternative hypothesis.
- The other possibility: The difference in sample proportions observed in the data are fairly likely (not unusual). In this case, it is not surprising to see proportions from random samples with larger absolute differences than we observed in the data. The probability is large enough that we don't think the data is unusual. It could come from populations with the same proportions. A large P-value suggests that we do not have evidence against the null hypothesis, so we cannot reject it in favor of the alternative hypothesis.

Before we get into the details of the hypothesis test, let's practice using the P-value to make a decision.

## Learn By Doing

Recall the Abecedarian Project. In this experiment researchers randomly assigned infants to a treatment or control group. The treatment group received 5-years of high quality pre-school. We previously stated the following hypotheses to test the claim that a larger proportion of children who received the treatment will attend college.

$H_0 : p_1 - p_2 = 0$
$H_a : p_1 - p_2 > 0$

$p_1$ is the the proportion of children who attend a quality preschool that eventually go to college.

$p_2$ is the the proportion of children who did not attend a quality preschool that eventually go to college.

https://assessments.lumenlearning.co...sessments/3940

https://assessments.lumenlearning.co...sessments/3657

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3645

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.14: Hypothesis Test for Difference in Two Population Proportions (2 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.15: Hypothesis Test for Difference in Two Population Proportions (3 of 6)

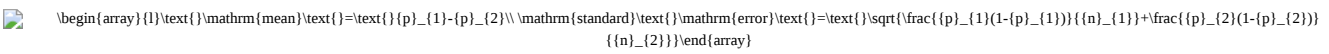
## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test for comparing two population proportions or two treatments. State a conclusion in context.
- Interpret the P-value as a conditional probability.

## Details of This Hypothesis Test

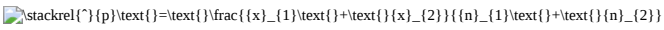
In a hypothesis test, we base our conclusion on the P-value. Where does the P-value come from? The P-value comes from a normal model of the sampling distribution of differences in sample proportions. In “Distribution of Differences in Sample Proportions,” we saw that a normal model is a good fit for the sampling distribution if each sample has at least 10 successes and failures.

We learned that the sampling distribution has the following center and spread.

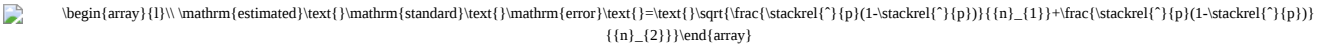



In the hypothesis test, we do not make a claim about either population proportion, so we do not have values for $p_1$ and $p_2$. For a confidence interval, we used the sample proportions, and , to estimate those values. Here we use a different estimate. Since the null hypothesis states that the population proportions are equal, we use the same estimate for both population proportions.

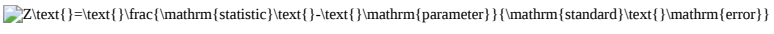
To do this, we combine the samples to create a *pooled proportion*. Here, $x_1$ and $x_2$ are the numbers of successes in the respective samples of sizes $n_1$ and $n_2$. We use the pooled proportion as an estimate for both population proportions.

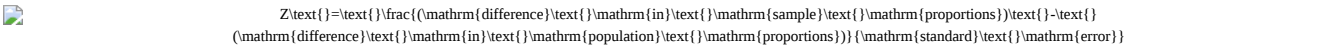


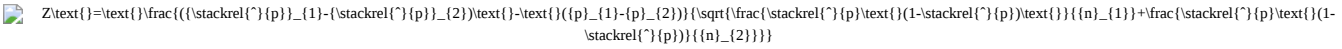

In a hypothesis test, we use the pooled proportion to estimate the standard error.



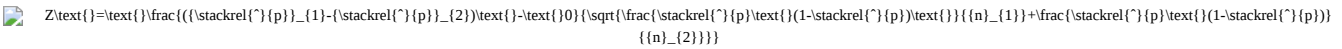

We use the estimated standard error to calculate the Z-test statistic.




Since $p_1 - p_2 = 0$ in the null hypothesis, the Z-test statistic simplifies to the following:




After we calculate the Z-test statistic, we use a simulation or other technology to find the P-value from the standard normal curve.

## Example

## Comparing Wal-Mart’s with Other Firms’ Insurance Coverage

Recall the 2003 press release by the AFL-CIO:

> *Wal-Mart exemplifies the harmful trend among America’s large employers to shirk health insurance responsibilities at the cost of their workers and community….Fewer than half of Wal-Mart workers are insured under the company plan – just 46 percent. This rate is dramatically lower than the 66 percent of workers at large private firms who are insured under their companies’ plans, according to a new Commonwealth Fund study released today.*

This press release claims that there is a 20% difference in the proportion of workers with insurance when we compare Wal-Mart to other large private firms. In hypothesis testing for two population proportions, we cannot test a claim about a specific difference between two population proportions. Instead, we test a claim that the proportion of Wal-Mart workers with health insurance is less than the proportion of workers at large private firms with health insurance.

Suppose we select a random sample of 50 Wal-Mart workers and find 23 have health insurance. Suppose also that a random sample of 70 workers of large private firms had 43 with health insurance.

For this test, we choose a 5% level of significance ($\alpha = 0.05$).

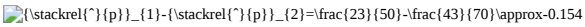
**Step 1: State the hypotheses.**

Let $p_1$ and $p_2$ represent the proportions of workers with health insurance among Wal-Mart and large private company employees respectively.

The null hypothesis is a claim of "no difference": $H_0$: $p_1 - p_2 = 0$. The alternative hypothesis states that the population proportion is lower for Wal-Mart employees: $p_1 < p_2$. The difference is less than zero, so it is negative: $H_a$: $p_1 - p_2 < 0$.

**Step 2: Collect the data.**

Of 50 Wal-Mart workers, 23 have health insurance. Of 70 workers from large private firms, 33 have health insurance. From the data, we can calculate the difference in sample proportions.



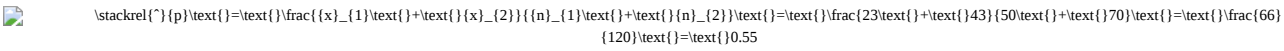
**Step 3: Assess the evidence.**

*Check the Normality Criteria*

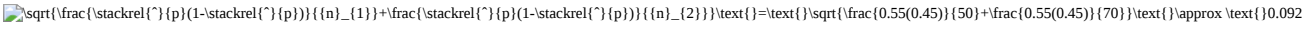
Determine if a normal model is a good fit for the sampling distribution. Verify that there are at least 10 successes and failures in each sample. Here, a *success* is an employee with health insurance. In the Wal-Mart sample, there are 23 successes and 50 − 23 = 27 failures. In the large private firms sample, there are 43 successes and 27 failures. Each of these is at least 10, so we can use the normal model.

*Compute the Test Statistic (only if the normal model is a good fit)*

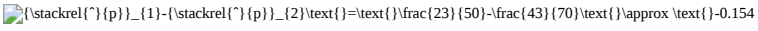
The test statistic requires the standard error. To compute the standard error, we first compute the pooled proportion.



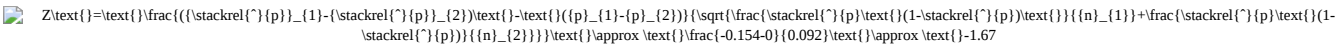
We use the pooled proportion to estimate the standard error.



Recall the difference in sample proportions from the data.



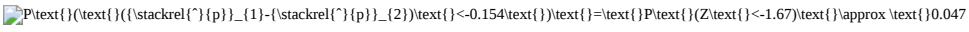
We use the *z*-score to determine how many standard errors −0.154 is from the mean of 0.

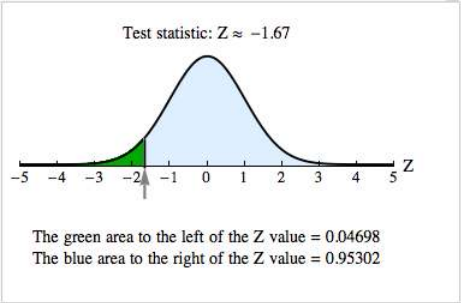


Note: A *z*-score of −1.67 tells us that the observed difference of = −0.154 is 1.67 standard errors below the assumed difference of zero. Does this suggest that the observed difference is statistically significant? Since we stated a significance level of 5%, we need to find the P-value and compare it to 0.05.

*Identify the P-Value*

We use a simulation. We want the probability that the difference in sample proportions is less than −0.154. This corresponds to the probability that *Z* is less than −1.67. So we use the area to the left of the Z-test statistic. The P-value is about 0.047. If you like symbols, we can write this in mathematical notation.






The P-value is small, about 4.7%. It means that if there is no difference in the population proportions, there is about a 4.7% chance that random samples will have a difference less than −0.154. The difference we observed in our samples, then, is fairly unlikely. We do not think this difference is due to chance. We see that the P-value is less than 5%, so we conclude that the difference we observed is statistically significant.

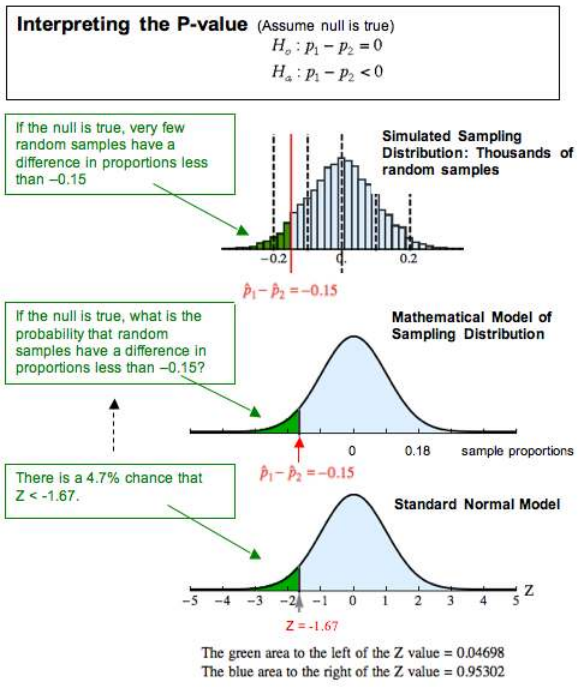
**Step 4: State a conclusion.**

*Use the P-Value to Make a Decision*

A P-value less than the significance level means we reject the null hypothesis. So we support the alternative hypothesis, $p_1 - p_2 < 0$, or more simply, $p_1 < p_2$. The given sample data support the claim that the proportion of Wal-Mart workers with health insurance is lower than the proportion of workers for large private companies.

## Comment

If a normal model is a good fit for the sampling distribution, we use it to find the P-value. But let's look at a simulation of the sampling distribution to remind ourselves what the P-value really means.




The simulation can help us understand the P-value. In the simulation, we assume that the population proportions are the same, so the difference is 0. This is the null hypothesis. We assume the null hypothesis is true and select thousands of random samples from populations with the same proportion of successes. The mean of the sampling distribution is 0 (as predicted by the null hypothesis). We see this in the simulated sampling distribution on the left.

We mark the difference in the sample proportions from our data. It is 23/50 − 43/70 = -0.15. This difference has a *z*-score of −1.67. In the simulation of the sampling distribution, we can see that a difference smaller than −0.15 is unlikely. Very few samples have a

difference less than −0.15. The normal model shows that the probability is about 4.7%.

Putting this all together, we have the formal definition of the P-value. The P-value is the probability that random samples have results at least as extreme as the data if the null hypothesis is true. We can also describe the P-value in terms of *z*-scores. The P-value is the probability that the test statistic has a value more extreme than that associated with the data if the null hypothesis is true.

### Learn By Doing

## Are There Gender Differences in Teen Depression Rates?

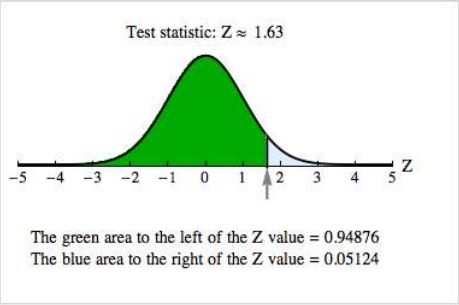
Previous studies suggest that female teens are more likely than male teens to be depressed. Define the depression rates for the female and male teens as $p_1$ and $p_2$ respectively. If we claim that the depression rate is higher for female teens ($p_1 > p_2$), the null and alternative hypotheses are:

$H_0$: $p_1 - p_2 = 0$
$H_a$: $p_1 - p_2 > 0$

Let's test the hypotheses at a 5% significance level. Suppose we randomly select 100 female teens and determine that 14 are clinically depressed. Among 200 randomly selected male teens, 16 are clinically depressed.

https://assessments.lumenlearning.co...sessments/3646

Since the normal model is a good fit, we can use the standard normal curve to find the P-value. We used a simulation. The P-value is about 0.051.




https://assessments.lumenlearning.co...sessments/3647

https://assessments.lumenlearning.co...sessments/3660

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.15: Hypothesis Test for Difference in Two Population Proportions (3 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.16: Hypothesis Test for Difference in Two Population Proportions (4 of 6)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test for comparing two population proportions or two treatments. State a conclusion in context.
- Given the description of a statistical study, evaluate whether conclusions are reasonable.

## A Reminder about Finding P-values

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.16: Hypothesis Test for Difference in Two Population Proportions (4 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.17: Hypothesis Test for Difference in Two Population Proportions (5 of 6)

## Learning Objectives

- Given the description of a statistical study, evaluate whether conclusions are reasonable.

## Thinking Critically about Conclusions from Statistical Studies

It is not uncommon to see debate over the conclusions and implications of statistical studies. When we read summaries of statistical studies, it is important to evaluate whether the conclusions are reasonable. Here we discuss two common pitfalls in drawing conclusions from statistical studies.

1. The conclusion is not appropriate to the study design.
2. The conclusion confuses statistical significance with practical importance.

We discuss these pitfalls in general, then look at examples that involve an inference about the difference between two population proportions or two treatments. But these pitfalls can happen with conclusions drawn from any inference procedure.

## The Conclusion Is Not Appropriate to the Study Design

Here are several examples of this common pitfall.

1. **The study makes an inference based on nonrandom data.**If the data come from a sample that is not randomly selected or from groups that are not randomly assigned, we should not use the data in inference procedures. Why? Well, all inference procedures are based on probability. We can make probability statements only about random events, so the data must come from randomly selected or randomly assigned individuals if we want to make a statement about the population on the basis of the data. With nonrandom data, our main option is to analyze the data using exploratory data analysis (the ideas from Modules 2, 3, and 4).
2. **The study makes inappropriate cause-and-effect conclusions.**We can make cause-and-effect conclusions only with data from a randomized comparative experiment. If data comes from a single observational study, we cannot make cause-and-effect conclusions.
3. **The study overgeneralizes its conclusions.**If researchers randomly assign individuals to one of two treatments in an experiment, statistically significant results suggest the treatment is effective. This is an appropriate causal conclusion if the experiment is well designed. But if the original group of individuals is not randomly selected, then we should be cautious about generalizing this conclusion to a broader population.

The following table summarizes these ideas.

## Learn By Doing

## Do Energy Drink “Cocktails” Lead to Increased Injury Risk?

The following excerpt is from “Energy Drink ‘Cocktails’ Lead to Increased Injury Risk, Study Shows” *(Science Daily, wwwesciencedaily.com, Nov. 6, 2007)*.

> *College students who drink alcohol mixed with so-called “energy” drinks are at dramatically higher risk for injury and other alcohol-related consequences, compared to students who drink alcohol without energy drinks.…The researchers found that students who consumed alcohol mixed with energy drinks were twice as likely to be hurt or injured, twice as likely to require medical attention, and twice as likely to ride with an intoxicated driver.*

The study collected data from a Web-based survey of 4,271 students at 10 North Carolina colleges and universities.

Following are two summaries of this study. Are these summaries appropriate? Explain.

https://assessments.lumenlearning.co...sessments/3945

https://assessments.lumenlearning.co...sessments/3946

## The Conclusion Confuses Statistical Significance with Practical Importance

Is a statistically significant difference always large enough to be important on a practical level? The answer is no.

Recall that when a P-value is less than the level of significance, we say the results are statistically significant. It means that the results are not due to chance. In the case of a difference in sample proportions, we are saying that the observed difference is larger than we expect to see in random samples from populations with the same population proportions. But this does not necessarily mean the difference is large enough to be important in real life.

We also know that the P-value depends on the size of the sample. Results from large samples vary less, so an observed difference is more likely to be statistically significant if the samples are large. This means that a very small difference in population parameters can be detected by the hypothesis test as statistically significant. In this case, the population difference may be too small to be important to decisions we might make in real life. On the other hand, small random samples can have a lot of variability in their results. In this case, a large population difference may go undetected by the hypothesis test because a large sample difference may not be statistically significant. We have to be cautious that we don't confuse statistical significance with practical importance.

### Example

## Controversy about HPV Vaccine

Recall the earlier example about the debate between Republican presidential candidates in 2011. Michele Bachmann, one of the candidates, implied that the vaccine for human papillomavirus (HPV) is unsafe for children and can cause mental retardation. In response, *USA Today* published an article on September 19, 2011, titled "No Evidence HPV Vaccines Are Dangerous." The article describes two studies by the Centers for Disease Control and Prevention (CDC) that track the safety of the vaccine. Here is an excerpt from the article.

> *First, the CDC monitors reports to the Vaccine Adverse Event Reporting System, a database to which anyone can report a suspected side effect. CDC officials then investigate to see whether reported problems could possibly be caused by vaccines or are simply a coincidence. Second, the CDC has been following girls who receive the vaccine over time, comparing them with a control group of unvaccinated girls….Again, the HPV vaccine has been found to be safe.*

We now examine "fake" data to demonstrate a couple of points. Suppose the CDC conducts a clinical trial to study the safety of the vaccine. Researchers select a random sample of girls and assign girls randomly to two groups: 1,000 girls get the vaccine, and 1,000 girls do not. Suppose that 6 girls in the vaccinated group develop serious health problems, and 1 girl in the unvaccinated group develops serious health problems.

*Is this difference statistically significant at the 5% level?*

Yes. If we use a statistical software package to find the P-value, we get a P-value of about 0.03. So the data supports the claim that the proportion of serious side effects is greater in the vaccine group ($P$ is about 0.03).

*Is the difference of practical importance?* We investigate ways to think about this next.

### Learn By Doing

## Controversy about HPV Vaccine

Suppose a headline summarizing this experiment says "Vaccine Leads to Significantly Higher Risk of Serious Health Problems for Girls." Indicate whether each critique of this headline is valid or invalid.

https://assessments.lumenlearning.co...sessments/3947

https://assessments.lumenlearning.co...sessments/3638

https://assessments.lumenlearning.co...sessments/3641

Someone reading the headline "Vaccine Leads to Significantly Higher Risk of Serious Health Problems for Girls" might think that it is very risky to have the vaccine. Practical importance here involves how someone judges the risk of vaccination.

Here are two headlines that make statements about the risk associated with the vaccine. Indicate whether the statements are valid or invalid.

https://assessments.lumenlearning.co...sessments/3642

https://assessments.lumenlearning.co...sessments/3643

https://assessments.lumenlearning.co...sessments/3948

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.17: Hypothesis Test for Difference in Two Population Proportions (5 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.18: Hypothesis Test for Difference in Two Population Proportions (6 of 6)

## Learning Objectives

- Identify type I and type II errors and select an appropriate significance level based on an analysis of the consequences of each type of error.

## Review of Type I and Type II Errors

Inference is based on probability, so there is always some chance of making a wrong decision. Recall that two types of wrong decisions can be made in hypothesis testing. When we reject a null hypothesis that is true, we commit a type I error. When we fail to reject a null hypothesis that is false, we commit a type II error.

The following table summarizes the logic behind type I and type II errors.

| | We Reject $H_0$. (accept $H_a$) | We Fail to Reject $H_0$ (not enough evidence to accept $H_a$) |
|---|---|---|
| $H_0$ is true. | Type I Error | Correct Decision |
| $H_0$ is false. ($H_a$ is true) | Correct Decision | Type II Error |

It is possible to have some influence over the likelihoods of committing these errors, but decreasing the chance of a type I error increases the chance of a type II error. We have to decide which error is more serious for a given situation. Sometimes a type I error is more serious, and other times a type II error is more serious.

## Learn By Doing

## Teens and Antidepressants

Recall the description of a clinical trial in which researchers study the effect of a new antidepressant on teens. Researchers design a randomized, controlled, double-blind experiment to study the effect of the antidepressant Fluoxetine combined with psychiatric therapy. The control group receives a placebo and psychiatric therapy. The response variable is *improvement*, which means symptoms of depression improve.

The hypotheses are as follows, with $p_1$ = proportion of teens who improve in the treatment group (Fluoxetine and psychiatric therapy) and $p_2$ = proportion of teens who improve in the control group (placebo and psychiatric therapy).

$H_0$: $p_1 - p_2 = 0$
$H_a$: $p_1 - p_2 > 0$

https://assessments.lumenlearning.co...sessments/3949

https://assessments.lumenlearning.co...sessments/3950

## Decreasing the Chance of Type I or Type II Error

How can we decrease the chance of a type I or type II error? Because decreasing the chance of a type I error increases the chance of a type II error, we have to weigh the consequences of these errors before deciding how to proceed.

Recall that the probability of committing a type I error is α. Why? Well, when we choose a level of significance (α) , we are choosing a benchmark for rejecting the null hypothesis. If the null hypothesis is true, then the probability that we will reject a true null hypothesis is α. So the smaller α is, the smaller the probability of a type I error.

It is more complicated to calculate the probability of a type II error. The best way to reduce the probability of a type II error is to increase the sample size. But once the sample size is set, larger values of α will decrease the probability of a type II error (while increasing the probability of a type I error).

Following are general guidelines for choosing a level of significance:

- If the consequences of a type I error are more serious, choose a small level of significance (α).

- If the consequences of a type II error are more serious, choose a larger level of significance (α). But remember that the level of significance is the probability of committing a type I error.
- In general, we pick the largest level of significance that we can tolerate as the chance of a type I error.

Note: It is not always the case that one type of error is worse than the other.

### Learn By Doing

## Hormone Replacement Therapy

Recall the experiment that investigated the side effects of hormone replacement therapy (HRT) for women with menopausal symptoms. The experiment randomly assigned over 16,000 U.S. women to receive a hormone treatment or a placebo. The experiment was double blind. After 5 years, a larger proportion of the hormone group had breast cancer and heart disease. This observed difference was statistically significant. Researchers were so alarmed by the results that the experiment was ended early to prevent further harm to the health of the women participating in the hormone group.

https://assessments.lumenlearning.co...sessments/3663

The type I error in this situation is that we conclude that HRT increases the risk of breast cancer and heart disease, but it does not. The type II error is that we conclude that HRT does not increase the risk of breast cancer and heart disease, but it does.

*Identify the type of error associated with each consequence.*

https://assessments.lumenlearning.co...sessments/3664

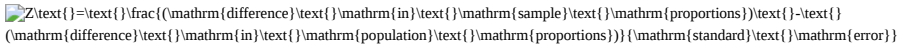
https://assessments.lumenlearning.co...sessments/3665

https://assessments.lumenlearning.co...sessments/3666

https://assessments.lumenlearning.co...sessments/3667

### Let's Summarize

- Hypothesis tests for two proportions can answer research questions about two populations or two treatments that involve categorical data.
- The alternative hypothesis is one of the following.

  $H_a$: $p_1 - p_2 < 0$, or
  $H_a$: $p_1 - p_2 > 0$, or
  $H_a$: $p_1 - p_2 \neq 0$
- 

  This statistic is approximately normal in its distribution if each sample has at least ten successes and failures. Note that the standard error is estimated with pooled proportion.
- The normal distribution may be used to provide P-values for a two-proportions test if each sample has at least 10 successes and failures.
- When the P-value in a two-proportions test is less than the level of significance (α), we should reject the null hypothesis in favor of the alternative. In this case, we say that the differences are statistically significant.
- Two types of errors can be made when conducting a hypothesis test.
  - A type I error occurs when we reject a true null hypothesis.
  - A type II error occurs when we fail to reject a false null hypothesis.
  - The level of significance, α, is the probability of a type I error.
  - Increasing the sample size lowers the probability of a type II error.
  - After considering the consequences of the type I and II errors, we should choose the largest value for α that we can tolerate, because increasing α decreases the probability of a type II error.
- (2) The conclusion confuses statistical significance with practical importance.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 9.18: Hypothesis Test for Difference in Two Population Proportions (6 of 6) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.19: Putting It Together- Inference for Two Proportions

## Let's Summarize

In *Inference for Two Proportions*, we learned two inference procedures to draw conclusions about a difference between two population proportions (or about a treatment effect): (1) a confidence interval when our goal is to estimate the difference and (2) a hypothesis test when our goal is to test a claim about the difference. Both types of inference are based on the sampling distribution.

### The Distribution of the Differences in Sample Proportions

In the section "Distribution of Differences in Sample Proportions," we learned about the sampling distribution of differences between sample proportions.

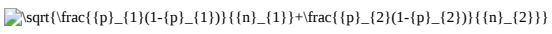
We used simulation to observe the behavior of sample differences when we select random samples from two populations. Every simulation began with an assumption about the difference between the two population proportions. From the simulated sampling distribution, we could determine if a sample difference observed in the data was likely or unlikely. A data result that is unlikely to occur in the sampling distribution provides evidence that our original assumption about the difference in the population proportions is probably incorrect. This logic is similar to the logic of hypothesis testing.

Because samples vary, we do not expect sample differences to always equal the population difference. Every sample difference has some error. We used simulations to observe the amount of error we expected to see in sample differences. The "typical" amount of error in the sampling distribution connects to the margin of error in a confidence interval.

We also used simulations to describe the shape, center, and spread of the sampling distribution. Later we developed a mathematical model for the sampling distribution with formulas for the mean of the sample differences and the standard deviation of the sample differences. We call this standard deviation *the standard error* because it represents an estimate for the average error we see in sample differences.

The mean of sample differences between sample proportions is equal to the difference between the population proportions, $p_1 - p_2$.

The standard error of differences between sample proportions is related to the population proportions and the sample sizes.



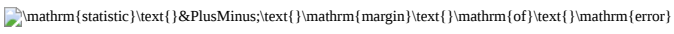
A normal model is a good fit for the sampling distribution of differences between sample proportions under certain conditions. We use a normal model if the counts of expected successes and failures are at least 10. For those who like formulas, this translates into saying the following four calculations must all be at least 10.

$n_1 p_1$
$n_1(1 - p_1)$
$n_2 p_2$
$n_2(1 - p_2)$

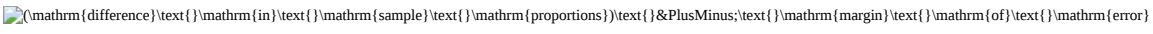
### Estimating the Difference between Two Population Proportions

In the section "Estimate the Difference between Population Proportions," we learned how to calculate a confidence interval to estimate the difference between two population proportions (or to estimate a treatment effect).

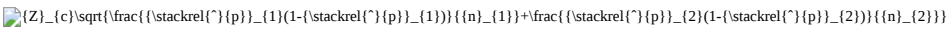
Every confidence interval has the form:



To estimate a difference in population proportions (or a treatment effect), the statistic is a difference in sample proportions. So the confidence interval is



Also, since we do not know the values of the population proportions, we estimate the standard error by using sample proportions in the formula for the margin of error.



Here are the critical $Z$-values for commonly used confidence levels.

| Confidence Level | Critical Value $Z_c$ |
|---|---|
| 90% | 1.645 |
| 95% | 1.960 |
| 99% | 2.576 |

The connection between the confidence level and critical $Z$-value depends on the use of a normal model. We use a normal model if each sample has at least 10 successes and failures.

We practiced interpreting confidence intervals and confidence levels. For example, we say we are "95% confident" that the population difference lies within the calculated confidence interval. We do not say there is a 95% chance that the population difference lies within the calculated interval. 95% confident means that in the long run 95% of the confidence intervals will contain the population differences.

### Hypothesis Test for a Difference in Two Population Proportions

In the section "Hypothesis Test for a Difference in Population Proportions," we tested claims regarding the difference between two population proportions (or a treatment effect).

In testing such claims, the null hypothesis is

$H_0$: $p_1 - p_2 = 0$

The alternative hypothesis is one of three:

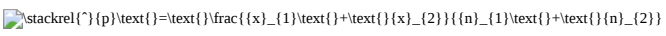
$H_a$: $p_1 - p_2 < 0$, or
$H_a$: $p_1 - p_2 > 0$, or
$H_a$: $p_1 - p_2 \neq 0$

These are equivalent to the following comparisons of $p_1$ and $p_2$.

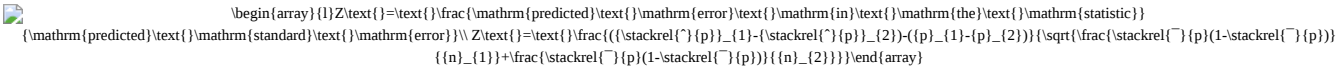
$H_a$: $p_1 < p_2$
$H_a$: $p_1 > p_2$
$H_a$: $p_1 \neq p_2$

We use the same criteria for determining if a normal model is a good fit for the sampling distribution: each sample must have at least 10 successes and failures.

In a hypothesis test, we assume the null hypothesis is true. Since we do not have values for $p_1$ and $p_2$, we again use sample data to estimate them. In the null hypothesis the population proportions are equal, so we create a single-value estimate for the population proportions using the pooled proportion.



With this pooled proportion, we estimate the standard error to compute the $Z$-test statistic for the hypothesis test. We can always view the $z$-score as the error in the statistic divided by the standard error. In a hypothesis test, we predict the error on the basis of the null, and we estimate the standard error.



If the conditions for approximate normality are met, this standardized statistic is approximately normal. This fact allows us to determine a P-value using computer software.

Whenever the P-value is less than or equal to the level of significance, we reject the null hypothesis in favor of the alternative. Otherwise, we fail to reject (but do not support) the null hypothesis.

Because our conclusions are based on probability, there is always a chance that our data will lead us to an incorrect conclusion. We make a type I error when we reject a true null hypothesis. We make a type II error when we fail to reject a false null hypothesis. These errors are not the result of a mistake. They are due to chance.

The level of significance, α, is the probability of a type I error. The probability of a type II error is harder to calculate. We did not learn to calculate type II error. Small values of α increase the probability of a type II error. Larger samples sizes decrease the probability of a type II error.

Finally, we should always remember “garbage in, garbage out.” If random selection or random assignment is not used to produce the data, we should not do inference.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.19: Putting It Together- Inference for Two Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 9.20: Assignment- A Statistical Investigation using Software

**Risk Factors for Low Birth Weight**

Rates of infant mortality, birth defect, and premature labor are high for babies with low birth weight. There are many factors that may contribute to low birth weight.

In this activity, we use data from a random sample of women who participated in a study in 1986 at the Baystate Medical Center in Springfield, MA. (*Source: Hosmer and Lemeshow (2000), Applied Logistic Regression: Second Edition.*)

For the 30 women in the study with a history of premature labor, a proportion of 18/30 = 0.60 (60%) had babies with low birth weight. For the remaining 159 women, a proportion of 41/159 = 0.26 (26%) had babies with low birth weight.

We now investigate the following research question: do the data provide evidence that the proportion of babies born with low birth weight is higher for women with a history of premature labor? This question is answered with a hypothesis test. To conduct the test we use a 1% level of significance.

### Question 1:

Is this study observational or experimental?

### Question 2:

Before analyzing the data, use your own experience and intuition to predict what the data will show. Do you think the proportion of babies with low birth weight is higher for women with a history of premature labor?

### Question 3:

We will test the claim that the proportion of women with low birth weight babies is higher among women with a history of premature labor. What are the null and alternative hypotheses?

### Question 4:

Are the criteria for approximate normality satisfied?

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 5:

State the test statistic and P-value. Interpret these values.

### Question 6:

Give a conclusion in context, and discuss whether a causal conclusion is appropriate.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 9.20: Assignment- A Statistical Investigation using Software is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 10: Inference for Means


10.1: Why It Matters- Inference for Means
10.2: Introduction- Distribution of Sample Means
10.3: Distribution of Sample Means (1 of 4)
10.4: Distribution of Sample Means (2 of 4)
10.5: Distribution of Sample Means (3 of 4)
10.6: Distribution of Sample Means (4 of 4)
10.7: Introduction- Estimating a Population Mean
10.8: Estimating a Population Mean (1 of 3)
10.9: Estimating a Population Mean (2 of 3)
10.10: Estimating a Population Mean (3 of 3)
10.11: Introduction- Hypothesis Test for a Population Mean
10.12: Hypothesis Test for a Population Mean (1 of 5)
10.13: Hypothesis Test for a Population Mean (2 of 5)
10.14: Hypothesis Test for a Population Mean (3 of 5)
10.15: Hypothesis Test for a Population Mean (4 of 5)
10.16: Hypothesis Test for a Population Mean (5 of 5)
10.17: Introduction- Inference for a Difference in Two Population Means
10.18: Inference for a Difference in Two Population Means
10.19: Hypothesis Test for a Difference in Two Population Means (1 of 2)
10.20: Hypothesis Test for a Difference in Two Population Means (2 of 2)
10.21: Estimating the Difference in Two Population Means
10.22: Putting It Together- Inference for Means
10.23: StatTutor- Analyzing Data From a Course's Grade Book
10.24: Assignment- Distribution of Sample Means
10.25: Assignment- Connection between Confidence Intervals and Sampling Distributions
10.26: Assignment- Hypothesis Testing for the Population Mean
10.27: Assignment- Matched Pairs
10.28: Assignment- Checking Conditions
10.29: Assignment- Two Independent Samples



This page titled 10: Inference for Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.1: Why It Matters- Inference for Means

## Learning Objectives

- Recognize when to use a hypothesis test or a confidence interval to draw a conclusion about a population mean.

In *Inference for Means*, we learn to make inferences about population means. Here are the types of research questions we focus on. Notice that we are working with quantitative variables for the first time in our inference work.

| Type of Question | Examples | Variable Type | Unit |
|---|---|---|---|
| Make an estimate about the population | What proportion of all U.S. adults support the death penalty? | Categorical variable | Inference for One Proportion |
| | **What is the average number of hours that community college students work each week?** | **Quantitative variable** | **Inference for Means** |
| Test a claim about the population | Do the majority of community college students qualify for federal student loans? | Categorical variable | Inference for One Proportion |
| | **Has the average birth weight in a town decreased from 3,500 grams?** | **Quantitative variable** | **Inference for Means** |
| Compare two populations | Are teenage girls more likely to suffer from depression than teenage boys? | Categorical variable | Inference for Two Proportions |
| | **In community colleges do female students have a higher average GPA than male students?** | **Quantitative variable** | **Inference for Means** |

Here again is the Big Picture. We have highlighted ideas new to this module in purple.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3668

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.1: Why It Matters- Inference for Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.2: Introduction- Distribution of Sample Means

## What you'll learn to do: Describe the sampling distribution of sample means.

### LEARNING OBJECTIVES

- Recognize when to use a hypothesis test or a confidence interval to draw a conclusion about a population mean.
- Describe the sampling distribution of sample means.
- Draw conclusions about a population mean from a simulation.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.2: Introduction- Distribution of Sample Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.3: Distribution of Sample Means (1 of 4)

## Learning Objectives

- Describe the sampling distribution of sample means.
- Draw conclusions about a population mean from a simulation.

## How Sample Means Vary in Random Samples

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.3: Distribution of Sample Means (1 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.4: Distribution of Sample Means (2 of 4)

## Learning Objectives

- Describe the sampling distribution of sample means.

Our next goal is to determine how the size of the sample affects the variability we see in sample means.

## Example

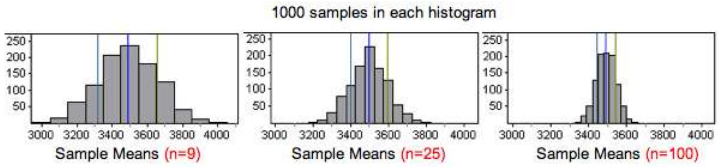
## Sample Size Affects Variability of Sample Means

We assumed that the population of individual babies has a mean of $\mu = 3{,}500$ grams and a standard deviation of $\sigma = 500$ grams. We selected a random sample of babies to test our assumptions about the population.

We saw previously that for this population of babies, it was not surprising to see a random sample of 9 babies with a mean birth weight of 3,400 grams. So a sample with a mean of 3,400 does not suggest that the town's mean birth weight is less than 3,500 grams this year.

*What if we increase the sample size? Will our conclusion change? That is, if the mean birth weight of 3,400 grams comes from a larger sample of babies, does the sample provide stronger evidence that the town's mean birth weight is less than 3,500 grams?*

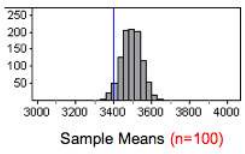
To investigate this question, we ran the simulation for different sample sizes. For each sample size, we collected 1,000 random samples and recorded the sample means.




When we compare the histograms of sample means, we notice the following:

- **Center:** The center is not affected by sample size. The mean of the sample means is always approximately the same as the population mean $\mu = 3{,}500$.
- **Spread:** The spread is smaller for larger samples, so the standard deviation of the sample means decreases as sample size increases. This is not surprising because we observed a similar trend with sample proportions.
- **Shape:** The sampling distributions all appear approximately normal. This is not surprising because the distribution of birth weights in the population has a normal shape.

Based on the histograms, it appears that sample size will change our conclusion about the population's mean birth weight this year. Suppose our sample mean of 3,400 grams came from a random sample of 100 babies. Means from samples this large did not vary much. We marked this sample result in a histogram for samples of size 100.




For $n = 100$, a sample mean of 3,400 grams is an unlikely result. It gives fairly strong evidence that the population's mean birth weight is less than 3,500 grams.

From advanced probability theory, we have a probability model for the sampling distribution of sample means. The model reinforces what we have already observed about the center and gives more precise information about the relationship between sample size and spread.

## The Theoretical Probability Model for the Sampling Distribution of Sample Means

Suppose a population has a mean μ and a standard deviation of σ. The distribution of all possible sample means from this population will have a mean of μ and a standard deviation of $\sigma\text{}/\sqrt{n}$.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3677

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3735

https://assessments.lumenlearning.co...sessments/3678

Let's compare and contrast what we know about the sampling distributions for sample means and sample proportions.

| | | | Sampling Distribution | | |
|---|---|---|---|---|---|
| Variable | Parameter | Statistic | Center | Spread | Shape |
| Categorical (example: left-handed or not) | $p$ = population proportion | $\stackrel{^}{p}$ = sample proportion | $p$ | $\sqrt{\frac{p(1-p)}{n}}$ | Normal when $np \geq 10$ and $n(1-p) \geq 10$ |
| Quantitative (example: age) | μ = population mean, σ = population standard deviation | $\overline{x}$ = sample mean | μ | $\frac{\sigma }{\sqrt{n}}$ | *When will the distribution of sample means be approximately normal?* |

We investigate the conditions that guarantee a normal sampling distribution for sample means on the next page.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.4: Distribution of Sample Means (2 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.5: Distribution of Sample Means (3 of 4)

## Learning Objectives

- Describe the sampling distribution of sample means.

## Shape of the Sampling Distribution of Means

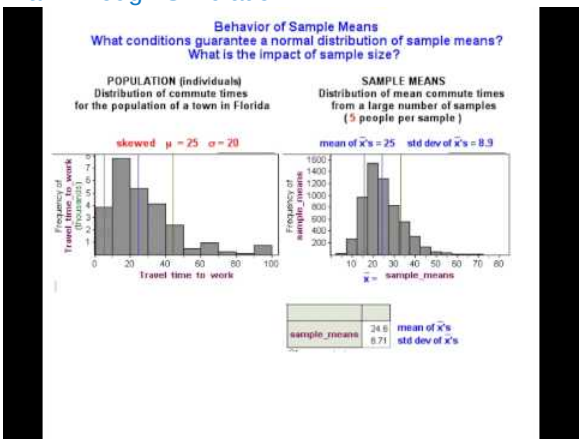
Now we investigate the shape of the sampling distribution of sample means. When we discussed the sampling distribution of sample proportions, we learned that this distribution is approximately normal if $np \geq 10$ and $n(1-p) \geq 10$. In other words, we had a guideline based on sample size for determining the conditions under which we could use a normal curve to do probability calculations for sample proportions.

Now we investigate these questions:

- *When will the distribution of sample means be approximately normal?*
- *Does it depend on the size of the sample?*
- *What happens if the distribution of the variable in the population is heavily skewed?*

The following simulation video helps us investigate these questions.

## WalkThrough Simulation




A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=382

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3736

https://assessments.lumenlearning.co...sessments/3679

## Comment

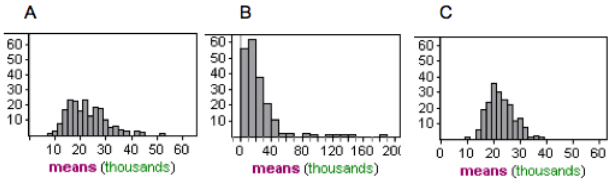
Are you surprised that a variable with a skewed distribution in the population can have a sampling distribution that is approximately normal? This discovery is probably the single most important result presented in introductory statistics courses. It is called the **central limit theorem**, which says that for large samples, the sampling distribution of sample means is approximately normal. This theorum is important! Inference procedures, such as hypothesis tests and confidence intervals, are based on a normal model for the sampling distribution. The central limit theorem assures us that we can use a normal probability model for sample means without knowing anything about the shape of the distribution of the variable in the population. All we have to do is collect large samples.

*How large a sample size do we need to assume that sample means will be normally distributed?* It really depends on the population distribution, as we saw in the simulation. The more skewed the distribution in the population, the larger the samples we need in order to use a normal model for the sampling distribution.

The general guideline is that samples of size greater than 30 will have a fairly normal distribution regardless of the shape of the distribution of the variable in the population. But if a population is strongly skewed, it is safer to use larger samples.

## Learn By Doing

The distribution of incomes is strongly skewed to the right for individuals in the U.S. The following histograms represent mean income from 200 samples randomly selected from the U.S. population. One histogram is based on samples of size of $n = 4$, one on samples of size of $n = 40$, and one on samples of size of $n = 100$.




https://assessments.lumenlearning.co...sessments/3680

https://assessments.lumenlearning.co...sessments/3681

## Summary

- Let's say we have a quantitative data set from a population with mean μ and standard deviation σ.The model for the theoretical sampling distribution of means of all random samples of size *n* has the following properties:
  - The mean of the sampling distribution of means is μ.
  - The standard deviation of the sampling distribution of means is $\frac{\sigma}{\sqrt{n}}$.
    - Notice that as *n* grows, the standard deviation of the sampling distribution of means shrinks.
- For large enough sample size, the sampling distribution of means is approximately normal (even if population is not normal).
  - If a variable has a skewed distribution for individuals in the population, a larger sample size is needed to ensure that the sampling distribution has a normal shape.
  - The general rule is that if *n* is more than 30, then the sampling distribution of means will be approximately normal. However, if the population is already normal, then any sample size will produce a normal sampling distribution.

## Comment

Notice that the size of the population is not mentioned in our discussion of sampling distributions. From our discussion, we know the following:

- The means from larger samples have less variability, so larger samples give more accurate estimates of the population mean.
- The means from larger samples have a distribution with a shape that is closer to normal.

These statements are true regardless of the size of the population as long as the population is large. To illustrate this point, we compare a distribution of sample means from two populations of different sizes. Population A has 10,000 newborns. Population B has 20,000 newborns. For each population, the mean and standard deviation of individual birth weights is the same: $\mu = 3{,}500$ and $\sigma = 500$.

We selected 525 random samples of 100 babies from each population and made a histogram of the sample means. We did this twice for population A, so two of the histograms represent 525 samples from the same population. As expected, there are some differences in the samples collected due to random chance. Comparing these two histograms gives us a sense of how much variation we can expect from the process of selecting random samples. Notice that the histogram of sample means from the larger population B has a similar shape, center, and spread to the histograms from population A.

## What's the Main Point?

The size of the population does not affect the variability of the sample means. Size matters if we are talking about sample size for random samples, but size does not matter if we are talking about population size as long as the population is large.

https://assessments.lumenlearning.co...sessments/3682

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.5: Distribution of Sample Means (3 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.6: Distribution of Sample Means (4 of 4)

## Learning Objectives

- We need to be careful before using the normal model to find probabilities associated with sample means.
  - Note: The logic of inference in this module is familiar. We make a claim about a population mean. We use a random sample to test our claim. We determine whether it is probable that random samples have means as extreme as the actual sample. If this is very unlikely, then we conclude this sample probably could not have come from this population and that the claim about the population mean is probably false. We used logic like this in Modules 7, 8, and 9 in the context of proportions. In this module, we further develop this idea in the context of means.

### Let's Summarize

- Many questions regarding quantitative variables require us to say something about the mean of a large population. It is often necessary to compute statistics from a random sample and use them to make an estimate or an inference about the population mean.
- We need to be able to compute the probability that the mean of a random sample falls in a given range. This probability allows us to draw an inference about the population parameter. To compute this probability, we need to understand the distribution of all sample means.
- Let's say we have a quantitative data set from a population with mean μ and standard deviation σ. The model for the theoretical sampling distribution of means of all random samples of size *n* has the following properties:
  - The mean of the sampling distribution of means is μ.
  - The standard deviation of the sampling distribution of means is $\sigma\text{}\/\sqrt{n}$.
    - Notice that as *n* grows, the standard deviation of the sampling distribution of means shrinks. It means that larger samples give more accurate estimates of population means.
- The central limit theorem states that for large enough sample sizes, the sampling distribution of means is approximately normal, even if the population is not normal.
  - If a variable has a skewed distribution for individuals in the population, a larger sample size is needed to ensure that the sampling distribution has a normal shape.
  - The general rule is that if *n* is more than 30, the sampling distribution of means will be approximately normal. However, if the population is already normal, then any sample size will produce a normal sampling distribution.
- The mechanics of finding a probability associated with a range of sample means usually proceeds as follows.
  - Convert a sample mean $\stackrel{\bar{}}{X}$ into a *z*-score: $Z\text{}=\text{}\frac{\stackrel{\bar{}}{X}-\mu}{\sigma\/\sqrt{n}}$.
  - Use technology to find a probability associated with a given range of *z*-scores.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.6: Distribution of Sample Means (4 of 4) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.7: Introduction- Estimating a Population Mean

## What you'll learn to do: Construct and interpret a confidence interval to estimate a population mean when conditions are met.

### LEARNING OBJECTIVES

- Construct a confidence interval to estimate a population mean when conditions are met. Interpret the confidence interval in context.
- Interpret the meaning of a confidence level associated with a confidence interval.
- Adjust the margin of error by making changes to the confidence level or sample size.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.7: Introduction- Estimating a Population Mean is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.8: Estimating a Population Mean (1 of 3)

## Learning Objectives

- Construct a confidence interval to estimate a population mean when conditions are met. Interpret the confidence interval in context.
- Interpret the meaning of a confidence level associated with a confidence interval.

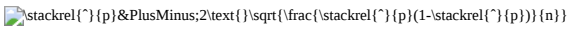
In “Estimating a Population Mean,” we focus on how to use a sample mean to estimate a population mean. This is the type of thinking we did in Modules 7 and 8 when we used a sample proportion to estimate a population proportion. Let’s take a moment to review what we learned in the modules *Linking Probability to Statistical Inference* and *Inference for One Proportion*, and then we’ll see how it relates to the current module.

- In *Linking Probability to Statistical Inference*, we noted that random samples vary, so we expect to see variability in sample proportions. In the section “Distribution of Sample Means” in that module, we made the same observations about sample means. In both cases, a normal model is a good fit for the sampling distribution when appropriate conditions are met.
- We also noted in that module that a sample proportion is an estimate for the population proportion. We do not expect the sample proportion to equal the population proportion, so there is some error. The error is due to random chance. Likewise, a sample mean is an estimate for the population mean, but there will be some error due to random chance.

| CONCEPT | FOR PROPORTIONS | FOR MEANS |
|---|---|---|
| When we take lots of random samples, we use the standard deviation of the sample statistics to describe the error that is due to random chance. This is called the standard error. The standard error depends on a population parameter and sample size. | Standard Error = $\sqrt{\frac{p(1-p)}{n}}$ | Standard Error = $\sigma/\sqrt{n}$ |
| If a normal model is a good fit for the sampling distribution, then 95% of sample statistics estimate the population parameter within 2 standard errors. This describes the margin of error. | Margin of Error = $2\sqrt{\frac{p(1-p)}{n}}$ | Margin of Error = $2\sigma/\sqrt{n}$ |
| We create an interval to estimate the population parameter. We say we are 95% confident that the confidence interval contains the population parameter. | Confidence interval is $\hat{p} \pm \textit{margin of error}$ $\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}$ | Confidence interval is $\bar{X} \pm \textit{margin of error}$ $\bar{X} \pm 2\sigma/\sqrt{n}$ |
| These confidence intervals require a normal model. So we can only use these formulas when the normality criteria are met. | Both expected successes and failures at least 10 | Variable is normally distributed in the population OR sample size is more than 30 |

## Comment

Recall that, in *Inference for One Proportion*, we adjusted the standard error by replacing $p$ with the sample proportion. Doing so made sense because the goal of the confidence interval is to estimate $p$. So the margin of error in the confidence interval formula changed. Here is the adjusted formula.




This adjustment changed the normality conditions. We use this adjusted confidence interval to estimate $p$ when the successes and failures in the actual sample are at least 10.

We will eventually have to adjust the standard error for the sampling distribution of sample means, too. It makes sense because in many situations we will not know the population standard deviation, σ. This adjustment is more complicated than the adjustment to standard error for sample proportions, so before we do it, let’s practice finding the confidence interval for μ assuming we know σ.

Assuming we know σ is realistic when a lot of previous research has been done. For example, when we are estimating height, weight, or scores on a standardized test, previous research gives us reliable values for σ.

## Example

## Estimating Mean SAT Math Score

The SAT is the most widely used college admission exam. (Most community colleges do not require students to take this exam.) The mean SAT math score varies by state and by year, so the value of μ depends on the state and the year. But let's assume that the shape and spread of the distribution of individual SAT math scores in each state is the same each year. More specifically, assume that individual SAT math scores consistently have a normal distribution with a standard deviation of 100.

An educational researcher wants to estimate the mean SAT math score (μ) for his state this year. The researcher chooses a random sample of 650 exams in his state. The average score is 475 (so $\stackrel{¯}{x}$ = 475). Estimate the mean SAT math score in this state for this year.

We answer this question by computing and interpreting a confidence interval.

**Checking conditions:**

From our work in "Distribution of Sample Means," we know that a normal model is a good fit for the distribution of sample means from random samples if one of two conditions is met:

- The population of individual values is normal (in which case the sample size is not important).
- If we do not know if the population of individual values is normal, then we must have a large sample size (more than 30).

Because we assume that the distribution of individual SAT math scores is normal in this example, a normal model is also a good fit for the distribution of sample means. Even if the population distribution had not been normal, the sample size is large enough that the normal distribution would still apply to the sample means. So we can use the confidence interval formula given above.

**Finding the margin of error:**

Keep in mind that the sample mean, $\stackrel{¯}{x}$, is only a single-value estimate for the population mean, μ. Because it comes from a random sample, we expect there to be some error in the estimate. *But how much error should we expect?*

We know that the sample distribution of means is approximately normal because conditions are met. Recall that in a normal model, 95% of the values fall within 2 standard deviations of the mean, so we use 2 standard errors for our margin of error. This was part of the empirical rule from the module *Probability and Probability Distribution*.

$$\begin{array}{l}\mathrm{standard}\text{ }\mathrm{error}\text{ }\mathrm{is}\text{ }\sigma\text{ }\sqrt{n}=100\text{ }\sqrt{650}\approx 3.9\\ \mathrm{margin}\text{ }\mathrm{of}\text{ }\mathrm{error}\text{ }\mathrm{is}\text{ }2(3.9)\approx 7.8\end{array}$$

**Finding the confidence interval:**

We are 95% confident that $\stackrel{¯}{x}$ falls within 7.8 points of μ. This also means that we are 95% confident that μ falls within 7.8 points of $\stackrel{¯}{x}$. So we construct a 95% confidence interval from this sample mean by adding and subtracting 7.8 points. The 95% confidence interval is shown.

$$\begin{array}{l}\stackrel{¯}{x}\pm \mathrm{margin}\text{ }\mathrm{of}\text{ }\mathrm{error}\\ \stackrel{¯}{x}\pm 2(\mathrm{standard}\text{ }\mathrm{error})\\ 475\pm 7.8\\ (467.2,484.8)\end{array}$$

**Conclusion:**

We are 95% confident that the mean SAT math score in this state this year is between 467.2 and 484.8. Recall from our previous work that being *95% confident* means this method, in the long run, captures the true population mean (μ) about 95% of the time.

### Summary

If we want to estimate μ, a population mean, we want to calculate a confidence interval. The 95% confidence interval is:

$$\stackrel{¯}{x}\pm 2\frac{\sigma}{\sqrt{n}}$$

We can use this formula only if a normal model is a good fit for the sampling distribution of sample means. If the sample size is large ($n$ > 30), we can use a normal model. If the sample size is not greater than 30, then we can use a normal model only if the variable is normally distributed in the population. As always, we must have a random sample. If the sample is not random, we cannot use it to estimate μ.

We say we are 95% confident that this interval contains μ, which means that in the long run, 95% of these confidence intervals contain μ.

## Constructing a Confidence Interval for Pregnancy Length

Is smoking during pregnancy associated with premature births? To investigate this question, researchers selected a random sample of 114 pregnant women who were smokers. The average pregnancy length for this sample of smokers was 260 days. From a large body of research, it is known that length of human pregnancy has a standard deviation of 16 days. The researchers assume that smoking does not affect the variability in pregnancy length.

https://assessments.lumenlearning.co...sessments/3751

https://assessments.lumenlearning.co...sessments/3752

### Comment

In our work with confidence intervals for estimating a population mean, μ, we require the population standard deviation, σ, to be known. In practice, σ usually is unknown. However, in some situations, especially when a lot of research has been done on the quantitative variable whose mean we are estimating (such as IQ, height, weight, scores on standardized tests), it is reasonable to assume that σ is known. On the next page, we learn how to proceed when σ is unknown.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

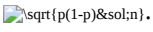
This page titled 10.8: Estimating a Population Mean (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

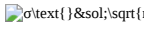
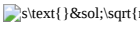
# 10.9: Estimating a Population Mean (2 of 3)

## Learning Objectives

- Construct a confidence interval to estimate a population mean when conditions are met. Interpret the confidence interval in context.
- Interpret the meaning of a confidence level associated with a confidence interval.
- Adjust the margin of error by making changes to the confidence level or sample size.

On the previous page, we used a confidence interval to estimate the population mean, μ. For this confidence interval, we had to supply a guess for the population standard deviation, σ, based on previous studies. It may have occurred to you that if we do not know μ, it is unlikely that we know σ. So we now take a different approach. We estimate σ using the sample standard deviation, *s*.

This is the same type of adjustment we used in *Inference for One Proportion* when we had to adjust our model of the sampling distribution. The standard error of the sampling distribution is  . (If we knew , then we wouldn't need to build a confidence interval.) We approximate by the sample proportion, .

Our process for adjusting the confidence interval for estimating μ is similar. We use the sample standard deviation, *s*, to estimate σ. The standard error for the sampling distribution  becomes  . So we adjust the margin of error in the confidence interval formula, but this adjustment is not as straightforward as our work with proportions.

This estimate for σ introduces more uncertainty in the process. The problem is worse with smaller samples because the sample standard deviations vary more. For small samples, *s* is a worse approximation for σ. Unfortunately, this approximation makes the normal model a bad fit and inappropriate for determining critical values. We instead use what is called a **t-model** for this purpose.

## Introduction to the T-Model

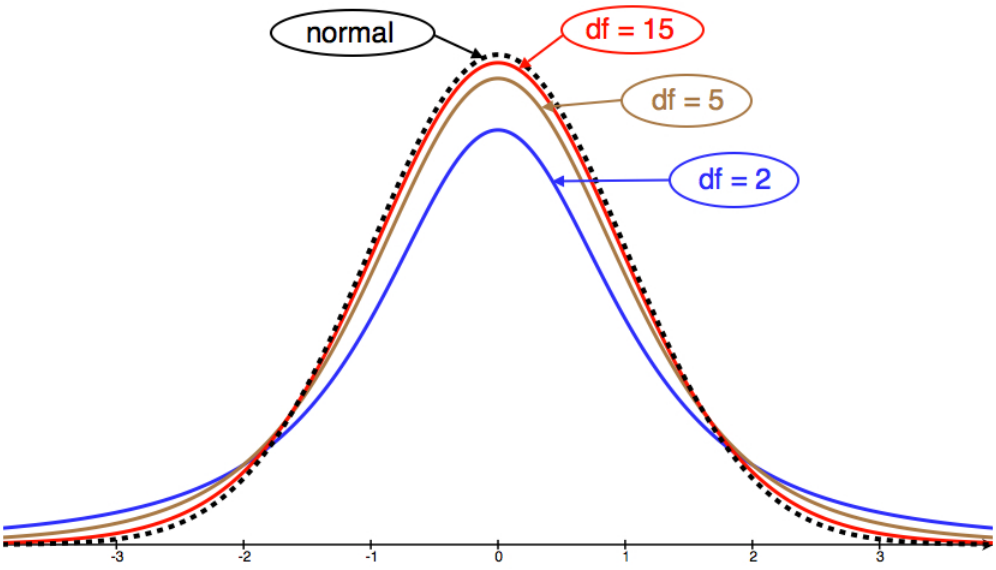
Here is the formula for the T-score. We also include the *z*-score for comparison. The formulas are very similar.

The distribution of *z*-scores is the standard normal curve, with mean of 0 and standard deviation of 1. The distribution of T-scores depends on the sample size, *n*. There is a different T-model for every *n*. So the T-model is a family of curves.

Instead of referring to *n* to specify which T-model to use, we refer to the **degrees of freedom**, or *df* for short. For Topics 10.2 and 10.3, the number of degrees of freedom is 1 less than the sample size. That is, $df = n - 1$.

In summary, a normal model is defined by its mean and standard deviation. A T-model is a family of curves defined by the degrees of freedom.

Let's take a look at a few T-model curves (for various *df*) to see how they compare to the normal model.




We can see from the picture that as *df* grows, the T-model gets closer to the standard normal model.

**Similarities between T-model and standard normal model:**

- Symmetric with a central peak, bell-shaped.
- Centered at 0.
- The larger the degrees of freedom, the closer the T-model is to the standard normal model.

**Difference between T-model and standard normal model:**

- The T-model has more spread than the standard normal model.
- The T-model has more probability in the tails and less in the center than the standard normal model. We can see this in the fatter tails and lower central peak of the T-model.

**When is a T-model a good fit for the sampling distribution of sample means?**

*Check these conditions before using the T-model:*

- Use the T-model if σ (the population standard deviation) is unknown. If σ is known, then use the normal model instead of the T-model.
- Use the T-model if variable values are normally distributed in the population. If this is not true, then make sure the sample size is large (more than 30).

### Example

## Cable Strength

A group of engineers developed a new design for a steel cable. They need to estimate the amount of weight the cable can hold. The weight limit will be reported on cable packaging.

The engineers take a random sample of 45 cables and apply weights to each of them until they break. The mean breaking weight for the 45 cables is $\stackrel{-}{x}$ = 768.2 lb. The standard deviation of the breaking weight for the sample is $s$ = 15.1 lb.

*What should the engineers report as the mean amount of weight held by this type of cable?*

Let's use these sample statistics to construct a 95% confidence interval for the mean breaking weight of this type of cable.

**Checking conditions:**

Since we do not know the standard deviation of breaking weights of all of the cables (the population parameter σ), we use the sample standard deviation ($s$) as an approximation for σ. Since we don't know σ, we must use the T-distribution to model the

sampling distribution of means.

*Is the T-model a good fit for the sampling distribution?*

Yes, because the conditions are met:

- σ is unknown.
- The sample size is large enough.

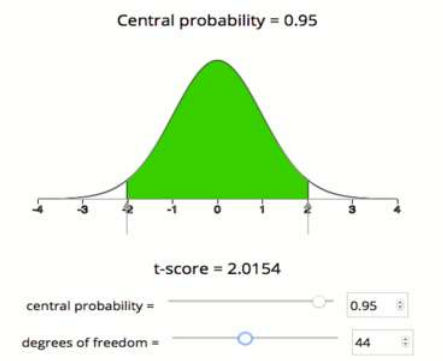
**Finding the standard error:**

As usual, we start by estimating the standard error. This estimate comes from the formula $\sigma/\sqrt{n}$. However, since we don't know σ, we use $s = 15.1$ as an approximation for σ. So our estimate for the standard error of all sample means is $s/\sqrt{n}=15.1/\sqrt{45}\approx 2.25$.

**Finding the margin of error:**

To find the margin of error, we need to find the critical T-value that corresponds to a 95% confidence level. This is just like the critical Z-value when we built confidence intervals for proportions, except that it comes from the T-model instead of the standard normal model.

We will use technology to find the critical T-value. There are a number of tools for doing this. Some books will also give you the option to use printed tables of values. Here we will use a simulation that gives the T-model based on degrees of freedom. We want the T-values that cut off the central 95% of the area under the curve. It will look as follows.




Using the simulation, we see that the critical T-value for a 95% confidence interval with 44 degrees of freedom is $T_c = 2.015$, which means our margin of error for this confidence interval is

$$\begin{array}{l}\text{Estimated standard error is } s/\sqrt{n}\approx 2.25\\ \text{Margin of error is } T_c \cdot s/\sqrt{n}\approx 2.015(2.25)\approx 4.53\end{array}$$

Note: For 95% confidence, the empirical rule approximates the critical Z-value as 2. The empirical rule is based on the normal model. Using the T-model for $df = 44$, the critical T-value (2.015) is very close to 2. This makes sense because for larger $df$, the T-model is very close to the standard normal model. We will see that the critical T-value differs more from the critical Z-value when the sample sizes are small.

**Finding the confidence interval:**

We have all the pieces to build the confidence interval. In our example, the confidence interval is

$$\begin{array}{l}\bar{x}\pm \text{margin of error}\\ \bar{x}\pm T_c\cdot\frac{s}{\sqrt{n}}\\ 768.2\pm 4.53\\ (763.7,772.7)\end{array}$$

**Conclusion:**

We are 95% confident that the mean breaking weight for all cables of this type is between 763.7 lb and 772.7 lb.

Confidence intervals at the 95% confidence level are common in practice. But 95% is not the only confidence level we use. Particularly in situations that involve safety issues, such as the previous example, people often prefer to estimate population means

with 99% confidence intervals. Let's do some exploration with technology to see how changes in the confidence level affect the confidence interval.

Click here to open this simulation in its own window.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=390

### Learn By Doing

## How Much Alcohol Do College Students Drink?

According to the website www.collegedrinkingprevention.gov, "About 25 percent of college students report academic consequences of their drinking including missing class, falling behind, doing poorly on exams or papers, and receiving lower grades overall." A statistics student is curious about drinking habits of students at his college. He wants to estimate the mean number of alcoholic drinks consumed each week by students at his college. He plans to use a 90% confidence interval. He surveys a random sample of 71 students. The sample mean is 3.93 alcoholic drinks per week. The sample standard deviation is 3.78 drinks.

https://assessments.lumenlearning.co...sessments/3684

https://assessments.lumenlearning.co...sessments/3753

https://assessments.lumenlearning.co...sessments/3685

https://assessments.lumenlearning.co...sessments/3686

https://assessments.lumenlearning.co...sessments/3687

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3754

https://assessments.lumenlearning.co...sessments/3755

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3691

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.9: Estimating a Population Mean (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

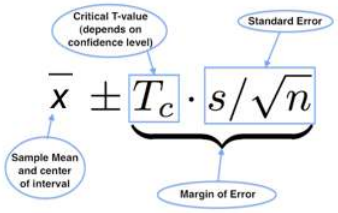
# 10.10: Estimating a Population Mean (3 of 3)

## Learning Objectives

- Construct a confidence interval to estimate a population mean when conditions are met. Interpret the confidence interval in context.
- Adjust the margin of error by making changes to the confidence level or sample size.

## Structure of a Confidence Interval

Let's take a closer look at the parts of the confidence interval. Remember that this is a confidence interval for a population mean. We use this formula when the population standard deviation is unknown.

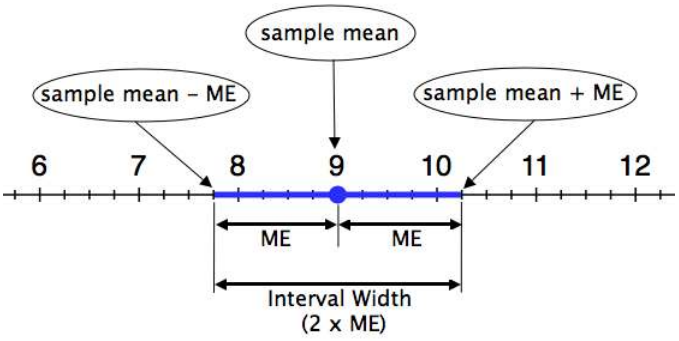



Let's remind ourselves how the confidence interval formula relates to the graph of the confidence interval on a number line.

The confidence interval shown below is a 95% confidence interval for a sample of size $n$ = 25 (so $df$ = 24), with sample mean $\overline{x}$ = 9 and sample standard deviation of $s$ = 3. The critical T-value for a 95% confidence interval with a $df$ = 24 is 2.064.

$$\begin{array}{l}\mathrm{Standard}\text{ }\mathrm{error}\text{ }\mathrm{is}\text{ }3\text{ }&sqrt{25}\text{ }=\text{ }0.6\\ \mathrm{Margin}\text{ }\mathrm{of}\text{ }\mathrm{error}\text{ }(\mathrm{ME})\text{ }\mathrm{is}\text{ }2.064(3)\text{ }&sqrt{}\text{ }\sqrt{25}\text{ }\approx \text{ }1.24\end{array}$$

The confidence interval is 9 ± 1.24. We are 95% confident that μ lies between 7.76 and 10.24.




Note:

- The sample mean (9 in this example) is at the center of the interval.
- The margin of error (labeled ME and equal to 1.24 in this example) is the distance that the interval extends to the left and right of the sample mean.
- The interval width is the length of the entire interval on the number line. The interval width is always twice the margin of error.

Let's quickly review how the *precision* of a confidence interval relates to the margin of error:

- An interval gives a *more precise* estimate when the interval is narrower. In other words, the margin of error is smaller.
- An interval gives a *less precise* estimate when the interval is wider. In other words, the margin of error is larger.

We know that a higher confidence level gives a larger margin of error, so confidence level is also related to precision.

- Increasing the confidence in our estimate makes the confidence interval wider and therefore less precise.
- Decreasing the confidence in our estimate makes the confidence interval narrower, and therefore more precise.

Confidence interval estimates are useful when they have the right balance of confidence and precision. Typical confidence levels used in practice are 90%, 95%, and 99%. When we need to be really sure about our estimates, such as in life-and-death situations, we choose a 99% confidence level. So if nothing else changes, we settle for less precise estimates when we need a high level of confidence.

In our discussion about the structure of confidence intervals, we said choosing a higher level of confidence means that we sacrifice some precision. This is true only if nothing else changes. But there is one way to keep a high level of confidence without sacrificing precision: Increase the sample size. We investigate the impact of sample size on the confidence interval next.

### Example

## Cable Strength Revisited

Recall the engineers who are trying to determine the breaking weight of a cable. In that example, we had a random sample of 45 cables with a mean breaking weight of 768.2 lb and a standard deviation of 15.1 lb. From that sample we computed a 95% confidence interval for the mean breaking weight of all such cables. Here are the important numbers we found from that calculation on the previous page:

$$\begin{array}{l}\mathrm{standard}\text{ }\mathrm{error}:\text{ }s\text{ }/\sqrt{n}\text{ }=\text{ }15.1\text{ }/\sqrt{45}\text{ }\approx \text{ }2.25\\ \mathrm{critical}\text{ }\text{T-value}:\text{ }{T}_{c}\text{ }=\text{ }2.015\text{ }(\mathrm{we}\text{ }\mathrm{found}\text{ }\mathrm{this}\text{ }\mathrm{using}\text{ }\mathrm{the}\text{ }\mathrm{simulation})\\ \mathrm{margin}\text{ }\mathrm{of}\text{ }\mathrm{error}:\text{ }{T}_{c}\text{ }\cdot \text{ }s\text{ }/\sqrt{n}\text{ }=\text{ }2.015(2.25)\text{ }=\text{ }4.53\\ \mathrm{confidence}\text{ }\mathrm{interval}:\text{ }768.2\text{ }±\text{ }4.53\text{ }\mathrm{or}\text{ }(763.67,772.73)\end{array}$$

Now let's increase the sample size and investigate the impact on the confidence interval. We calculate the confidence interval for a larger sample of 101 cables ($n$ = 101).

Sample size affects our calculations in two ways:

- The sample size ($n$) appears in our formula for standard error.
- The critical T-value depends on degrees of freedom, and $df = n - 1$.

**Finding the standard error:**

We approximate the standard error of all sample means as follows:

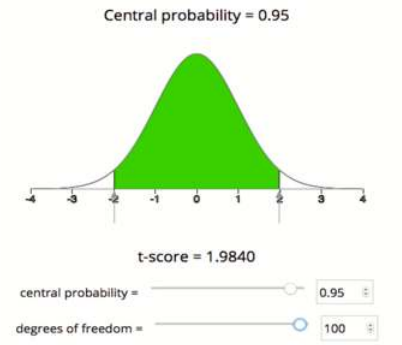


$$s\text{ }/\sqrt{n}\text{ }=\text{ }15.1\text{ }/\text{ }\sqrt{101}\text{ }\approx \text{ }1.50$$

Note: The standard error is smaller when the sample size is larger. We were expecting this because we know there is less variability in sample means when the samples are larger.

**Finding the critical T-value:**

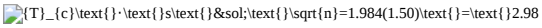
To find the critical T-value, we use the simulation. We set the $df$ to 100 and the central probability to 0.95. We see that the critical T-value is 1.984.




Note: Increasing the sample size decreased the critical T-value (the T-value went from 2.015 to 1.984 when we increased the sample size). You might also notice that both of the critical T-values for 95% confidence are larger than the critical Z-value for 95% confidence, which is approximately 1.96. This makes sense because the T-models are wider than the the standard normal curve.

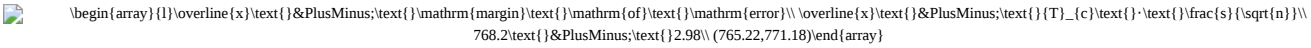
**Finding the margin of error.**

Here is the margin of error calculation:




**Finding the confidence interval.**

Here is the confidence interval calculation:




**Side-by-side comparison:**

Let's take a look at these two intervals to study the effects of changing the sample size.

| Sample Size | n = 45 | n = 101 |
|---|---|---|
| Standard Error | 2.25 | 1.50 |
| Critical T-value | 2.015 | 1.984 |
| Margin of Error | 4.53 | 2.98 |
| Confidence Interval | (763.67, 772.73) | (765.22, 771.18) |

Increasing the sample size had the following effects on the confidence interval estimate:

- Decreased standard error
- Decreased critical T-value
- Decreased margin of error and hence decreased the interval width
- Improved interval precision

### Comment

In the real world, increasing the sample size is not always possible. Sometimes collecting a sample is very expensive. If the study has budgetary constraints, which is usually the case, selecting a larger sample may be too expensive.

### Learn By Doing

## Appropriate Conclusions

For each of the following situations, decide if it is valid or invalid to use a confidence interval to estimate the population mean.

https://assessments.lumenlearning.co...sessments/3688

https://assessments.lumenlearning.co...sessments/3689

https://assessments.lumenlearning.co...sessments/3690

## Let's Summarize

- A confidence interval approximates a population mean by giving us a range of values that likely contains the population mean, μ. The general form of the confidence interval is $\stackrel{¯}{x}\text{}\pm\text{}\mathrm{margin}\text{}\mathrm{of}\text{}\mathrm{error}$.
- To say that we are "95% confident that the population mean falls within our confidence interval" really means that "about 95% of all confidence intervals computed in this way will capture the true population mean."
- We can use a sample mean to build a confidence interval as an estimate for μ. There are two possible cases:
  - Suppose the population standard deviation, σ, is known. We check the conditions for use of the normal model. Conditions: The variable must be normally distributed in the population, or the sample size is large enough ($n$ > 30). In this case, the confidence interval has the form $\stackrel{¯}{x}\text{}\pm\text{}{Z}_{c}\text{}\cdot\text{}\sigma\text{}/\text{}\sqrt{n}$.
  - Suppose the population standard deviation, σ, is not known. Then we use the sample standard deviation, $s$, as an approximation for σ. We check the conditions for use of the T-model. Conditions are the same: The variable must be normally distributed in the population, or the sample size is large enough ($n$ > 30). In this case, the confidence interval has the form $\stackrel{¯}{x}\text{}\pm\text{}{T}_{c}\text{}\cdot\text{}s\text{}/\text{}\sqrt{n}$ .When using the T-model to find the critical value, we need to select an appropriate number of degrees of freedom ($df$). The number of degrees of freedom is 1 less than the sample size ($df = n - 1$).
- As we have seen with other confidence intervals, the width of a confidence interval is twice the margin of error. The smaller the margin of error, the more narrow the confidence interval and the more precise the estimate of μ.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.10: Estimating a Population Mean (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.11: Introduction- Hypothesis Test for a Population Mean

## What you'll learn to do: Conduct and interpret results from a hypothesis test about a population mean.

### LEARNING OBJECTIVES

- Under appropriate conditions, conduct a hypothesis test about a population mean. State a conclusion in context.
- Under appropriate conditions, conduct a hypothesis test about a mean for a matched pairs design. State a conclusion in context.
- Interpret the P-value as a conditional probability.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.11: Introduction- Hypothesis Test for a Population Mean is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.12: Hypothesis Test for a Population Mean (1 of 5)

## Learning Objectives

- Recognize when to use a hypothesis test or a confidence interval to draw a conclusion about a population mean.
- Under appropriate conditions, conduct a hypothesis test about a population mean. State a conclusion in context.

## Introduction

In *Inference for Means*, our focus is on inference when the variable is quantitative, so the parameters and statistics are means. In "Estimating a Population Mean," we learned how to use a sample mean to calculate a confidence interval. The confidence interval estimates a population mean. In "Hypothesis Test for a Population Mean," we learn to use a sample mean to test a hypothesis about a population mean.

We did hypothesis tests in earlier modules. In *Inference for One Proportion*, each claim involved a single population proportion. In *Inference for Two Proportions*, the claim was a statement about a treatment effect or a difference in population proportions. In "Hypothesis Test for a Population Mean," the claims are statements about a population mean. But we will see that the steps and the logic of the hypothesis test are the same. Before we get into the details, let's practice identifying research questions and studies that involve a population mean.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3669

## Example

## Cell Phone Data

Cell phones and cell phone plans can be very expensive, so consumers must think carefully when choosing a cell phone and service. This decision is as much about choosing the right cellular company as it is about choosing the right phone. Many people use the data/Internet capabilities of a phone as much as, if not more than, they use voice capability. The data service of a cell company is therefore an important factor in this decision. In the following example, a student named Melanie from Los Angeles applies what she learned in her statistics class to help her make a decision about buying a data plan for her smartphone.

Melanie read an advertisement from the Cell Phone Giants (CPG, for short, and yes, we're using a fictitious company name) that she thinks is too good to be true. The CPG ad states that customers in Los Angeles get average data download speeds of 4 Mbps. With this speed, the ad claims, it takes, on average, only 12 seconds to download a typical 3-minute song from iTunes.

Only 12 seconds on average to download a 3-minute song from iTunes! Melanie has her doubts about this claim, so she gathers data to test it. She asks a friend who uses the CPG plan to download a song, and it takes 13 seconds to download a 3-minute song using the CPG network. Melanie decides to gather more evidence. She uses her friend's phone and times the download of the same 3-minute song from various locations in Los Angeles. She gets a mean download time of 13.5 seconds for her sample of downloads.

What can Melanie conclude? Her sample has a mean download time that is greater than 12 seconds. Isn't this evidence that the CPG claim is wrong? Why is a hypothesis test necessary? Isn't the conclusion clear?

Let's review the reason Melanie needs to do a hypothesis test before she can reach a conclusion.

**Why should Melanie do a hypothesis test?**

Melanie's data (with a mean of 13.5 seconds) suggest that the average download time overall is greater than the 12 seconds claimed by the manufacturer. But wait. We know that samples will vary. If the CPG claim is correct, we don't expect all samples to have a mean download time exactly equal to 12 seconds. There will be variability in the sample means. But if the overall average download time is 12 seconds, how much variability in sample means do we expect to see? We need to determine if the difference Melanie observed can be explained by chance.

We have to judge Melanie's data against random samples that come from a population with a mean of 12. For this reason, we must do a simulation or use a mathematical model to examine the sampling distribution of sample means. Based on the sampling

distribution, we ask, *Is it likely that the samples will have mean download times that are greater than 13.5 seconds if the overall mean is 12 seconds?* This probability (the P-value) determines whether Melanie's data provides convincing evidence against the CPG claim.

Now let's do the hypothesis test.

**Step 1: Determine the hypotheses.**

As always, hypotheses come from the research question. The null hypothesis is a hypothesis that the population mean equals a specific value. The alternative hypothesis reflects our claim. The alternative hypothesis says the population mean is "greater than" or "less than" or "not equal to" the value we assume is true in the null hypothesis.

Melanie's hypotheses:

$H_0$: It takes 12 seconds on average to download Melanie's song from iTunes with the CPG network in Los Angeles.
$H_a$: It takes more than 12 seconds on average to download Melanie's song from iTunes using the CPG network in Los Angeles.

We can write the hypotheses in terms of μ. When we do so, we should always define μ. Here μ = the average number of seconds it takes to download Melanie's song on the CPG network in Los Angeles.

$H_0$: μ = 12
$H_a$: μ > 12

**Step 2: Collect the data.**

To conduct a hypothesis test, Melanie knows she has to use a t-model of the sampling distribution. She thinks ahead to the conditions required, which helps her collect a useful sample.

Recall the conditions for use of a t-model.

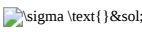
- There is no reason to think the download times are normally distributed (they might be, but this isn't something Melanie could know for sure). So the sample has to be large (more than 30).
- The sample has to be random. Melanie decides to use one phone but randomly selects days, times, and locations in Los Angeles.

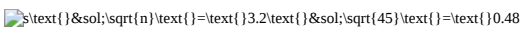
Melanie collects a random sample of 45 downloads by using her friend's phone to download her song from iTunes according to the randomly selected days, times, and locations.

Melanie's sample of size 45 downloads has an average download time of 13.5 seconds. The standard deviation for the sample is 3.2 seconds. Now Melanie needs to determine how unlikely this data is if CPG's claim is actually true.

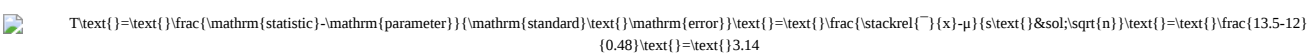
**Step 3: Assess the evidence.**

*Assuming the average download time for Melanie's song is really 12 seconds, what is the probability that 45 random downloads of this song will have a mean of 13.5 seconds or more?*

This is a question about sampling variability. Melanie must determine the standard error. She knows the standard error of random sample means is  . Since she has no way of knowing the population standard deviation, σ, Melanie uses the sample standard deviation, $s$ = 3.2, as an approximation. Therefore, Melanie approximates the standard error of all sample means ($n$ = 45) to be



Now she can assess how far away her sample is from the claimed mean in terms of standard errors. That is, she can compute the t-score of her sample mean.



The sample mean for Melanie's random sample is approximately 3.14 standard errors above the overall mean of 12. We know from previous experience that a sample mean this far above μ is very unlikely. With a t-score this large, the P-value is very small. We use a simulation of the t-model for 44 degrees of freedom to verify this.

We want the probability that the sample mean is greater than 13.5. This corresponds to the probability that T is greater than 3.14. The P-value is 0.0015.

**Step 4: State a conclusion.**

Here the logic is the same as for other hypothesis tests. We use the P-value to make a decision. The P-value helps us determine if the difference we see between the data and the hypothesized value of μ is statistically significant or due to chance. One of two outcomes can occur:

- One possibility is that results similar to the actual sample are extremely unlikely. This means the data does not fit with results from random samples selected from the population described by the null hypothesis. In this case, it is unlikely that the data came from this population. The probability as measured by the P-value is small, so we view this as strong evidence against the null hypothesis. We reject the null hypothesis in favor of the alternative hypothesis.
- The other possibility is that results similar to the actual sample are fairly likely (not unusual). This means the data fits with typical results from random samples selected from the population described by the null hypothesis. The probability as measured by the P-value is large. In this case, we do not have evidence against the null hypothesis, so we cannot reject it in favor of the alternative hypothesis.

Melanie's data is very unlikely if μ = 12. The probability is essentially zero (P-value = 0.0015). This means we will rarely see sample means greater than 13.5 if μ = 12. So we reject the null and accept the alternative hypothesis. In other words, this sample provides strong evidence that CPG has overstated the speed of its data download capability.

The following activities give you an opportunity to practice parts of the hypothesis testing process for a population mean. Later you will have the opportunity to practice the hypothesis test from start to finish.

### Learn By Doing

For the following scenarios, give the null and alternative hypotheses and state in words what μ represents in your hypotheses. A good definition of μ describes both the variable and the population.

https://assessments.lumenlearning.co...sessments/3764

https://assessments.lumenlearning.co...sessments/3765

https://assessments.lumenlearning.co...sessments/3766

### Comment

In the previous example, Melanie did not state a significance level for her test. If she had, the logic is the same as we used for hypothesis tests in Modules 8 and 9. To come to a conclusion about $H_0$, we compare the P-value to the significance level α.

- If $P \leq \alpha$, we reject $H_0$. We conclude there is significant evidence in favor of $H_a$.
- If $P > \alpha$, we fail to reject $H_0$. We conclude the sample does not provide significant evidence in favor of $H_a$.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3693

Use this simulation when needed to answer questions below.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3767

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3694

https://assessments.lumenlearning.co...sessments/3768

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.12: Hypothesis Test for a Population Mean (1 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.13: Hypothesis Test for a Population Mean (2 of 5)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a population mean. State a conclusion in context.

## More on Checking Conditions for a T-Test

In practice, you will often see the use of a t-test with small samples. Technically, we can use the t-test with small samples only if we know the variable has a normal distribution in the population. But this is hard to verify. In addition, no variable has a perfect normal distribution. So what does the requirement that the "variable be normally distributed in the population" really mean?

We call a confidence interval or a hypothesis test **robust** if the confidence level or P-value does not change very much when the conditions for use of the procedure are not met.

T-procedures are robust when the variable is not normally distributed in the population, as long as the distribution is not heavily skewed. But how can we determine if the distribution of the variable in the population is heavily skewed? In this introductory course, we examine the distribution of the variable in the sample and make an educated guess about what is going on in the population.

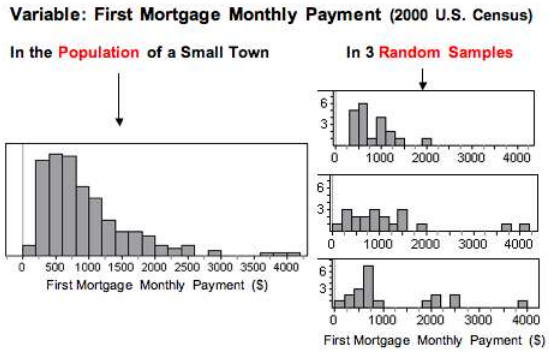
Now we investigate this question: *Can we tell from a sample whether the variable is normally distributed in the population?*

## Example

## Variable Skewed in the Population

Let's start with a skewed distribution in the population. Can we tell that this distribution is not normal by looking at random samples?

The following figure shows the monthly payment on first home mortgages for 5,000 people, as reported in the 2000 U.S. Census. Think of this as data from the population of a small town. From this population, we randomly selected 20 people. We did this three times. Notice that for each random sample, the shape of the distribution of the monthly payments in the sample is skewed to the right, just like the distribution in the population. In 2 of the 3 samples, we also see outliers, just as we see in the population. So by looking at the sample, we can get a pretty good sense that the variable is not normally distributed in the population.



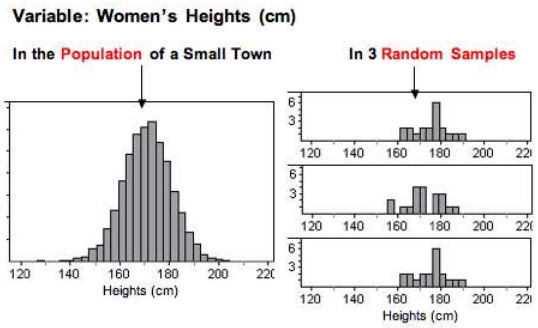

In this example, the sample size is less than 30. We can use the t-test only if the variable is normally distributed in the population. The shape of the distribution in any one of these samples suggests that the variable has a skewed distribution in the population, so we would not conduct a t-test with any of these samples.

## Example

## Variable Normal in the Population

Now we look at a variable that has an approximately normal distribution in the population. Can we tell that this distribution is approximately normal by looking at random samples?

The following graphs show the heights (in centimeters) of 5,000 women. Think of this as data from the population of a small town. From this population, we randomly selected 20 women. We did this three times. Notice that for each random sample, the shape of the distribution of the heights in the sample is not skewed, and there are no outliers. By looking at the sample, we can get a pretty good sense that the variable is not skewed in the population, which suggests that the variable may be somewhat normally distributed in the population.




In this example, the sample size is less than 30. We can only use the t-test if the variable is normally distributed in the population. The shape of the distribution in any one of these samples indicates that the variable does not have a skewed distribution in the population, suggesting that the distribution in the population is somewhat normal. Since the t-procedures are robust, we would conduct a t-test with any of these samples.

## What's the Main Point?

We previously stated the conditions for use of the t-procedures as follows:

(1) If the variable is normally distributed in the population, you can always use the t-procedures.

(2) If the variable is not normally distributed in the population (or you can't determine this factor), the sample size must be greater than 30 for safe use of the t-procedures.

We are now loosening these conditions somewhat because the t-procedures are robust.

(3) If the sample is small ($n \leq 30$), plot the data. If the distribution in the sample is not heavily skewed and does not have outliers, then we assume the variable is somewhat normally distributed in the population, so we use t-procedures.

## Comment

If we use a t-procedure for a small sample ($n \leq 30$), it is good practice to include a disclaimer with the conclusion. We might say something like, "On the basis of the sample, we are assuming that the variable is distributed without strong skew or extreme outliers in the population. The conclusion from this test is valid only if this assumption is true."

## Learn By Doing

Each histogram in the following questions represents a random sample. We do not know if the variable has a normal distribution in the population, but we want to run a t-test to test a claim about the population mean. For each histogram, choose the option that best describes how to proceed with the hypothesis test.

https://assessments.lumenlearning.co...sessments/3695

https://assessments.lumenlearning.co...sessments/3696

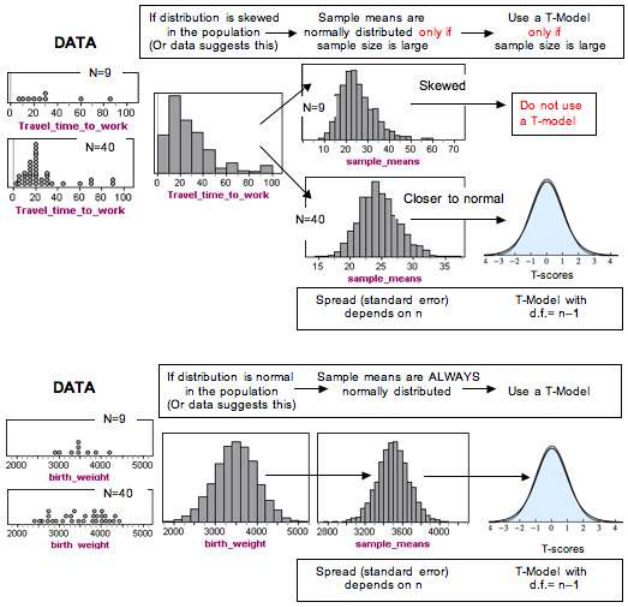
https://assessments.lumenlearning.co...sessments/3697

## Comment

Recall that the sample mean and standard deviation are not resistant to outliers. An outlier in the data can make the mean and standard deviation poor measures of center and spread. So why can we use data from large samples even if the data has an outlier? Well, if the sample is large enough, the distribution of sample means will still be approximately normal. And the t-model will be a good fit when we estimate the standard error of the sample means using the sample standard deviation. This is the important point.

The P-value and confidence level come from a model of the sampling distribution, not from a model of the population's distribution.

## Summary in a Diagram







CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.13: Hypothesis Test for a Population Mean (2 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.14: Hypothesis Test for a Population Mean (3 of 5)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a mean for a matched pairs design. State a conclusion in context.

Another common use of the t-test for a population mean is in "before and after" situations. In this situation, we have two quantitative measurements from a single sample of individuals. This is an example of a **matched-pairs** design.

## Example

## Drinking and Driving

The Centers for Disease Control and Prevention (CDC) website cites studies from the National Highway Traffic Safety Administration to support this statement: "Alcohol use slows reaction time and impairs judgment and coordination, which are all skills needed to drive a car safely. The more alcohol consumed, the greater the impairment." All states in the United States have adopted a blood alcohol concentration of 0.08% (80 mg/dL) as the legal limit for operating a motor vehicle. The CDC website continues, "Note: Legal limits do not define a level below which it is safe to operate a vehicle or engage in some other activity. Impairment due to alcohol use begins to occur at levels well below the legal limit."

It is this last statement that may be surprising to drivers. Interviews with drunk drivers who were involved in accidents reveal that drunk drivers do not realize how drunk they are. "I only had one or two drinks – I am okay to drive" is a common sentiment. Suppose a college conducts a study to call attention to this issue. Researchers use a random sample of 20 college students to examine the effect of drinking two beers on reaction times. They use a driving simulator to measure each student's reaction time before and after drinking two beers. The reaction time is the time it takes the student to hit the brakes in the simulator when an obstacle appears in the road.

Click here to see the data.

In this situation, we have two quantitative measurements for each student. To measure the effect of the two beers, we subtract the two reaction times to create one measurement of "change" or "effect." This controls the effects of individual characteristics that could influence reaction time, such as driving experience or natural quickness.

Here is a partial list of the data. We define the difference as "before minus after." If the "after" reaction time is longer, then the difference is negative. A negative value means the reaction time is slower after drinking.

Note: It is common to define the difference in measurements as "before minus after." But we could also define the difference the other way around as "after minus before." In this definition, if the "after" reaction time is longer, then the difference is positive, so a slower reaction time after drinking corresponds to a positive value. This makes less intuitive sense to us. We want "negative" to mean "drinking has a negative effect," so we used the other definition, "before minus after."

**Step 1: Determine the hypotheses.**

The null hypothesis is a claim of "no change" or "no effect." The alternative hypothesis reflects the claim.

$H_0$: Drinking two beers has no effect on reaction time.
$H_a$: Drinking two beers slows reaction time.

If drinking two beers has no effect on reaction time, then the mean of the differences in reaction times (before minus after) will be zero. If drinking two beers slows reaction time, then the mean of the differences in reaction times (before minus after) will be negative. So we can rewrite our hypotheses as follows.

$H_0$: $\mu = 0$
$H_a$: $\mu < 0$

where $\mu$ is the mean of the difference in reaction time (before minus after) for all students at this college after drinking two beers.

Suppose researchers set a significance level of 5%.

*Note:* if we defined the difference in reverse order as "after minus before," then a positive difference corresponds to a slower reaction time. This changes the alternative hypothesis to $H_a$: $\mu > 0$. This will not effect the P-value or our conclusion. You just have

to make sure the alternative hypothesis says what you want it to say given the way you define the difference.

*Note:* In some textbooks and other statistical materials, you will see the "mean of the difference" written as $\mu_d$.

**Step 2: Collect the data.**

Here is the (made-up) data from this sample of 20 college students. The mean of the differences (before minus after) is approximately −0.46 seconds, and the standard deviation is approximately 0.87 seconds.

**Step 3: Assess the evidence.**

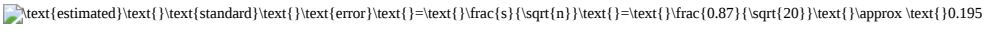
*Check the Criteria for Use of a T-Model*

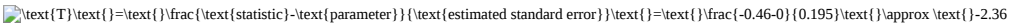
The sample size is only 20, and we do not know if these differences would be normally distributed in general when comparing these two treatments in the population of all college students. We therefore do not meet the conditions for use of a t-model. Some researchers would stop here and not complete the hypothesis test. Others would check the shape of the distribution of differences in the sample. If the sample is approximately normal (or at least not heavily skewed), then they view this as a hopeful indication that the distribution in the population will also be approximately normal, and they continue with the hypothesis test, adding a disclaimer to the conclusion.

Here is a histogram of the differences in the sample.

The data is not heavily skewed, so we are willing to proceed with the t-test.

*Compute the Test Statistic*




*Find the P-value.*

We use the simulation with the t-model for 19 degrees of freedom ($df = n - 1 = 20 - 1 = 19$). The P-value is approximately 0.015.

**Step 4: State a conclusion.**

The P-value (0.015) is less than the significance level (0.05), so we reject the null and accept the alternative hypothesis that $\mu < 0$.

This study suggests that reactions times when driving are significantly slower after drinking two beers for students at this college. ($P = 0.015$).

It is good practice to include a disclaimer with this conclusion because the sample is small. We might add the following to our conclusion if we were publicly presenting these results: "The sample was too small to formally met the requirements for a t-test. On the basis of the data, we are assuming that the difference in reaction times would be normally distributed in general when comparing these two treatments in the population of all college students. The conclusion from this test is valid only if this assumption is true."

## Comment

*From this study, can we generalize to a larger population of "all college students" or "all drivers"?* Technically, such a generalization requires that the sample be randomly selected from the more general population. Is this type of random sampling done in practice? Not always. The matched pairs design controls for individual differences that would otherwise confound such a generalization. This is one of the reasons that researchers run hypothesis tests with data gathered by groups like the National Highway Traffic Safety Administration, even when participants are not randomly selected. But ideally we should select samples randomly from the population of interest.

*Can we make a cause-and-effect conclusion from this study?*

We should be cautious about a cause-and-effect conclusion here because there is no random assignment. We take measurements from every student in both the treatment and the control setting without randomizing the treatment order. Every participant did the driving test sober, then drank two beers and did the driving test again. Technically, we need a more sophisticated study design that uses random assignment in order to make cause-and-effect conclusions. For example, we could randomly assign the treatment order for each student by flipping a coin, assigning some students do the first driving test while sober and others do the first driving

test after drinking two beers. Then everyone comes back on another day to complete the part of the experiment that he or she had not done.

The following activities give you an opportunity to practice parts of the hypothesis testing process. In each of these activities, the study is a matched-pairs design, so the population mean represents a mean of differences in paired measurements. Later you will have the opportunity to practice the hypothesis test from start to finish.

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3701

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3702

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.14: Hypothesis Test for a Population Mean (3 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.15: Hypothesis Test for a Population Mean (4 of 5)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a population mean. State a conclusion in context.
- Under appropriate conditions, conduct a hypothesis test about a mean for a matched pairs design. State a conclusion in context.

This page contains four opportunities for practicing the hypothesis test for a population mean from start to finish. The last two activities guide you through this hypothesis test using data and your statistical software package.

## How Much Are College Students Sleeping?

(not including naps in class!)

**Scenario:** Americans average 6.9 hours of sleep on weeknights, according to a report released in 2011 by the National Sleep Foundation. A student in a statistics class at Los Medanos College wondered if the average amount of sleep on weeknights is different for LMC students. She collected data from a survey of 43 randomly selected students at LMC. Respondents averaged 7.12 hours of sleep a night with a standard deviation of 1.45 hours.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3698

https://assessments.lumenlearning.co...sessments/3769

https://assessments.lumenlearning.co...sessments/3770

https://assessments.lumenlearning.co...sessments/3771

https://assessments.lumenlearning.co...sessments/3699

https://assessments.lumenlearning.co...sessments/3772

Use this simulation when needed to answer questions above.

## Texting and Concentration

**Scenario:** An instructor at Los Medanos College conducted an experiment with her statistics class to study the effect of texting on concentration. She created two audio clips in which she read two different lists of words. The treatment required students to send a short text message to a friend while listening to one of the audio clips. In the control setting, students simply listened to one of the audio clips. Everyone wore earphones and listened to the audio clips in the same order. But a coin flip determined who was and was not texting each time. After each listening session, students had 2 minutes to write down all the words they could remember. Twenty three students participated in the experiment.

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3703

https://assessments.lumenlearning.co...sessments/3773

https://assessments.lumenlearning.co...sessments/3774

https://assessments.lumenlearning.co...sessments/3775

https://assessments.lumenlearning.co...sessments/3776

https://assessments.lumenlearning.co...sessments/3704

https://assessments.lumenlearning.co...sessments/3777

Use this simulation when needed to answer questions above.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.15: Hypothesis Test for a Population Mean (4 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.16: Hypothesis Test for a Population Mean (5 of 5)

## Learning Objectives

- Interpret the P-value as a conditional probability.

We finish our discussion of the hypothesis test for a population mean with a review of the meaning of the P-value, along with a review of type I and type II errors.

## Review of the Meaning of the P-value

At this point, we assume you know how to use a P-value to make a decision in a hypothesis test. The logic is always the same. If we pick a level of significance (α), then we compare the P-value to α.

- If the P-value ≤ α, reject the null hypothesis. The data supports the alternative hypothesis.
- If the P-value > α, do not reject the null hypothesis. The data is not strong enough to support the alternative hypothesis.

In fact, we find that we treat these as "rules" and apply them without thinking about what the P-value means. So let's pause here and review the meaning of the P-value, since it is the connection between probability and decision-making in inference.

## Example

## Birth Weights in a Town

Let's return to the familiar context of birth weights for babies in a town. Suppose that babies in the town had a mean birth weight of 3,500 grams in 2010. This year, a random sample of 50 babies has a mean weight of about 3,400 grams with a standard deviation of about 500 grams. Here is the distribution of birth weights in the sample.

Obviously, this sample weighs less on average than the population of babies in the town in 2010. A decrease in the town's mean birth weight could indicate a decline in overall health of the town. *But does this sample give strong evidence that the town's mean birth weight is less than 3,500 grams this year?*

We now know how to answer this question with a hypothesis test. Let's use a significance level of 5%.

Let μ = mean birth weight in the town this year. The null hypothesis says there is "no change from 2010."

$H_0$: μ < 3,500
$H_a$: μ = 3,500

Since the sample is large, we can conduct the T-test (without worrying about the shape of the distribution of birth weights for individual babies.)

Statistical software tells us the P-value is 0.082 = 8.2%. Since the P-value is greater than 0.05, we fail to reject the null hypothesis.

**Our conclusion:** This sample does not suggest that the mean birth weight this year is less than 3,500 grams (*P*-value = 0.082). The sample from this year has a mean of 3,400 grams, which is 100 grams lower than the mean in 2010. But this difference is not statistically significant. It can be explained by the chance fluctuation we expect to see in random sampling.

## What Does the P-Value of 0.082 Tell Us?

A simulation can help us understand the P-value. In a simulation, we assume that the population mean is 3,500 grams. This is the null hypothesis. We assume the null hypothesis is true and select 1,000 random samples from a population with a mean of 3,500 grams. The mean of the sampling distribution is at 3,500 (as predicted by the null hypothesis.) We see this in the simulated sampling distribution.

In the simulation, we can see that about 8.6% of the samples have a mean less than 3,400. Since probability is the relative frequency of an event in the long run, we say there is an 8.6% chance that a random sample of 500 babies has a mean less than 3,400 if the population mean is 3,500. We can see that the corresponding area to the left of $T = -1.41$ in the T-model (with $df = 49$) also gives us a good estimate of the probability. This area is the P-value, about 8.2%.

If we generalize this statement, we say the P-value is the probability that random samples have results more extreme than the data if the null hypothesis is true. (By more extreme, we mean further from value of the parameter, in the direction of the alternative hypothesis.) We can also describe the P-value in terms of T-scores. The P-value is the probability that the test statistic from a random sample has a value more extreme than that associated with the data if the null hypothesis is true.

### Learn By Doing

## What Does a P-Value Mean?

Do women who smoke run the risk of shorter pregnancy and premature birth? The mean pregnancy length is 266 days. We test the following hypotheses.

$H_0$: $\mu = 266$
$H_a$: $\mu < 266$

Suppose a random sample of 40 women who smoke during their pregnancy have a mean pregnancy length of 260 days with a standard deviation of 21 days. The P-value is 0.04.

What probability does the P-value of 0.04 describe? Label each of the following interpretations as valid or invalid.

https://assessments.lumenlearning.co...sessments/3654

https://assessments.lumenlearning.co...sessments/3655

https://assessments.lumenlearning.co...sessments/3656

### Review of Type I and Type II Errors

We know that statistical inference is based on probability, so there is always some chance of making a wrong decision. Recall that there are two types of wrong decisions that can be made in hypothesis testing. When we reject a null hypothesis that is true, we commit a type I error. When we fail to reject a null hypothesis that is false, we commit a type II error.

The following table summarizes the logic behind type I and type II errors.

It is possible to have some influence over the likelihoods of committing these errors. But decreasing the chance of a type I error increases the chance of a type II error. We have to decide which error is more serious for a given situation. Sometimes a type I error is more serious. Other times a type II error is more serious. Sometimes neither is serious.

Recall that if the null hypothesis is true, the probability of committing a type I error is α. Why is this? Well, when we choose a level of significance (α), we are choosing a benchmark for rejecting the null hypothesis. If the null hypothesis is true, then the probability that we will reject a true null hypothesis is α. So the smaller α is, the smaller the probability of a type I error.

It is more complicated to calculate the probability of a type II error. The best way to reduce the probability of a type II error is to increase the sample size. But once the sample size is set, larger values of α will decrease the probability of a type II error (while increasing the probability of a type I error).

**General Guidelines for Choosing a Level of Significance**

- If the consequences of a type I error are more serious, choose a small level of significance (α).
- If the consequences of a type II error are more serious, choose a larger level of significance (α). But remember that the level of significance is the probability of committing a type I error.
- In general, we pick the largest level of significance that we can tolerate as the chance of a type I error.

### Learn By Doing

Let's return to the investigation of the impact of smoking on pregnancy length.

**Recap of the hypothesis test:** The mean human pregnancy length is 266 days. We test the following hypotheses.

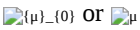
$H_0$: $\mu = 266$
$H_a$: $\mu < 266$

https://assessments.lumenlearning.co...sessments/3778

## Let's Summarize

In this "Hypothesis Test for a Population Mean," we looked at the four steps of a hypothesis test as they relate to a claim about a population mean.

**Step 1: Determine the hypotheses.**

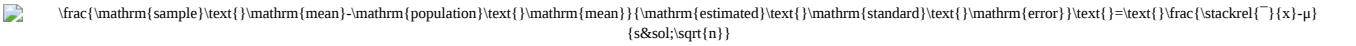
- The hypotheses are claims about the population mean, μ.
- The null hypothesis is a hypothesis that the mean equals a specific value, $\mu_0$.
- The alternative hypothesis is the competing claim that μ is less than, greater than, or not equal to the .
  - When  is  or  , the test is a one-tailed test.
  - When  is  , the test is a two-tailed test.

**Step 2: Collect the data.**

Since the hypothesis test is based on probability, random selection or assignment is essential in data production. Additionally, we need to check whether the t-model is a good fit for the sampling distribution of sample means. To use the t-model, the variable must be normally distributed in the population *or* the sample size must be more than 30. In practice, it is often impossible to verify that the variable is normally distributed in the population. If this is the case and the sample size is not more than 30, researchers often use the t-model if the sample is not strongly skewed and does not have outliers.

**Step 3: Assess the evidence.**

- If a t-model is appropriate, determine the t-test statistic for the data's sample mean.



- Use the test statistic, together with the alternative hypothesis, to determine the P-value.
- The P-value is the probability of finding a random sample with a mean at least as extreme as our sample mean, assuming that the null hypothesis is true.
- As in all hypothesis tests, if the alternative hypothesis is greater than, the P-value is the area to the right of the test statistic. If the alternative hypothesis is less than, the P-value is the area to the left of the test statistic. If the alternative hypothesis is not equal to, the P-value is equal to double the tail area beyond the test statistic.

**Step 4: Give the conclusion.**

The logic of the hypothesis test is always the same. To state a conclusion about $H_0$, we compare the P-value to the significance level, α.

- If $P \leq \alpha$, we reject $H_0$. We conclude there is significant evidence in favor of $H_a$.
- If $P > \alpha$, we fail to reject $H_0$. We conclude the sample does not provide significant evidence in favor of $H_a$.
- We write the conclusion in the context of the research question. Our conclusion is usually a statement about the alternative hypothesis (we accept $H_a$ or fail to accept$H_a$) and should include the P-value.

### Other Hypothesis Testing Notes

- Remember that the P-value is the probability of seeing a sample mean at least as extreme as the one from the data if the null hypothesis is true. The probability is about the random sample; it is not a "chance" statement about the null or alternative hypothesis.
- Hypothesis tests are based on probability, so there is always a chance that the data has led us to make an error.
  - If our test results in rejecting a null hypothesis that is actually true, then it is called a type I error.
  - If our test results in failing to reject a null hypothesis that is actually false, then it is called a type II error.

- If rejecting a null hypothesis would be very expensive, controversial, or dangerous, then we really want to avoid a type I error. In this case, we would set a strict significance level (a small value of α, such as 0.01).
- Finally, remember the phrase “garbage in, garbage out.” If the data collection methods are poor, then the results of a hypothesis test are meaningless.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.16: Hypothesis Test for a Population Mean (5 of 5) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.17: Introduction- Inference for a Difference in Two Population Means

## What you'll learn to do: Conduct a hypothesis test or construct a confidence interval to investigate a difference between two population means. Interpret results in context.

### LEARNING OBJECTIVES

- Under appropriate conditions, conduct a hypothesis test about a difference between two population means. State a conclusion in context.
- Construct a confidence interval to estimate a difference in two population means (when conditions are met). Interpret the confidence interval in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.17: Introduction- Inference for a Difference in Two Population Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.18: Inference for a Difference in Two Population Means

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a difference between two population means. State a conclusion in context.

## Introduction

In this section, we learn to make inferences about a difference between two population means. Our work here parallels our work in inference about a difference between two population proportions. Recall the following slogan from the previous module, *Inference for Two Proportions*.

*It's Not about the Values – It's about How They Are Related!*

So just as in that module, the value of the population means is not the focus of inference. Instead, we want to develop tools for determining the relationship between two unknown population means. We select independent random samples from two different populations and find the difference in the sample means. We use the sample difference either to conduct a hypothesis test about the difference in population means or to estimate the difference using a confidence interval.

## Example

## Beer and Reaction Time

Suppose researchers study the effect of low levels of alcohol on drivers' reaction time. Consider the following two study designs.

**Design 1:**

Researchers select a random sample of 19 drivers and assign them randomly to one of two treatments. The 9 drivers in the treatment group each drink two beers. The 10 drivers assigned to the control group do not drink beer. The response variable is *reaction time* (measured in seconds). The reaction time is the time it takes the driver to hit the brakes in a driving simulator when an obstacle appears in the road. The random assignment guarantees, at least in theory, that the two groups are independent.

In this design, we calculate a mean and standard deviation in response time for each group. We use the difference in the sample means to either test a hypothesis about, or calculate a confidence interval for, a difference in two population means or two treatments.

**Design 2:**

Researchers randomly select 8 drivers. The experiment is a matched-pairs design with two measurements taken for each driver. The researchers measure the reaction times in the driving simulator *before and then after* the consumption of two beers.

In this design, we first calculate the differences in the two measurements for each driver. Then we calculate the mean and standard deviation of this one list of numbers. We use the single sample mean to either test a hypothesis about, or calculate a confidence interval for, a single population or a treatment effect. This is one of the types of inference we did in the previous section, "Hypothesis Test for a Population Mean."

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3709

## Learn By Doing

Identify the situations that involve inference about a difference between two population means by choosing "valid" or "invalid."

https://assessments.lumenlearning.co...sessments/3710

https://assessments.lumenlearning.co...sessments/3711

https://assessments.lumenlearning.co...sessments/3712

https://assessments.lumenlearning.co...sessments/3713

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.18: Inference for a Difference in Two Population Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.19: Hypothesis Test for a Difference in Two Population Means (1 of 2)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a difference between two population means. State a conclusion in context.

## The Hypothesis Test for a Difference in Two Population Means

The general steps of this hypothesis test are the same as always. As expected, the details of the conditions for use of the test and the test statistic are unique to this test (but similar in many ways to what we have seen before.)

**Step 1: Determine the hypotheses.**

The hypotheses for a difference in two population means are similar to those for a difference in two population proportions. The null hypothesis, $H_0$, is again a statement of "no effect" or "no difference."

$H_0$: $\mu_1 - \mu_2 = 0$, which is the same as $H_0$: $\mu_1 = \mu_2$

The alternative hypothesis, $H_a$, can be any one of the following.

$H_a$: $\mu_1 - \mu_2 < 0$, which is the same as $H_a$: $\mu_1 < \mu_2$
$H_a$: $\mu_1 - \mu_2 > 0$, which is the same as $H_a$: $\mu_1 > \mu_2$
$H_a$: $\mu_1 - \mu_2 \neq 0$, which is the same as $H_a$: $\mu_1 \neq \mu_2$

**Step 2: Collect the data.**

As usual, how we collect the data determines whether we can use it in the inference procedure. We have our usual two requirements for data collection.

- Samples must be random to remove or minimize bias.
- Samples must be representative of the populations in question.

We use this hypothesis test when the data meets the following conditions.

- The two *random* samples are *independent*.
- The variable is *normally distributed in both populations*. If this variable is not known, *samples of more than 30* will have a difference in sample means that can be modeled adequately by the t-distribution. As we discussed in "Hypothesis Test for a Population Mean," t-procedures are robust even when the variable is not normally distributed in the population. If checking normality in the populations is impossible, then we look at the distribution in the samples. If a histogram or dotplot of the data does not show extreme skew or outliers, we take it as a sign that the variable is not heavily skewed in the populations, and we use the inference procedure. (Note: This is the same condition we used for the one-sample t-test in "Hypothesis Test for a Population Mean.")

**Step 3: Assess the evidence.**

If the conditions are met, then we calculate the t-test statistic. The t-test statistic has a familiar form.

$$T\text{ }=\text{ }\frac{(\mathrm{Observed}\text{ }\mathrm{difference}\text{ }\mathrm{in}\text{ }\mathrm{sample}\text{ }\mathrm{means})-(\mathrm{Hypothesized}\text{ }\mathrm{difference}\text{ }\mathrm{in}\text{ }\mathrm{population}\text{ }\mathrm{means})}{\mathrm{Standard}\text{ }\mathrm{error}}$$

$$T\text{ }=\text{ }\frac{({\stackrel{¯}{x}}_{1}-{\stackrel{¯}{x}}_{2})-({\mu}_{1}-{\mu}_{2})}{\sqrt{\frac{{{s}_{1}}^{2}}{{n}_{1}}+\frac{{{s}_{2}}^{2}}{{n}_{2}}}}$$

Since the null hypothesis assumes there is no difference in the population means, the expression $(\mu_1 - \mu_2)$ is always zero.

As we learned in "Estimating a Population Mean," the t-distribution depends on the **degrees of freedom (df)**. In the one-sample and matched-pair cases $df = n - 1$. For the two-sample t-test, determining the correct *df* is based on a complicated formula that we do not cover in this course. We will either give the *df* or use technology to find the *df*. With the t-test statistic and the degrees of freedom, we can use the appropriate t-model to find the P-value, just as we did in "Hypothesis Test for a Population Mean." We can even use the same simulation.

**Step 4: State a conclusion.**

To state a conclusion, we follow what we have done with other hypothesis tests. We compare our P-value to a stated level of significance.

- If the P-value ≤ α, we reject the null hypothesis in favor of the alternative hypothesis.
- If the P-value > α, we fail to reject the null hypothesis. We do not have enough evidence to support the alternative hypothesis.

As always, we state our conclusion in context, usually by referring to the alternative hypothesis.

### Example

## "Context and Calories"

Does the company you keep impact what you eat? This example comes from an article titled "Impact of Group Settings and Gender on Meals Purchased by College Students" *(Allen-O'Donnell, M., T. C. Nowak, K. A. Snyder, and M. D. Cottingham, Journal of Applied Social Psychology 49(9), 2011, onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.2011.00804.x/full)*. In this study, researchers examined this issue in the context of gender-related theories in their field. For our purposes, we look at this research more narrowly.

**Step 1: Stating the hypotheses.**

In the article, the authors make the following hypothesis. "The attempt to appear feminine will be empirically demonstrated by the purchase of fewer calories by women in mixed-gender groups than by women in same-gender groups." We translate this into a simpler and narrower research question: *Do women purchase fewer calories when they eat with men compared to when they eat with women?*

Here the two populations are "women eating with women" (population 1) and "women eating with men" (population 2). The variable is the calories in the meal. We test the following hypotheses at the 5% level of significance.

The null hypothesis is always $H_0$: $\mu_1 - \mu_2 = 0$, which is the same as $H_0$: $\mu_1 = \mu_2$.

The alternative hypothesis $H_a$: $\mu_1 - \mu_2 > 0$, which is the same as $H_a$: $\mu_1 > \mu_2$.

Here $\mu_1$ represents the mean number of calories ordered by women when they were eating with other women, and $\mu_2$ represents the mean number of calories ordered by women when they were eating with men.

Note: It does not matter which population we label as 1 or 2, but once we decide, we have to stay consistent throughout the hypothesis test. Since we expect the number of calories to be greater for the women eating with other women, the difference is positive if "women eating with women" is population 1. If you prefer to work with positive numbers, choose the group with the larger expected mean as population 1. This is a good general tip.

**Step 2: Collect Data.**

As usual, there are two major things to keep in mind when considering the collection of data.

- Samples need to be representative of the population in question.
- Samples need to be random in order to remove or minimize bias.

*Representative Samples?*

The researchers state their hypothesis in terms of "women." We did the same. But the researchers gathered data by watching people eat at the HUB Rock Café II on the campus of Indiana University of Pennsylvania during the Spring semester of 2006. Almost all of the women in the data set were white undergraduates between the ages of 18 and 24, so there are some definite limitations on the scope of this study. These limitations will affect our conclusion (and the specific definition of the population means in our hypotheses.)

*Random Samples?*

The observations were collected on February 13, 2006, through February 22, 2006, between 11 a.m. and 7 p.m. We can see that the researchers included both lunch and dinner. They also made observations on all days of the week to ensure that weekly customer patterns did not confound their findings. The authors state that "since the time period for observations and the place where [they] observed students were limited, the sample was a convenience sample." Despite these limitations, the researchers conducted

inference procedures with the data, and the results were published in a reputable journal. We will also conduct inference with this data, but we also include a discussion of the limitations of the study with our conclusion. The authors did this, also.

*Do the data met the conditions for use of a t-test?*

The researchers reported the following sample statistics.

- In a sample of 45 women dining with other women, the average number of calories ordered was 850, and the standard deviation was 252.
- In a sample of 27 women dining with men, the average number of calories ordered was 719, and the standard deviation was 322.

One of the samples has fewer than 30 women. We need to make sure the distribution of calories in this sample is not heavily skewed and has no outliers, but we do not have access to a spreadsheet of the actual data. Since the researchers conducted a t-test with this data, we will assume that the conditions are met. This includes the assumption that the samples are independent.

**Step 3: Assess the evidence.**

As noted previously, the researchers reported the following sample statistics.

- In a sample of 45 women dining with other women, the average number of calories ordered was 850, and the standard deviation was 252.
- In a sample of 27 women dining with men, the average number of calories ordered was 719, and the standard deviation was 322.

To compute the t-test statistic, make sure sample 1 corresponds to population 1. Here our population 1 is “women eating with other women.” So $x_1 = 850$, $s_1 = 252$, $n_1 = 45$, and so on.

$$T\text{ }=\text{ }\frac{\text{ }\stackrel{¯}{x}\text{ }_{1}\text{ }\text{ }-\text{ }\text{ }\stackrel{¯}{x}\text{ }_{2}\text{ }}{\sqrt{\frac{\text{ }s\text{ }_{1}\text{ }^{2}\text{ }}{\text{ }n\text{ }_{1}\text{ }}+\frac{\text{ }s\text{ }_{2}\text{ }^{2}\text{ }}{\text{ }n\text{ }_{2}\text{ }}}}\text{ }=\text{ }\frac{850\text{ }\text{ }-\text{ }719}{\sqrt{\frac{252\text{ }^{2}\text{ }}{45}+\frac{322\text{ }^{2}\text{ }}{27}}}\text{ }\approx \text{ }\frac{131}{72.47}\text{ }\approx \text{ }1.81$$

Using technology, we determined that the degrees of freedom are about 45 for this data. To find the P-value, we use our familiar simulation of the t-distribution. Since the alternative hypothesis is a “greater than” statement, we look for the area to the right of $T = 1.81$. The P-value is 0.0385.

**Step 4: State a conclusion.**

*Generic Conclusion*

The hypotheses for this test are $H_0$: $\mu_1 - \mu_2 = 0$ and $H_a$: $\mu_1 - \mu_2 > 0$. Since the P-value is less than the significance level ($0.0385 < 0.05$), we reject $H_0$ and accept $H_a$.

*Conclusion in context*

At Indiana University of Pennsylvania, the mean number of calories ordered by undergraduate women eating with other women is greater than the mean number of calories ordered by undergraduate women eating with men (P-value = 0.0385).

## A Comment about Conclusions

In the conclusion above, we did not generalize the findings to all women. Since the samples included only undergraduate women at one university, we included this information in our conclusion. But our conclusion is a cautious statement of the findings. The authors see the results more broadly in the context of theories in the field of social psychology. In the context of these theories, they write, “Our findings support the assertion that meal size is a tool for influencing the impressions of others. For traditional-age, predominantly White college women, diminished meal size appears to be an attempt to assert femininity in groups that include men.” This viewpoint is echoed in the following summary of the study for the general public on National Public Radio (npr.org).

> *Both men and women appear to choose larger portions when they eat with women, and both men and women choose smaller portions when they eat in the company of men, according to new research published in the Journal of Applied Social Psychology. The study, conducted among a sample of 127 college students, suggests that both men and women are influenced by unconscious scripts about how to behave in each other’s company. And these scripts change the way men and women eat when they eat together and when they eat apart.*

Should we be concerned that the findings of this study are generalized in this way? Perhaps. But the authors of the article address this concern by including the following disclaimer with their findings: “While the results of our research are suggestive, they should be replicated with larger, representative samples. Studies should be done not only with primarily White, middle-class college students, but also with students who differ in terms of race/ethnicity, social class, age, sexual orientation, and so forth.” This is an example of good statistical practice. It is often very difficult to select truly random samples from the populations of interest. Researchers therefore discuss the limitations of their sampling design when they discuss their conclusions.

In the following activities, you will have the opportunity to practice parts of the hypothesis test for a difference in two population means. On the next page, the activities focus on the entire process and also incorporate technology.

## Learn By Doing

### National Health and Nutrition Survey

https://assessments.lumenlearning.co...sessments/3705

https://assessments.lumenlearning.co...sessments/3782

https://assessments.lumenlearning.co...sessments/3706

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.19: Hypothesis Test for a Difference in Two Population Means (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.20: Hypothesis Test for a Difference in Two Population Means (2 of 2)

## Learning Objectives

- Under appropriate conditions, conduct a hypothesis test about a difference between two population means. State a conclusion in context.

On this page, we practice the hypothesis test for a difference in two population means (also called the two-sample t-test).

## Example

## Using Technology to Run the Hypothesis Test

*When dating someone, what matters more to you: looks or personality?* This question was the focus of a community college student's class project for an introductory statistics course. She devised a 25-point scale. An answer of 1 means "personality matters most and looks don't matter at all." A score of 25 means "looks matter most and personality does not matter at all." Her hypothesis is that the mean scores for males and females will differ, but she does not have an opinion about which population will have a higher mean score.

Here are her hypotheses.

$H_0$: $\mu_1 - \mu_2 = 0$
$H_a$: $\mu_1 - \mu_2 \neq 0$

We can also write the hypotheses as follows.

$H_0$: $\mu_1 = \mu_2$
$H_a$: $\mu_1 \neq \mu_2$

She chose a random sample of 10 classes from the schedule at Los Medanos College and distributed surveys in those classes. Survey respondents totaled 239 students: 150 females and 85 males.

We used her data to run a hypothesis test for a difference in two population means.

Here is the relevant output for our example:

```
> t.test(looks$Score~looks$Gender)

        Welch Two Sample t-test

data:  looks$Score by looks$Gender
t = -4.6574, df = 182.973, p-value = 6.143e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.695865 -1.496292
sample estimates:
mean in group Female   mean in group Male
            10.73333             13.32941
```

According to R, the P-value of this test is so small that it is essentially 0. How do we interpret this?

A P-value that is practically 0 means that it would be almost impossible to get data like that observed (or even more extreme) had the null hypothesis been true.

More specifically to our example, if there were no differences between females and males with respect to value they place on looks versus personality, it would be almost impossible (probability approximately 0) to get data where the difference between the sample means of females and males is -2.6 (that difference is 10.73 – 13.33 = -2.6) or higher.

## Learn By Doing

## Identify the P-value

Remember to use the printout of the results in the above example to answer the questions below.

https://assessments.lumenlearning.co...sessments/3707

https://assessments.lumenlearning.co...sessments/3708

https://assessments.lumenlearning.co...sessments/3783

https://assessments.lumenlearning.co...sessments/3784

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.20: Hypothesis Test for a Difference in Two Population Means (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.21: Estimating the Difference in Two Population Means

## Learning Objectives

- Construct a confidence interval to estimate a difference in two population means (when conditions are met). Interpret the confidence interval in context.

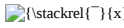
## Confidence Interval to Estimate $\mu_1 - \mu_2$

In a hypothesis test, when the sample evidence leads us to reject the null hypothesis, we conclude that the population means differ or that one is larger than the other. An obvious next question is *how much larger?* In practice, when the sample mean difference is statistically significant, our next step is often to calculate a confidence interval to estimate the size of the population mean difference.

The confidence interval gives us a range of reasonable values for the difference in population means $\mu_1 - \mu_2$. We call this the *two-sample T-interval* or the *confidence interval* to estimate a difference in two population means. The form of the confidence interval is similar to others we have seen.

$$\begin{array}{l}(\mathrm{sample}\text{ }\mathrm{statistic})\text{ }±\text{ }(\mathrm{margin}\text{ }\mathrm{of}\text{ }\mathrm{error})\\ (\mathrm{sample}\text{ }\mathrm{statistic})\text{ }±\text{ }(\mathrm{critical}\text{ }\mathrm{T\text{-}value})(\mathrm{standard}\text{ }\mathrm{error})\end{array}$$

**Sample Statistic**

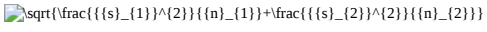
Since we're estimating the difference between two population means, the sample statistic is the difference between the means of the two independent samples:  .

**Critical T-Value**

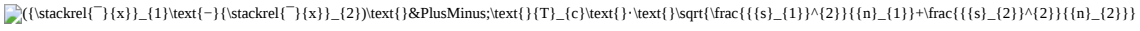
The critical T-value comes from the T-model, just as it did in "Estimating a Population Mean." Again, this value depends on the degrees of freedom (*df*). For two-sample T-test or two-sample T-intervals, the *df* value is based on a complicated formula that we do not cover in this course. We either give the *df* or use technology to find the *df*.

**Standard Error**

The estimated standard error for the two-sample T-interval is the same formula we used for the two-sample T-test. (As usual, $s_1$ and $s_2$ denote the sample standard deviations, and $n_1$ and $n_2$ denote the sample sizes.)



Putting all this together gives us the following formula for the two-sample T-interval.



**Conditions for Use**

The conditions for using this two-sample T-interval are the same as the conditions for using the two-sample T-test.

- The two *random* samples are *independent* and *representative.*
- The variable is *normally distributed in both populations*. If it is not known, *samples of more than 30* will have a difference in sample means that can be modeled adequately by the T-distribution. As we discussed in "Hypothesis Test for a Population Mean," T-procedures are robust even when the variable is not normally distributed in the population. If checking normality in the populations is impossible, then we look at the distribution in the samples. If a histogram or dotplot of the data does not show extreme skew or outliers, we take it as a sign that the variable is not heavily skewed in the populations, and we use the inference procedure.

## Example

## Confidence Interval for the "Calories and Context" Study

In the preceding few pages, we worked through a two-sample T-test for the "calories and context" example. In this example, we use the sample data to find a two-sample T-interval for $\mu_1 - \mu_2$ at the 95% confidence level.

**Recap of the Situation**

Population 1: Let $\mu_1$ be the mean number of calories purchased by women eating with other women.

- Population 1: Let $\mu_1$ be the mean number of calories purchased by women eating with other women.
- Population 2: Let $\mu_2$ be the mean number of calories purchased by women eating with men.

**Sample Statistics**

| | Size (n) | $\mathrm{Mean}\text{}(\stackrel{¯}{x})$ | SD (s) |
|---|---|---|---|
| Sample 1 | 45 | 850 | 252 |
| Sample 2 | 27 | 719 | 322 |

**Standard Error**

We found that the standard error of the sampling distribution of all sample differences is approximately 72.47.

$$\sqrt{\frac{{s}_{1}}^{2}}{{n}_{1}}+\frac{{{s}_{2}}^{2}}{{n}_{2}}}\text{}=\text{}\sqrt{\frac{{252}^{2}}{45}+\frac{{322}^{2}}{27}}\text{}\approx \text{}72.47$$

**Critical T-value**

For these two independent samples, *df* = 45. We find the critical T-value using the same simulation we used in "Estimating a Population Mean."

Reading from the simulation, we see that the critical T-value is 1.6790.

**Confidence Interval**

We can now put all this together to compute the confidence interval:

$$({\stackrel{¯}{x}}_{1}-{\stackrel{¯}{x}}_{2})\text{}\±\text{}{T}_{c}\text{}\cdot\text{}\mathrm{SE}\text{}=\text{}(850-719)\text{}\±\text{}(1.6790)(72.47)\text{}\approx \text{}131\text{}\±\text{}122$$

Expressing this as an interval gives us:

$$(\mathrm{9,\; 253})$$

**Interpretation**

We are 95% confident that the true value of $\mu_1 - \mu_2$ is between 9 and 253 calories. We can be more specific about the populations. We are 95% confident that at Indiana University of Pennsylvania, undergraduate women eating with women order between 9.32 and 252.68 more calories than undergraduate women eating with men.

In this next activity, we focus on interpreting confidence intervals and evaluating a statistics project conducted by students in an introductory statistics course.

## Learn By Doing

## Improving Children's Math Skills

Students in an introductory statistics course at Los Medanos College designed an experiment to study the impact of subliminal messages on improving children's math skills. The students were inspired by a similar study at City University of New York, as described in David Moore's textbook *The Basic Practice of Statistics(4th ed., W. H. Freeman, 2007)*. The participants were 11 children who attended an afterschool tutoring program at a local church. The children ranged in age from 8 to 11. All received tutoring in arithmetic skills. At the beginning of each tutoring session, the children watched a short video with a religious message that ended with a promotional message for the church.

The statistics students added a slide that said, "I work hard and I am good at math." This slide flashed quickly during the promotional message, so quickly that no one was aware of the slide. Children who attended the tutoring sessions on Mondays watched the video with the extra slide. Children who attended the tutoring sessions on Wednesday watched the video without the extra slide. The experiment lasted 4 weeks. The children took a pretest and posttest in arithmetic. Here are some of the results:

| | Math Test 1 (pretest) | | Math Test 2 (post-test) | | Improvement | |
|---|---|---|---|---|---|---|
| | Mean | Standard deviation | Mean | Standard deviation | Mean | Standard deviation |
| Treatment (n = 5) | 36 | 20.74 | 62 | 8.37 | 26 | 15.17 |
| Control (n = 6) | 40 | 20 | 56.67 | 19.66 | 16.67 | 13.66 |
| Overall (n=11) | 38.18 | 19.4 | 59.09 | 15.14 | 20.91 | 14.46 |

https://assessments.lumenlearning.co...sessments/3714

https://assessments.lumenlearning.co...sessments/3715

https://assessments.lumenlearning.co...sessments/3781

## Let's Summarize

Hypothesis tests and confidence intervals for two means can answer research questions about two populations or two treatments that involve quantitative data. In "Inference for a Difference between Population Means," we focused on studies that produced two independent samples. Previously, in "Hpyothesis Test for a Population Mean," we looked at matched-pairs studies in which individual data points in one sample are naturally paired with the individual data points in the other sample.

The hypotheses for two population means are similar to those for two population proportions.

The null hypothesis, $H_0$, is a statement of "no effect" or "no difference."

$H_0$: $\mu_1 - \mu_2 = 0$, which is the same as $H_0$: $\mu_1 = \mu_2$

The alternative hypothesis, $H_a$, takes one of the following three forms:

$H_a$: $\mu_1 - \mu_2 < 0$, which is the same as $H_a$: $\mu_1 < \mu_2$
$H_a$: $\mu_1 - \mu_2 > 0$, which is the same as $H_a$: $\mu_1 > \mu_2$
$H_a$: $\mu_1 - \mu_2 \neq 0$, which is the same as $H_a$: $\mu_1 \neq \mu_2$

As usual, how we collect the data determines whether we can use it in the inference procedure. We have our usual two requirements for data collection.

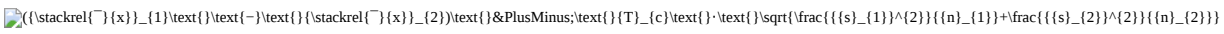
- Samples must be random in order to remove or minimize bias.
- Sample must be representative of the population in question.

We use the two-sample hypothesis test and confidence interval when the following conditions are met:

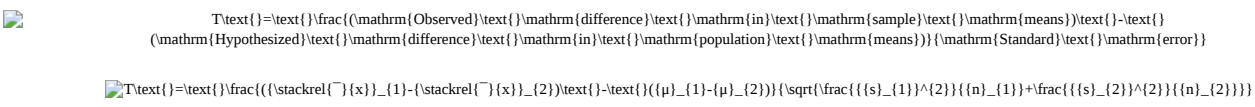
- The two random samples are independent.
- The variable is normally distributed in both populations. If this variable is not known, samples of more than 30 will have a difference in sample means that can be modeled adequately by the t-distribution. As we discussed in "Hypothesis Test for a Population Mean," t-procedures are robust even when the variable is not normally distributed in the population. Therefore, if checking normality in the populations is impossible, then we look at the distribution in the samples. If a histogram or dotplot of the data does not show extreme skew or outliers, we take it as a sign that the variable is not heavily skewed in the populations, and we use the inference procedure.

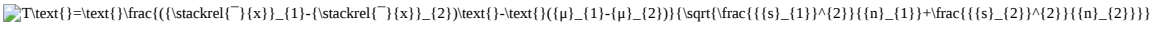
**Formulas:**

The confidence interval for $\mu_1 - \mu_2$ is




Hypothesis test for $H_0$: $\mu_1 - \mu_2 = 0$ is




We use technology to find the degrees of freedom to determine P-values and critical t-values for confidence intervals. (In most problems in this section, we provided the degrees of freedom for you.)

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.21: Estimating the Difference in Two Population Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.22: Putting It Together- Inference for Means

## Let's Summarize

The focus of this module, *Inference for Means*, is inference for a population mean or a difference between two populations means. We began this module with a discussion of the sampling distribution of sample means. We then developed a probability model based on this sampling distribution. We used the probability model with an actual sample mean to test a claim about population mean in a hypothesis test or to estimate a population mean with a confidence interval. We then moved to inference for a difference in two population means (or a treatment effect.)

## Sampling Distribution of Means

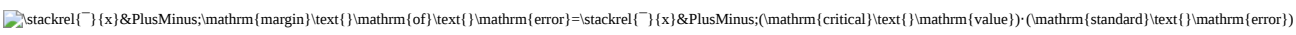
If we have a quantitative data set from a population with mean μ and standard deviation σ, the model for the theoretical sampling distribution of means of all random samples of size n has the following properties:

- The mean of the sampling distribution of means is μ.
- The standard deviation of the sampling distribution of means is .
  - Notice that as *n* grows, the standard error of the sampling distribution of means shrinks. That means that larger samples give more accurate estimates of a population mean.
- For large enough sample size, the sampling distribution of means is approximately normal (even if population is not normal). This is called the *central limit theorem*.
  - If a variable has a skewed distribution for individuals in the population, a larger sample size is needed to ensure that the sampling distribution has a normal shape.
  - The general rule is that if *n* is at least 30, then the sampling distribution of means will be approximately normal. However, if the population is already normal, then any sample size will produce a normal sampling distribution.
- We practiced finding a probability associated with a range of sample means, which is similar to finding a P-value in hypothesis testing. The process is as follows.
  - Convert a sample mean *X* into a *z*-score:
  - Use technology to find a probability associated with a given range of *z*-scores.

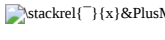
## Confidence Intervals

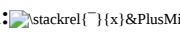
**The Form**

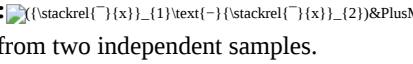
A confidence interval approximates a population mean by giving us a range of values that likely contains the population mean μ. The general form of the confidence interval is



We covered three different types of confidence intervals:

**One-sample Z-interval:**  , where σ is the population standard deviation (when it is known).

**One-sample T-interval:**  , where *s* is the sample standard deviation.

**Two-sample T-interval:**  , where we use the sample statistics from two independent samples.

**The T-Model**

When the standard deviation of the population is unknown, which is often the case, we use the T-model to find the critical values. When using the T-model to find critical values, we need to select an appropriate number of degrees of freedom.

- In the one-sample case, the number of degrees of freedom is 1 less than the sample size ($df = n - 1$).
- In the two-independent-sample case, the degrees of freedom come from a complicated formula, and we often use technology to find *df*.

**Conclusions**

To say we are 95% confident that the population mean falls within our confidence interval really means that about 95% of all confidence intervals computed in this way will capture the true population mean.

**Conditions**

The population must be normally distributed, or the sample size must be large enough (larger than 30). In the case of the two-sample T-interval, both populations/samples must meet these conditions. In practice, we use T-procedures with smaller samples if the distribution of the variable in the sample(s) is not heavily skewed and is without outliers. We take this as an indication that the variable has a fairly normal distribution in the population(s).

**Observations about Confidence Interval Structure**

- As we saw with other confidence intervals, the width of a confidence interval is twice the margin of error. The smaller the margin of error, the narrower the confidence interval and the more precise the estimate of the population parameter.
- Increasing the confidence level decreases the precision (larger margin of error, so wider interval). Decreasing the confidence level increases the precision (smaller margin of error, so narrower interval).
- Confidence intervals are useful estimates only when they provide a good balance of confidence level and precision. In order to increase precision without losing confidence, we must increase the sample size. In other words, larger samples provide more precise estimates without sacrificing confidence.

## Hypothesis Testing (Tests for Statistical Significance)

The process of any hypothesis test consists of four basic steps:

- Define the hypotheses
- Collect the data: We need random samples that are representative of the population. For the two-sample T-test, the samples must be independent.
- Assess the evidence: Assessment includes checking appropriate conditions, computing test statistics, and finding corresponding P-values.
- State the conclusion: We compare the P-value to $\alpha$, decide whether or not to reject $H_0$, then state conclusion in context.

**Hypotheses**

- **The null hypothesis ($H_0$):** The null hypothesis gives the value of the parameter we use to create the sampling distribution. In this way, the null hypothesis states what we assume to be true about the population.
- **The alternative hypothesis ($H_a$):** The alternative hypothesis usually reflects the claim in the research question about the value of the parameter. The alternative hypothesis says the parameter is greater than or less than or not equal to the value we assume to be true in the null hypothesis.
  - When $H_a$ is $\mu < \mu_0$ or $\mu > \mu_0$, the test is called a one-tailed test.
    - For the paired T-test, $H_0$ would look like $\mu < 0$ or $\mu > 0$ in the case of a one-tailed test.
    - For the two-sample T-test, $H_0$ would look like $\mu_1 - \mu_2 < 0$ or $\mu_1 - \mu_2 > 0$ in the case of a one-tailed test.
  - When $H_a$ is $\mu \neq \mu_0$, the test is called a two-tailed test.
    - For the paired T-test, $H_a$ would look like $\mu \neq 0$ in the case of a two-tailed test.
    - For the two-sample T-test, $H_a$ would look like $\mu_1 - \mu_2 \neq 0$ in the case of a two-tailed test.

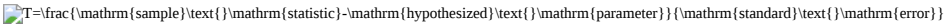
**Conditions**

Conditions that must be satisfied in order to carryout T-procedures are as follows:

- The population is normally distributed, *or* the sample is large (at least 30). This applies to both populations for the two-sample T-test.
- The samples must be random in order to avoid bias.
- The samples must be independent in the case of the two-sample T-test.

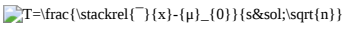
**Test Statistic**

The test T-statistic is given by

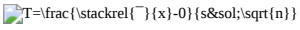


We've learned about three different types of T-tests:

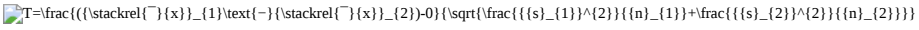
**One-sample T-test:**



**Paired T-test:** We calculate the differences, then find the mean and standard deviation.



**Two-sample T-test:**



**P-values**

The P-value is the probability of finding a random sample with a test statistic at least as extreme as ours, assuming that the null hypothesis is true. We find P-values by using the T-distribution.

To come to a conclusion about $H_0$, we compare the P-value to the significance level, $\alpha$.

- If $P \leq \alpha$, we reject $H_0$ and conclude there is significant evidence in favor of $H_a$.
- If $P > \alpha$, we fail to reject $H_0$ and conclude the sample does not provide significant evidence in favor of $H_a$.

**Error Types**

Hypothesis tests are based on random samples, so the conclusions are really statements about probabilities, and it is possible for the conclusions to be wrong.

- If our test results in rejecting a null hypothesis that is actually true, it is called a type I error.
- If our test results in failing to reject a null hypothesis that is actually false, it is called a type II error.

You are now ready to practice what you learned in this module by doing a StatTutor exercise. We design StatTutor exercises to help you apply what you have learned to a real-life data analysis question.

**Instructions:** One of the first few screens in StatTutor contains a link to download the data set for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click "Save," which allows you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you're using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you're using the TI Calculator.

If you are using StatCrunch, please see Additional Instructions for StatCrunch.

## Are You Ready for the Checkpoint?

If you completed all of the exercises in this module, you should be ready for the Checkpoint. To make sure that you are ready for the Checkpoint, use the My Response link below to evaluate your understanding of the learning objectives for this module and to submit questions that you may have.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.22: Putting It Together- Inference for Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.23: StatTutor- Analyzing Data From a Course's Grade Book

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=418

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.23: StatTutor- Analyzing Data From a Course's Grade Book is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.24: Assignment- Distribution of Sample Means

## Question 1:

Scores on the math portion of the SAT (SAT-M) in a recent year have followed a normal distribution with mean $\mu = 507$ and standard deviation $\sigma = 111$.

What is the probability that the mean SAT-M score of a random sample of 4 students who took the test that year is more than 600? Explain why you can solve this problem, even though the sample size ($n = 4$) is very low.

## Question 2:

Bags of a certain brand of potato chips say that the net weight of the contents is 35.6 grams. Assume that the standard deviation of the individual bag weights is 5.2 grams.

A quality control engineer selects a random sample of 35 bags. The mean weight of these 35 bags turns out to be 33.6 grams.

If the mean and standard deviation of individual bags is reported correctly, what is the probability that a random sample of 35 bags has a mean weight of 33.6 grams or less.

## Question 3:

Does the sample provide strong evidence that the mean weight of the bags is lower than the 35.6 grams listed on the package?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.24: Assignment- Distribution of Sample Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.25: Assignment- Connection between Confidence Intervals and Sampling Distributions

The purpose of this activity is to help give you a better understanding of the underlying **reasoning** behind the interpretation of confidence intervals. In particular, you will gain a deeper understanding of why we say that we are **"95% confident** that the population mean is **covered** by the interval."

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=422

When the simulation loads you will see a normal-shaped distribution, which represents the **sampling distribution of the mean** (x-bar) for random samples of a particular fixed sample size, from a population with a fixed standard deviation of σ.

The green line marks the value of the population mean, μ.

To begin the simulation, click the very top **"sample"** button at the topmost right of the simulation. You will see a line segment appear underneath the distribution; you should see that the line segment has a tiny red dot in the middle.

You have used the simulation to select a single sample from the population; the simulation has automatically computed the mean (x-bar) of your sample; your x-bar value is represented by the little red dot in the middle of the line segment. The line segment represents a confidence interval. Notice that, by default, the simulation used a **95%** confidence level.

### Question 1:

Did your 95% confidence interval contain (or "cover") the population mean $\mu$ (the green line)?

If your confidence interval *did* cover the population mean μ, then the simulation will have recorded 1 "hit" on the right side of the simulation.

Now, click to select another single sample.

### Question 2:

Was your second sample mean x-bar (the new red dot) the same value as your 1st sample mean? (i.e., is it in the same relative location along the axis?) Why is this result to be expected?

### Question 3:

A new 95% confidence interval has also been constructed (the new line segment, centered at the location of your second x-bar). Does the new interval cover the population mean $\mu$?

Notice, under "total" on the right side of the simulation, the number of total selected samples has been tallied.

Now click "**sample 50**" repeatedly until the simulation tallies a "total" of around 1,000 samples. You will see that the simulation computes the "percent hit" for all the intervals.

### Question 4:

What percentage of the many 95% confidence intervals should cover the population mean $\mu$?

### Question 5:

Now let's summarize some key ideas.

Based on what you've seen on the simulation (with the level set at 95%), decide which of the following statements are true and which are false.

1. Each interval is centered at the population mean (μ).
2. Each interval is centered at the sample mean (x-bar).
3. The population mean (μ) changes when different samples are selected.
4. The sample mean (x-bar) changes when different samples are selected.

5. In the long run, 95% of the intervals will contain (or “cover”) the **sample** mean (x-bar).
6. In the long run, 95% of the intervals will contain (or “cover”) the **population** mean (μ).

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 10.25: Assignment- Connection between Confidence Intervals and Sampling Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.26: Assignment- Hypothesis Testing for the Population Mean

The purpose of this activity is to give you guided practice in going through the process of a t-test for the population mean, and teach you how to carry out this test using statistical software.

**Background:**

A group of 75 college students from a certain liberal arts college were randomly sampled and asked about the number of alcoholic drinks they have in a typical week. The file containing the data is linked below. The purpose of this study was to compare the drinking habits of the students at the college to the drinking habits of college students in general. In particular, the dean of students, who initiated this study, would like to check whether the mean number of alcoholic drinks that students at his college have in a typical week differs from the mean of U.S. college students in general, which is estimated to be 4.73.

## Question 1:

Let μ be the mean number of alcoholic beverages that students in the college drink in a typical week. State the hypotheses that are being tested in this problem.

## Question 2:

Here is a histogram of the data. Can we safely use the t-test with this data?

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 3:

State the test statistic, interpret its value and show how it was found.

## Question 4:

Based on the P-value, draw your conclusions in context.

## Question 5:

What would your conclusions be if the dean of students suspected that the mean number of alcoholic drinks that students in the college consume in a typical week is **lower** than the mean of U.S. college students in general? In other words, if this were a test of the hypotheses:

$H_0$: μ = 4.73 drinks per week

$H_a$: μ < 4.73 drinks per week

## Question 6:

Now suppose that instead of the 75 students having been randomly selected from the entire student body, the 75 students had been randomly selected **only** from the engineering classes at the college (for the sake of convenience).

Address the following two issues regarding the effect of such a change in the study design:

a. Would we still be mathematically justified in using the T-test for obtaining conclusions, as we did previously?

b. Would the resulting conclusions still address the question of interest (which, remember, was to investigate the drinking habits of the students at the college as whole)?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.26: Assignment- Hypothesis Testing for the Population Mean is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.27: Assignment- Matched Pairs

The purpose of this activity is to give you guided practice in carrying out the paired t-test and to teach you how to obtain the paired t-test output using statistical software. Here is some background for the historically important data that we are going to work with in this activity.

## Background: Gosset's Seed Plot Data

William S. Gosset was employed by the Guinness brewing company of Dublin. Sample sizes available for experimentation in brewing were necessarily small, and new techniques for handling the resulting data were needed. Gosset consulted Karl Pearson (1857-1936) of University College in London, who told him that the current state of knowledge was unsatisfactory. Gosset undertook a course of study under Pearson, and the outcome of his study was perhaps the most famous paper in statistical literature, "The Probable Error of a Mean" (1908), which introduced the t distribution.

Since Gosset was contractually bound by Guinness, he published under a pseudonym, "Student"; hence, the t distribution is often referred to as *Student's t distribution*.

As an example to illustrate his analysis, Gosset reported in his paper on the results of seeding 11 different plots of land with two different types of seed: regular and kiln-dried. There is reason to believe that drying seeds before planting will increase plant yield. Since different plots of soil may be naturally more fertile, this confounding variable was eliminated by using the matched pairs design and planting both types of seed in all 11 plots.

The resulting data (corn yield in pounds per acre) are as follows:

We use these data to test the hypothesis that kiln-dried seed yields more corn than regular seed.

Because of the nature of the experimental design (matched pairs), we are testing the difference in yield.

Note that the differences were calculated: regular – kiln-dried.

## Question 1:

State the appropriate hypotheses that are being tested here. Be sure to define the parameter that you are using.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 2:

Are the conditions that allow me to safely use the paired T-test satisfied? Support your answer by using appropriate visual displays.

## Question 3:

Based on the visual display that you produced for answering the previous question, does it seem like there is some evidence in the data in favor of the alternative hypothesis? Explain.

## Question 4:

Carry out the paired t-test, state the test statistic and P-value, and state your conclusion in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.27: Assignment- Matched Pairs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.28: Assignment- Checking Conditions

The purpose of this activity is to give you guided practice in checking whether the conditions that allow us to use the two-sample t-test are met. (Recall that the two-sample t-test is another name for the hypothesis test for a difference in two population means.)

### Background

A researcher wanted to study whether or not men and women differ in the amount of time they watch TV during a week. In each of the following cases, you'll have to decide whether we can use the two-sample t-test to test this claim or not.

## Case 1

A random sample of 40 adults was chosen (22 of whom were women and 18 of whom were men). At the end of the week, each of the 40 subjects reported the total amount of time (in minutes) that he/she watched TV during that week.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 1:

Can we use the two-sample T-test to test this claim?

## Case 2

A random sample of 400 adults was chosen (191 women and 209 men). At the end of the week, each of the 400 subjects reported the total amount of time (in minutes) that he or she watched TV during that week.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 2:

Can we use the two-sample T-test to test this claim?

## Case 3

A random sample of 25 women and another random sample of 25 men was chosen. At the end of the week, each of the 50 subjects reported the total amount of time (in minutes) that he or she watched TV during that week.

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 3:

Can we use the two-sample T-test to test this claim?

## Case 4

A random sample of 50 married couples was chosen, which was split into a sample of 50 men and a sample of 50 women. At the end of the week, each of the 100 subjects reported the total amount of time (in minutes) that he or she watched TV during that week.

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 4:

Can we use the two-sample T-test to test this claim?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.28: Assignment- Checking Conditions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 10.29: Assignment- Two Independent Samples

The purpose of this activity is to give you guided practice in obtaining and interpreting a 95% confidence interval for $\mu_1 - \mu_2$ following a two-sample T-test that rejected $H_0$. Recall our second example:

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question:

Obtain the 95% confidence interval for $\mu_1 - \mu_2$ and interpret it in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 10.29: Assignment- Two Independent Samples is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 11: 11- Chi-Square Tests


11.1: Why It Matters- Chi-Square Tests
11.2: Introduction- Chi-Square Test for One-Way Tables
11.3: Goodness-of-Fit (1 of 2)
11.4: Goodness-of-Fit (2 of 2)
11.5: Introduction- Chi-Square Tests for Two-Way Tables
11.6: Test of Independence (1 of 3)
11.7: Test of Independence (2 of 3)
11.8: Test of Independence (3 of 3)
11.9: Test of Homogeneity
11.10: Putting It Together- Chi-Square Tests
11.11: StatTutor- Risk Factors for Low Birth Weight
11.12: Assignment- Test of Independence Using Technology
11.13: Assignment- Using Technology with Data to Run a Hypothesis Test


---


This page titled 11: 11- Chi-Square Tests is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.1: Why It Matters- Chi-Square Tests

In this module, *Chi-Square Tests*, we again focus on inference with categorical variables. We learn three new hypothesis tests, two of which are an extension of hypothesis tests about proportions that we learned in the modules *Inference for One Proportion* and *Inference for Two Proportions*. This module does not focus on estimating a parameter, so there is nothing about confidence intervals in this module.

Here is the Big Picture of Statistics with the new material for *Chi-Square Tests* highlighted in purple.

Following are examples of research questions that procedures in this module can address:

**Goodness-of-Fit Test:** Test a claim about the distribution of a categorical variable in a population.

- During the presidential election of 2008, the Pew Research Center collected survey data that suggested that 24% of registered voters were liberal, 38% were moderate, and 38% were conservative. Is the distribution of political views different this year?
- The distribution of blood types for whites in the United States is 45% type O, 41% type A, 10% type B, and 4% type AB. Is the distribution of blood types different for Asian Americans?

**Test of Independence:** Test a claim about the relationship between two categorical variables in a population.

- For young adults in the United States, is gender related to body image?
- Is alcohol abuse by New York firefighters dependent on participation in the 9/11 rescue operation?
- In the United States, is race associated with political views (conservative, moderate, liberal)?

**Test of Homogeneity:** Test a claim about the distribution of a categorical variable in several populations.

- Does the use of steroids in collegiate athletics differ across the three NCAA divisions?
- Was the distribution of political views (liberal, moderate, conservative) different for the last three presidential elections in the United States?

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.1: Why It Matters- Chi-Square Tests is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.2: Introduction- Chi-Square Test for One-Way Tables

## What you'll learn to do: Conduct a chi-square goodness-of-fit test. Interpret the conclusion in context.

### LEARNING OBJECTIVES

- Conduct a chi-square goodness-of-fit test. Interpret the conclusion in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.2: Introduction- Chi-Square Test for One-Way Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.3: Goodness-of-Fit (1 of 2)

## Learning Objectives

- Conduct a chi-square goodness-of-fit test. Interpret the conclusion in context.

In this section, we learn a new hypothesis test called the *chi-square goodness-of-fit test*. A goodness-of-fit test determines whether or not the distribution of a categorical variable in a sample fits a claimed distribution in the population.

We can answer the following research questions with a chi-square goodness-of-fit test:

- According to the manufacturer of M&M candy, the color distribution for plain chocolate M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, and 16% green. Do the M&Ms in our sample suggest that the color distribution is different?
- During the presidential election of 2008, the Pew Research Center collected survey data that suggested that 24% of registered voters were liberal, 38% were moderate, and 38% were conservative. Is the distribution of political views different this year?
- The distribution of blood types for whites in the United States is 45% type O, 41% type A, 10% type B, and 4% type AB. Is the distribution of blood types different for Asian Americans?

The null hypothesis states a specific distribution of proportions for each category of the variable in the population. The alternative hypothesis says that the distribution is different from that stated in the null hypothesis. To test our hypotheses, we select a random sample from the population and determine the distribution of the categorical variable in the data. Of course, we need a method for comparing the observed distribution in the sample to the expected distribution stated in the null hypothesis.

## Example

### Distribution of Color in Plain M&M Candies

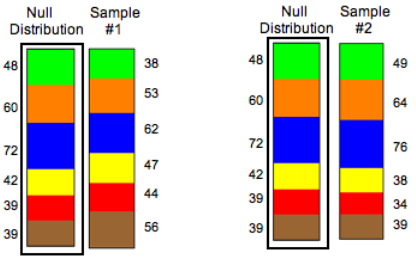
According to the manufacturer of M&M candy, the color distribution for plain chocolate M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green. This statement about the distribution of color in plain M&Ms is the null hypothesis. The alternative hypothesis says that this is not the distribution.

$H_0$: The color distribution for plain M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green.
$H_a$: The color distribution for plain M&Ms is different from the distribution stated in the null hypothesis.

We select a random sample of 300 plain M&M candies to test these hypotheses. If the sample has the distribution of color stated in the null hypothesis, then we expect 13% of the 300 to be brown, 13% of 300 to be red, 14% of 300 to be yellow, 24% of 300 to be blue, and so on. Here are the expected counts of each color for a sample of 300 candies:

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Expected | 0.13(300)=39 | 0.13(300)=39 | 0.14(300)=42 | 0.24(300)=72 | 0.20(300)=60 | 0.16(300)=48 |

Of course, the distribution of color will vary in different samples, so we need to develop a way to measure how far a sample distribution is from the null distribution, something analogous to a *z*-score or T-score. Before we discuss this new measure, let's look at two random samples selected from the null distribution to practice recognizing different amounts of variability. We can compare the distributions visually using ribbon charts.




Which random sample deviates the most from the null distribution? We address this question in the next activity.

## Observed Counts for Two Random Samples

Here are the observed counts for the two random samples shown above. This is the same information shown in the ribbon charts.

Random Sample #1 (n=300)

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 56 | 44 | 47 | 62 | 53 | 38 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |

Random Sample #2 (n=300)

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 39 | 34 | 38 | 76 | 64 | 49 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |

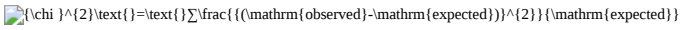
https://assessments.lumenlearning.co...sessments/3794

https://assessments.lumenlearning.co...sessments/3795

### Learn By Doing

https://assessments.lumenlearning.co...sessments/3796

Statisticians use the following formula to measure how far the observed data are from the null distribution. It is called the *chi-square test statistic*. The Greek letter chi is written χ.




**Notes about this formula:**

- Recall that the symbol ∑ means sum. Each category contributes a term to the sum, so the chi-square test statistic is based on the entire distribution. If the categorical variable has six categories, then the chi-square test statistic has six terms. If the categorical variable has three categories, then the chi-square test statistic has three terms, and so on.
- Notice that the difference "observed minus expected" for each category is part of the formula, but each difference is squared. This is necessary because the differences will add to 0, as we saw in the previous activity.
- Notice also that each squared difference is divided by the expected count for that category. The chi-square test statistic looks at the difference between the observed and expected counts *relative* to the size of the expected count.

### Example

## Calculating χ2

For Sample 1, the chi-square test statistic is approximately 12.94. For Sample 2, the chi-square test statistic is approximately 1.53. Usually, we use technology to calculate $\chi^2$, but here we show two calculations in detail to illustrate how the formula works. Notice that we are adding six terms. Each term represents the deviation for one color category.

Random Sample #1 (n=300)

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 56 | 44 | 47 | 62 | 53 | 38 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |

**Comment:** In Sample 1, notice that both blue and green observed counts deviate from the expected counts by 10 candies. But green contributes more to the chi-square test statistic. This makes sense because the chi-square test statistic measures relative difference. Relative to the expected count of 48 green candies, an absolute error of 10 is large. It is almost 20% of the expected count. (10/48 is about 0.20). The squared difference relative to the expected count is 100/48, about 2.08. Relative to the expected

count of 72 blue candies, an error of 10 candies is smaller. It is only about 14% of the expected count (10/72 is about 0.14). The squared difference relative to the expected count is 100/72, about 1.39.

Here is the chi-square calculation for Sample 2.

Random Sample #2 (n=300)

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 39 | 34 | 38 | 76 | 64 | 49 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |



\begin{array}
{1}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{brown}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{red}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{yellow}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{blue}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{orange}\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\text{ }\mathrm{green}\\ {\mathrm{\chi}}^{2}=\frac{{(39-39)}^{2}}{39}+\frac{{(34-39)}^{2}}{39}+\frac{{(38-42)}^{2}}{42}+\frac{{(76-72)}^{2}}{72}+\frac{{(64-60)}^{2}}{60}+\frac{{(49-48)}^{2}}{48}\\ \text{ }=\text{ }\frac{0}{39}\text{ }\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\text{ }\frac{16}{42}\text{ }\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\text{ }\frac{16}{72}\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\frac{16}{60}\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\frac{1}{48}=\\ \text{ }\text{ }\text{ }\text{ }\text{ }\approx
\text{ }\text{ }\text{ }0\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\text{ }0.64\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }0.38\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }0.22\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }0.27\text{ }\text{ }\text{ }\text{ }\text{ }+\text{ }\text{ }\text{ }\text{ }\text{ }0.02=1.53\text{ }\end{array}

## Calculating $\chi^2$

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3717

https://assessments.lumenlearning.co...sessments/3718

https://assessments.lumenlearning.co...sessments/3719

https://assessments.lumenlearning.co...sessments/3716

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 11.3: Goodness-of-Fit (1 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.4: Goodness-of-Fit (2 of 2)

## Learning Objectives

- Conduct a chi-square goodness-of-fit test. Interpret the conclusion in context.

Here we continue with the details of the chi-square goodness-of-fit hypothesis test. A goodness-of-fit test determines whether or not the distribution of a categorical variable in a sample fits a claimed distribution in the population. The chi-square test statistic is our measure of how much the sample distribution deviates from the population distribution.

As with other hypothesis tests, we need to be able to model the variability we expect in samples if the null hypothesis is true. Then we can determine whether the chi-square test statistic from the data is unusual or typical. An unusual $\chi^2$ value suggests that there are statistically significant differences between the sample data and the null distribution and provides evidence against the null hypothesis. This is the same logic we have been applying with hypothesis testing.

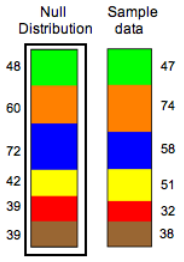
## Example

## Distribution of Color in Plain M&M Candies

Recall the claim made by the manufacturer of M&M candy: the color distribution for plain chocolate M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green. We used this distribution as our null hypothesis.

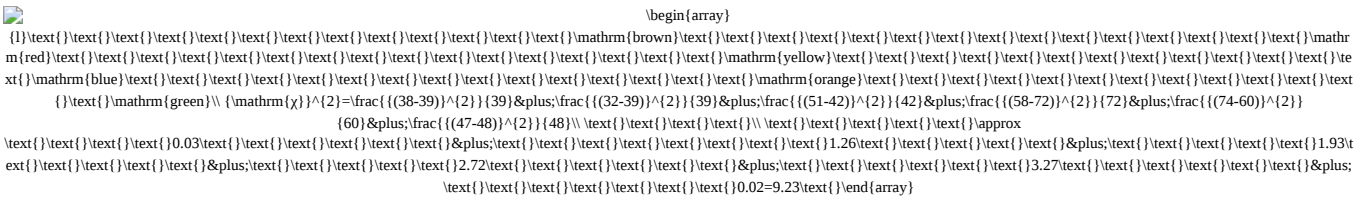
$H_0$: The color distribution for plain M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green.
$H_a$: The color distribution for plain M&Ms is different from the distribution stated in the null hypothesis.

Suppose we buy a large bag of plain M&M candies to test these hypotheses. We randomly select 300 from the bag and view this as a random sample from the population of all plain M&M candies. Our observed counts along with the expected counts are shown in the following ribbon chart and the table. Recall that the expected counts come from the null hypothesis.




We see that the sample distribution is very close to the null distribution for some colors and not others. The deviation appears largest for blue and orange. When we calculate the chi-square statistic, we see that these colors contribute the most to the chi-square value.

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 38 | 32 | 51 | 58 | 74 | 47 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |




What can we conclude? Is this chi-square value unusual or typical? To answer these questions, we must take many random samples from the population described by the null hypothesis. As we have done before, we use a simulation to take random samples. We do this in the next activity.

Learn By Doing

## Reasoning from the Chi-Square Sampling Distribution

Click here to open the simulation.

https://assessments.lumenlearning.co...sessments/3797

Learn By Doing

## Reasoning from the Chi-Square Sampling Distribution

Recall the distribution of political views for registered voters in 2008: 24% liberal, 38% moderate, and 38% conservative. We want to determine if the distribution is the same this year.

$H_0$: The distribution of political views this year is 0.24 liberal, 0.38 moderate, 0.38 conservative.
$H_a$: The distribution of political views this year differs from the 2008 distribution stated in the null hypothesis.

Previously, we used the data shown in the table to calculate the chi-square test statistic of 1.61.

| Political views | Liberal | Moderate | Conservative |
|---|---|---|---|
| Observed | 80 | 105 | 115 |
| Expected | 72 | 114 | 114 |

$$\chi^2 = \underset{\text{liberal}}{\frac{(80-72)^2}{72}} + \underset{\text{moderate}}{\frac{(105-114)^2}{114}} + \underset{\text{conservative}}{\frac{(115-114)^2}{114}}$$
$$\approx 0.89 + 0.71 + 0.01 = 1.61$$

What can we conclude?

Click here to open the simulation. Use this simulation to select at least 40 random samples from the null distribution.

Use the simulation below these next questions to select at least 40 random samples from the null distribution.

https://assessments.lumenlearning.co...sessments/3720

Now mark each conclusion valid or invalid.

https://assessments.lumenlearning.co...sessments/3721

https://assessments.lumenlearning.co...sessments/3722

https://assessments.lumenlearning.co...sessments/3723

https://assessments.lumenlearning.co...sessments/3724

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=440

In the previous activities, we based our conclusions on a relatively small number of random samples. If we continued taking random samples, the resulting distribution of chi-square statistics has a pattern that can be described by a mathematical model, called the *chi-square distribution*. As with other models for sampling distributions, this model is a probability model. The total area under the curve equals 1. We again use the area under the curve to represent the probability of sample results occurring if the null hypothesis is true. This means we again use the mathematical model with technology to find a P-value.

### The Chi-Square Distribution

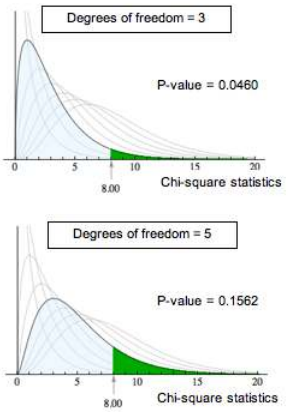
Unlike other sampling distributions we have studied, the chi-square model does not have a normal shape. It is skewed to the right. Like the T-model, the chi-square model is a family of curves that depend on degrees of freedom. For a chi-square goodness-of-fit test, the degrees of freedom is the number categories minus 1. (Sometimes this is written $(r - 1)$, where $r$ represents "rows" in the one-way table of observed counts.) The mean of the chi-square distribution is equal to the degrees of freedom.

A chi-square model is a good fit for the distribution of the chi-square test statistic only if the following conditions are met:

- The sample is randomly selected.
- All of the expected counts are 5 or greater.

If these conditions are met, we use the chi-square distribution to find the P-value. We use the same logic that we use in all hypothesis tests to draw a conclusion based on the P-value. If the P-value is at least as small as the significance level, we reject the null hypothesis and accept the alternative hypothesis.

The P-value is the likelihood that results from random samples have a $\chi^2$ value equal to or greater than that calculated from the data. As before, the P-value is a conditional probability based on the condition that the null hypothesis is true. For different degrees of freedom, the same $\chi^2$ value gives different P-values. For example, a chi-square value of 8 is statistically significant for $\alpha = 0.05$ with 3 degrees of freedom. This is not true for 5 degrees of freedom. As shown below, this is due to the change in the chi-square curve.




## Learn By Doing

## Hypothesis Test about the Color Distribution for Plain M&Ms

Recall the hypothesis test about the color distribution for plain M&Ms.

$H_0$: The color distribution for plain M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green.
$H_a$: The color distribution for plain M&Ms is different from the distribution stated in the null hypothesis.

From the null hypothesis, we determined the expected counts for a sample of 300. A random sample of 300 M&Ms gave the observed counts shown in the table. We calculated a chi-square statistic of 9.23.

| Color | Brown | Red | Yellow | Blue | Orange | Green |
|---|---|---|---|---|---|---|
| Observed | 38 | 32 | 51 | 58 | 74 | 47 |
| Expected | 39 | 39 | 42 | 72 | 60 | 48 |

Click here to open the simulation.

https://assessments.lumenlearning.co...sessments/3725

https://assessments.lumenlearning.co...sessments/3726

## Comment

Goodness-of-fit is an extension of the hypothesis test for one population proportion that we learned in *Inference for One Proportion*. Both of these hypothesis tests focus on a categorical variable in one population. In the hypothesis test for one population proportion, we focus on one category of the variable that we call “a success.” We make a claim about the proportion of “successes” in the population. For example, we previously investigated the claim that 20% of plain M&Ms are orange. In a chi-square goodness-of-fit test, we focus on the entire distribution of categories for the variable. So we investigate a claim that the

color distribution for plain M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green. The chi-square goodness-of-fit test does not give information about the deviation for specific categories. It gives a more general conclusion of "seems to fit the null distribution" or "does not fit the null distribution."

## Let's Summarize

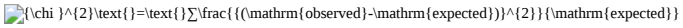
In "Chi-Square Test for One-Way Tables," we learned an inference procedure called the chi-square goodness-of-fit test. A goodness-of-fit test determines if the distribution of a categorical variable in a sample fits a claimed distribution in the population, or not.

We can answer the following research questions with a chi-square goodness-of-fit test:

- The distribution of blood types in the United States is 45% type O, 41% type A, 10% type B, and 4% type AB. Is the distribution of blood types the same in China?
- The Mars Company claims that 24% of M&M plain milk chocolate candies are blue, 13% brown, 16% green, 20% orange, 10% red, and 14% yellow. Do the M&Ms in our sample suggest that the color distribution is different?

## The Chi-Square Test Statistic and Distribution

The chi-square test statistic $\chi^2$ measures how far the observed data are from the null hypothesis by comparing observed counts and expected counts. Expected counts are the counts we expect to see if the null hypothesis is true.

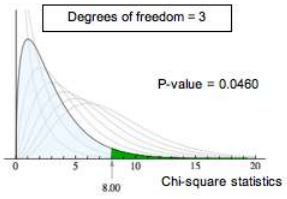


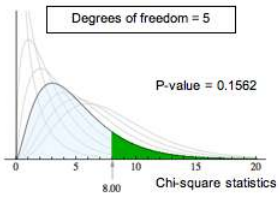

The chi-square model is a family of curves that depend on degrees of freedom. For a one-way table the degrees of freedom equals $(r - 1)$. All chi-square curves are skewed to the right with a mean equal to the degrees of freedom.

A chi-square model is a good fit for the distribution of the chi-square test statistic only if the following conditions are met:

- The sample is randomly selected.
- All expected counts are 5 or greater.

If these conditions are met, we use the chi-square distribution to find the P-value. We use the same logic that we use in all hypothesis tests to draw a conclusion based on the P-value. If the P-value is at least as small as the significance level, we reject the null hypothesis and accept the alternative hypothesis. The P-value is the likelihood that results from random samples have a $\chi^2$ value equal to or greater than that calculated from the data if the null hypothesis is true. For different degrees of freedom, the same $\chi^2$ value gives different P-values.




CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.4: Goodness-of-Fit (2 of 2) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.5: Introduction- Chi-Square Tests for Two-Way Tables

## What you'll learn to do: Conduct chi-square tests of independence and homogeneity.

### LEARNING OBJECTIVES

- Conduct a chi-square test of independence. Interpret the conclusion in context.
- Conduct a chi-square test of homogeneity. Interpret the conclusion in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.5: Introduction- Chi-Square Tests for Two-Way Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.6: Test of Independence (1 of 3)

## Learning Objectives

- Conduct a chi-square test of independence. Interpret the conclusion in context.

In this section, we learn two new hypothesis tests: a chi-square test of independence and a chi-square test of homogeneity. As the names imply, these two tests both use the same chi-square test statistic that we learned previously to compare observed and expected counts. In addition, P-values come from the same family of chi-square distributions. Therefore, much of what we learned in the previous section, "Chi-Square Test for One-Way Tables," will be useful here.

We begin with the chi-square test of independence. This test determines if there is a relationship between two categorical variables in the population. It is called a *test of independence* because "no relationship" means "independent." If there is a relationship between the two variables in the population, then they are dependent.

We can answer the following research questions with a chi-square test of independence:

- For young adults in the United States, is gender related to body image?
- Is alcohol abuse by New York firefighters dependent on participation in the 9/11 rescue operation?
- In the United States, is race associated with political views (conservative, moderate, liberal)?

The null hypothesis states that the two categorical variables are not related in the population. In other words, the variables are independent. The alternative hypothesis says that the two categorical variables are related in the population. In other words, the variables are dependent. To test our hypotheses, we select a random sample from the population and gather data on two categorical variables from each individual. As with all chi-square tests, the expected counts reflect the null hypothesis. So we need to determine what we expect to see in a sample if the variables are independent. As before, the chi-square test statistic measures the amount that the observed counts in the sample deviate from the expected counts.

## Example

## Gender and Body Image

What is your perception of your own body? Do you feel that you are overweight, underweight, or about right? A random sample of 1,200 U.S. college students answered this question as part of a larger survey. The following table shows part of the responses.

| Student | Gender | Body Image |
|---|---|---|
| Student 25 | M | Overweight |
| Student 26 | M | About right |
| Student 27 | F | Underweight |
| Student 28 | F | About right |
| Student 29 | M | About right |

Notice that the sample is a *random sample from a single population:* U.S. college students. Notice also that we collected data on *two categorical variables* for each student: gender and body image. This is the type of situation that is appropriate for a chi-square test of independence.

**Step 1: State the hypotheses.**

Here are two equivalent ways we can state the hypotheses for a test of independence.

$H_0$: There is no relationship between gender and body image for U.S. college students.
$H_a$: There is a relationship between gender and body image for U.S. college students.

We could also state the hypotheses like this:

$H_0$: Gender and body image are independent in the population of U.S. college students.
$H_a$: Gender and body image are dependent in the population of U.S. college students.

**Step 2: Collect and analyze the data.**

We summarized the data for this sample in a two-way table:

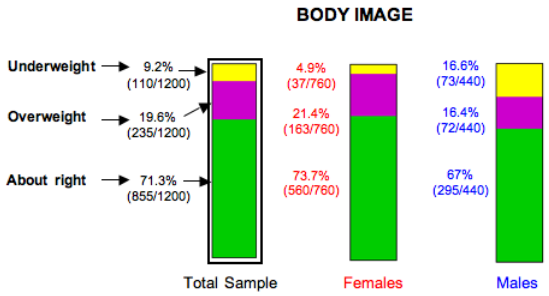
| Observed Counts | | | | |
|---|---|---|---|---|
| | About right | Overweight | Underweight | Row Totals |
| Female | 560 | 163 | 37 | 760 |
| Male | 295 | 72 | 73 | 440 |
| Column Totals | 855 | 235 | 110 | 1200 |

To investigate the relationship between gender and body image, we compare the percentage of males and females who gave each body image response. In *Relationships in Categorical Data with Intro to Probability*, we called these conditional percentages.

| Conditional Percents | | | | |
|---|---|---|---|---|
| | About right | Overweight | Underweight | Row Totals |
| Female | 560/760<br>73.68% | 163/760<br>21.45% | 37/760<br>4.87% | 760/760<br>100% |
| Male | 295/440<br>67.00% | 72/440<br>16.36% | 73/440<br>16.59% | 440/440<br>100% |
| Column Totals | 855/1200<br>71.25% | 235/1200<br>19.58% | 110/1200<br>9.17% | 1200/1200<br>100% |

**Comment:** In this situation, *gender* is the explanatory variable. *Body image* is the response variable. We compare the distribution of the response variable (body image) for the groups defined by the explanatory variable (males and females). This means that explanatory category totals are the denominator of each fraction. In this case, we use male and female totals in the denominators.

We graphed these conditional percentages using ribbon charts for a visual comparison. In this sample, we can see that a larger percentage of females (73.7%) than males (67%) perceive their body weight to be "about right." A larger percentage of females (21.4%) than males (16.4%) also perceive themselves to be overweight. Males are as likely to say they are overweight (16.4%) as underweight (16.6%), but females as much less likely to perceive themselves to be underweight (4.9%).




*What do these conditional percentages have to do with independence?*

Recall the definition of independence from *Probability and Probability Distribution*. Two events, A and B, are independent if the probability of A is the same as the probability of A when B has already occurred. We write this statement as $P(A) = P(A \mid B)$. In this context, if gender and body image are independent variables, then gender does not affect the probability that a student will give a specific answer to the body image question. We use relative frequencies from the sample to represent these probabilities.

For example, we would expect the following probabilities to be the same if gender and body image are independent variables.

$P$(about right) = 855/1,200 = 0.713.
$P$(about right | female) = 560/760 = 0.737, which is the conditional percent shown in the ribbon chart for females.
$P$(about right | male) = 295/440 = 0.67, also a conditional percent shown in the ribbon chart for males.

| Conditional Percents | | | | |
|---|---|---|---|---|
| | About right | Overweight | Underweight | Row Totals |
| Female | 560/760<br>73.68% | 163/760<br>21.45% | 37/760<br>4.87% | 760/760<br>100% |
| Male | 295/440<br>67.00% | 72/440<br>16.36% | 73/440<br>16.59% | 440/440<br>100% |
| Column Totals | 855/1200<br>71.25% | 235/1200<br>19.58% | 110/1200<br>9.17% | 1200/1200<br>100% |

If gender and body image are independent, then percents for a response category are the same.

Obviously, when we compare the conditional percentages for males and females in this sample, we see that they differ. But do they differ enough to conclude that variables are dependent? Or could these results have come from a population where gender and body image are independent? In which case, the differences are due to chance fluctuation that happens in random sampling. We need to conduct a test of independence to find out.

## Learn By Doing

### Alcoholism Risk in 9/11 Responders

Some firefighters and other first responders to the World Trade Center on September 11, 2001, have experienced symptoms of traumatic stress, depression, anxiety, and drinking problems. Cornell University researchers conducted a survey of a random sample of New York firefighters, some of whom had participated in the 9/11 rescue efforts. The report's title is "On the Front Line: The Work of First Responders in a Post-9/11 World." To see the report, click here. We use data from this report to investigate the question: *Are alcohol-related problems among New York firefighters associated with participation in the 9/11 rescue?*

https://assessments.lumenlearning.co...sessments/3804

https://assessments.lumenlearning.co...sessments/3805

https://assessments.lumenlearning.co...sessments/3806

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.6: Test of Independence (1 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.7: Test of Independence (2 of 3)

## Learning Objectives

- Conduct a chi-square test of independence. Interpret the conclusion in context.

Here we continue our chi-square test of independence for the variables *gender* and *body image* in the population of U.S. college students.

## Example

## Gender and Body Image Continued

**Step 1: State the hypotheses.**

Here are our hypotheses from the previous page:

$H_0$: There is no relationship between gender and body image for U.S. college students. (The variables are independent.)
$H_a$: There is a relationship between gender and body image for U.S. college students. (The variables are dependent.)

**Step 2: Collect and analyze the data.**

If the variables are independent, the percentage of males and females with a given response will be the same or at least close. Previously, we determined that in our sample, there are differences in the percentage of males and females who answer "about right," "overweight," or "underweight."

We need to determine if these differences are typical in random samples from a population where gender and body image are independent. Perhaps the differences we see in this sample are just fluctuations expected in random sampling. Or perhaps these differences are too large to be explained by chance. We will not know until we complete the hypothesis test.

**Step 3: Assess the evidence.**

We need to determine the expected values and the chi-square test statistic so that we can find the P-value.

*Calculating Expected Values for a Test of Independence*

Expected counts always describe what we expect to see in a sample if the null hypothesis is true. In this situation, if gender and body image are independent, then we expect the probability that a student answers "about right" in the sample to be the same probability that a male (or a female) student answers "about right" (similarly for "overweight" or "underweight" responses).

Here are the calculations of expected counts for the response "about right":

Probability that a student will answer "about right": $P$(about right) = (855/1,200) = 0.7125
Expected count of females in the sample who will answer "about right": 0.7125(760) = 541.5
Expected count of males in the sample who will answer "about right": 0.7125(440) = 313.5

Here are the calculations of expected counts for the response "overweight":

- Probability that a student will answer "overweight": $P$(overweight) = (235/1,200) = 0.1958
- Expected count of females in the sample who will answer "overweight": 0.1958(760) = 148.8
- Expected count of males in the sample who will answer "overweight": 0.1958(440) = 86.2

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3807

## Learn By Doing

https://assessments.lumenlearning.co...sessments/3808

## More on Gender and Body Image

*Checking Conditions*

The conditions for use of the chi-square distribution are the same as we learned previously:

- The sample is random.
- All of the expected counts are 5 or greater.

Since the data meets the conditions, we can proceed with calculating $\chi^2$ test statistic.

*Calculating the Chi-Square Test Statistic*

We calculate the chi-square test statistic as we did in "Chi-Square Test for One-Way Tables."

We will use technology to calculate the chi-square value. But for this sample, we will show the calculation.

*Finding Degrees of Freedom and the P-Value*

For a chi-square tests based on two-way tables, the degrees of freedom are

(number of explanatory categories – 1) × (number of response categories – 1)

You will also see this written: $(r - 1)(c - 1)$, where $r$ is the number of rows and $c$ is the number of columns in the two-way table (when we write the table without row and column totals). In this case the degrees of freedom are $(2 - 1)(3 - 1) = 2$.

We use the chi-square distribution with $df = 2$ to find the P-value. Note that the chi-square test statistic for this sample is so large that it is off the scale used in the simulation. So we conclude that the P-value is essentially zero.

**Step 4: Conclusion**

The relationship between gender and body image is statistically significant in this sample. We reject the null hypothesis and accept the alternative hypothesis. Gender and body image are dependent variables in the population of U.S. college students. (P-value is essentially 0.)

### Learn By Doing

## More on Gender and Body Piercing

A study was done on the relationship between gender and ear piercing among high-school students. A sample of 1,000 students was chosen, then classified according to both gender and whether or not they had either of their ears pierced. The following information is available:

| **Chi-Square Test: Pierced, No Pierced** | | | |
|---|---|---|---|
| | | | |
| **Observed Counts:** | | | |
| | | | |
| | Pierced | No Pierced | Total |
| Female | 576 | 64 | 640 |
| Male | 72 | 288 | 360 |
| Total | 648 | 352 | 1000 |
| | | | |
| **Expected Counts:** | | | |
| | | | |
| | Pierced | No Pierced | Total |
| Female | 414.72 | 225.28 | 640 |
| Male | 233.28 | 126.72 | 360 |
| Total | 648 | 352 | 1000 |
| | | | |
| **Chi-Square Contribution:** | | | |
| | | | |
| | Pierced | No Pierced | |
| Female | | 115.462 | |
| Male | 111.502 | 205.265 | |

https://assessments.lumenlearning.co...sessments/3727

https://assessments.lumenlearning.co...sessments/3728

https://assessments.lumenlearning.co...sessments/3729

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 11.7: Test of Independence (2 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.8: Test of Independence (3 of 3)

## Learning Objectives

- Conduct a chi-square test of independence. Interpret the conclusion in context.

On this page, we practice the chi-square test for independence in its entirety and learn how to use statistical software to conduct this test. We also investigate the effect of sample size on the chi-square test statistic.

## Learn By Doing

## A Real Court Case

In the early 1970s, a young man challenged an Oklahoma state law that prohibited the sale of 3.2% beer to males under age 21 but allowed its sale to females in the same age group. The case (*Craig v. Boren*, 429 U.S. 190, 1976) was ultimately heard by the U.S. Supreme Court. The state of Oklahoma argued that the law improved traffic safety. One of the three main pieces of data presented to the court was the result of a "random roadside survey." This survey gathered information on gender and whether or not the driver had been drinking alcohol in the previous 2 hours. A total of 619 drivers under 21 years of age were included in the survey.

Please click here to open the simulation for use in the following activity.

https://assessments.lumenlearning.co...sessments/3809

https://assessments.lumenlearning.co...sessments/3810

https://assessments.lumenlearning.co...sessments/3811

https://assessments.lumenlearning.co...sessments/3812

https://assessments.lumenlearning.co...sessments/3813

https://assessments.lumenlearning.co...sessments/3814

https://assessments.lumenlearning.co...sessments/3815

https://assessments.lumenlearning.co...sessments/3816

### Comment: The Effect of Sample Size on Chi-Square

With other hypothesis tests, we have seen that sample size can affect the P-value and our conclusion. This is also true for chi-square. To illustrate this idea, we multiplied all of the counts in the Oklahoma data by 3.

**ORIGINAL DATA**

| | Drank Alcohol in Last Two Hours? | | |
|---|---|---|---|
| | Yes | No | Totals |
| Male | 77 | 404 | 481 |
| Female | 16 | 122 | 138 |
| Totals | 93 | 526 | 619 |

**DATA x3**

| | Drank Alcohol in Last Two Hours? | | |
|---|---|---|---|
| | Yes | No | Totals |
| Male | 231 | 1212 | 1443 |
| Female | 48 | 366 | 414 |
| Totals | 279 | 1578 | 1857 |

**ORIGINAL DATA Conditional Percents**

| | Drank Alcohol in Last Two Hours? | | |
|---|---|---|---|
| | Yes | No | Totals |
| Male | 77/481<br>16.0% | 404/481<br>84.0% | 481 |
| Female | 16/138<br>11.6% | 122/138<br>88.4% | 138 |
| Totals | 93 | 526 | 619 |

**DATA x3 Conditional Percents**

| | Drank Alcohol in Last Two Hours? | | |
|---|---|---|---|
| | Yes | No | Totals |
| Male | 231/1443<br>16.0% | 1212/1443<br>84.0% | 1443 |
| Female | 48/414<br>11.6% | 366/414<br>88.4% | 414 |
| Totals | 279 | 1578 | 1857 |

Notice that the conditional percentages do not change, so the new "data" shows the same relationship between gender and drinking before driving. The probability that a driver under the age of 21 drinks alcohol before driving is still about 15.0% (279/1857). Males are still more likely to consume alcohol before driving (231/1443 = 16.0%) than are females (48/414 = 11.6%), with the same difference of 4.4% that we saw in the original data.

We used technology to find expected counts and the chi-square test statistic.

**ORIGINAL DATA**

| | Drank Alcohol Yes | Drank Alcohol No | Total |
|---|---|---|---|
| male | 77<br>72.27 | 404<br>408.7 | 481 |
| female | 16<br>20.73 | 122<br>117.3 | 138 |
| Total | 93 | 526 | 619 |

Chi-Square test:

| Statistic | DF | Value | P-value |
|---|---|---|---|
| Chi-square | 1 | 1.6365641 | 0.2008 |

**DATA X3**

| | Drank Alcohol Yes | Drank Alcohol No | Total |
|---|---|---|---|
| male | 231<br>216.8 | 1212<br>1226 | 1443 |
| female | 48<br>62.2 | 366<br>351.8 | 414 |
| Total | 279 | 1578 | 1857 |

Chi-Square test:

| Statistic | DF | Value | P-value |
|---|---|---|---|
| Chi-square | 1 | 4.9096923 | 0.0267 |

Notice that multiplying the observed counts by 3 also triples the expected counts and the chi-square value. This increase in the chi-square value gives a statistically significant P-value of 0.0267, which changes our conclusion. With this larger sample, the evidence is strong enough to reject the null hypothesis. We conclude that gender is associated with drinking alcohol before driving. The variables are dependent for drivers under the age of 21 in Oklahoma. With this sample size, the data provides evidence in support of the Oklahoma law that forbids sale of 3.2% beer to males and permits it to females with the goal of improving traffic safety.

**What's the point?** We see once again that sample size affects the P-value in a hypothesis test. This means that a small sample may not detect a relationship that exists between two categorical variables in a population. Conversely, a large sample may indicate that a relationship is statistically significant on the basis of differences in observed and expected counts that are not important in a practical sense.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.8: Test of Independence (3 of 3) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.9: Test of Homogeneity

## Learning Objectives

- Conduct a chi-square test of homogeneity. Interpret the conclusion in context.

We have learned the details for two chi-square tests, the goodness-of-fit test, and the test of independence. Now we focus on the third and last chi-square test that we will learn, the **test for homogeneity**. This test determines if two or more populations (or subgroups of a population) have the same distribution of a single categorical variable.

The test of homogeneity expands the test for a difference in two population proportions, which is the two-proportion Z-test we learned in *Inference for Two Proportions*. We use the two-proportion Z-test when the response variable has only two outcome categories and we are comparing two populations (or two subgroups.) We use the test of homogeneity if the response variable has two or more categories and we wish to compare two or more populations (or subgroups.)

We can answer the following research questions with a chi-square test of homogeneity:

- Does the use of steroids in collegiate athletics differ across the three NCAA divisions?
- Was the distribution of political views (liberal, moderate, conservative) different for last three presidential elections in the United States?

The null hypothesis states that the distribution of the categorical variable is the same for the populations (or subgroups). In other words, the proportion with a given response is the same in all of the populations, and this is true for all response categories. The alternative hypothesis says that the distributions differ.

Note: *Homogeneous* means the same in structure or composition. This test gets its name from the null hypothesis, where we claim that the distribution of the responses are the same (homogeneous) across groups.

To test our hypotheses, we select a random sample from each population and gather data on one categorical variable. As with all chi-square tests, the expected counts reflect the null hypothesis. We must determine what we expect to see in each sample if the distributions are identical. As before, the chi-square test statistic measures the amount that the observed counts in the samples deviate from the expected counts.

## Example

### Steroid Use in Collegiate Sports

In 2006, the NCAA published a report called “Substance Use: NCAA Study of Substance Use of College Student-Athletes.” We use data from this report to investigate the following question: *Does steroid use by student athletes differ for the three NCAA divisions?*

The data comes from a random selection of teams in each NCAA division. The sampling plan was somewhat complex, but we can view the data as though it came from a random sample of athletes in each division. The surveys are anonymous to encourage truthful responses.

To see the NCAA report on substance use, click here.

A note on NCAA divisions: The National Collegiate Athletic Association (NCAA) is divided into three divisions and oversees a wide range of collegiate sports. Division I schools have to sponsor more sports teams. These schools tend to be large universities with large athletic budgets supplemented by revenue from the games. They must offer athletic scholarships. Division II schools tend to be the smaller public universities and many private institutions. They have much smaller budgets that come solely from the college. The NCAA limits the amount Division II colleges can spend on athletic scholarships. Division III consists of colleges and universities that treat athletics as an extracurricular activity for students, instead of a source of revenue. These institutions do not offer athletic scholarships.

**Step 1: State the hypotheses.**

In the test of homogeneity, the null hypothesis says that the distribution of a categorical response variable is the same in each population. In this example, the categorical response variable is *steroid use* (yes or no). The populations are the three NCAA divisions.

$H_0$: The proportion of athletes using steroids is the same in each of the three NCAA divisions.

$H_0$: The proportion of athletes using steroids is the same in each of the three NCAA divisions.
$H_a$: The proportion of athletes using steroids is not same in each of the three NCAA divisions.

Note: These hypotheses imply that the proportion of athletes not using steroids is also the same in each of the three NCAA divisions, so we don't need to state this explicitly. For example, if 2% of the athletes in each division are using steroids, then 98% are not.

Here is an alternative way we could state the hypotheses for a test of homogeneity.

$H_0$: For each of the three NCAA divisions, the distribution of "yes" and "no" responses to the question about steroid use is the same.
$H_a$: The distribution of responses is not the same.

**Step 2: Collect and analyze the data.**

We summarized the data from these three samples in a two-way table.

We use percentages to compare the distributions of yes and no responses in the three samples. This step is similar to our data analysis for the test of independence.

We can see that Division I and Division II schools have essentially the same percentage of athletes who admit steroid use (about 1.2%). Not surprisingly, the least competitive division, Division III, has a slightly lower percentage (about 1.0%). Do these results suggest that the proportion of athletes using steroids is the same for the three divisions? Or is the difference seen in the sample of Division III schools large enough to suggest differences in the divisions? After all, the sample sizes are very large. We know that for large samples, a small difference can be statistically significant. Of course, we have to conduct the test of homogeneity to find out.

Note: We decided not to use ribbon charts for visual comparison of the three distributions because the percentage admitting steroid use is too small in each sample to be visible.

**Step 3: Assess the evidence.**

We need to determine the expected values and the chi-square test statistic so that we can find the P-value.

*Calculating Expected Values for a Test of Homogeneity*

Expected counts always describe what we expect to see in a sample if the null hypothesis is true. In this situation, we expect the percentage using steroids to be the same for each division. What percentage do we use? We find the percentage using steroids in the combined samples. This calculation is the same as we did when finding expected counts for a test of independence, though the logic of the calculation is subtly different.

Here are the calculations for the response "yes":

- Percentage using steroids in combined samples: 220/19,377 = 0.01135 = 1.135%

Expected count of steroid users for Division I is 1.135% of Division I sample:

- 0.01135(8,543) = 96.96

Expected count of steroid users for Division II is 1.135% of Division II sample:

- 0.01135(4,341) = 49.27

Expected count of steroid users for Division III is 1.135% of Division III sample:

- 0.01135(6,493) = 73.70

*Checking Conditions*

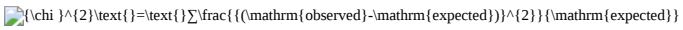
The conditions for use of the chi-square distribution are the same as we learned previously:

- A sample is randomly selected from each population.
- All of the expected counts are 5 or greater.

Since this data meets the conditions, we can proceed with calculating the $\chi^2$ test statistic.

*Calculating the Chi-Square Test Statistic*

There are no changes in the way we calculate the chi-square test statistic.



We use technology to calculate the chi-square value. For this example, we show the calculation. There are six terms, one for each cell in the 3 × 2 table. (We ignore the totals, as always.)

*Finding Degrees of Freedom and the P-Value*

For chi-square tests based on two-way tables (both the test of independence and the test of homogeneity), the degrees of freedom are $(r - 1)(c - 1)$, where $r$ is the number of rows and $c$ is the number of columns in the two-way table (not counting row and column totals). In this case, the degrees of freedom are $(3 - 1)(2 - 1) = 2$.

We use the chi-square distribution with $df = 2$ to find the P-value. The P-value is large (0.4561), so we fail to reject the null hypothesis.

**Step 4: Conclusion.**

The data does not provide strong enough evidence to conclude that steroid use differs in the three NCAA divisions (P-value = 0.4561).

### Learn By Doing

## First Use of Anabolic Steroids by NCAA Athletes

The NCAA survey includes this question: “When, if ever, did you start using anabolic steroids?” The response options are: have never used, before junior high, junior high, high school, freshman year of college, after freshman year of college. We focused on those who admitted use of steroids and compared the distribution of their responses for the years 1997, 2001, and 2005. (These are the years that the NCAA conducted the survey. Counts are estimates from reported percentages and sample size.) Recall that the NCAA uses random sampling in its sampling design.

Please click here to open the simulation for use in the following activity.

https://assessments.lumenlearning.co...sessments/3798

https://assessments.lumenlearning.co...sessments/3799

https://assessments.lumenlearning.co...sessments/3730

https://assessments.lumenlearning.co...sessments/3731

https://assessments.lumenlearning.co...sessments/3800

https://assessments.lumenlearning.co...sessments/3801

https://assessments.lumenlearning.co...sessments/3802

https://assessments.lumenlearning.co...sessments/3803

We now know the details for the chi-square test for homogeneity. We conclude with two activities that will give you practice recognizing when to use this test.

### Learn By Doing

## Gender and Politics

Consider these two situations:

A: Liberal, moderate, or conservative: Are there differences in political views of men and women in the United States? We survey a random sample of 100 U.S. men and 100 U.S. women.
B: Do you plan to vote in the next presidential election? We ask a random sample of 100 U.S. men and 100 U.S. women. We look for differences in the proportion of men and women planning to vote.

https://assessments.lumenlearning.co...sessments/3732

https://assessments.lumenlearning.co...sessments/3733

Learn By Doing

## Steroid Use for Male Athletes in NCAA Sports

We plan to compare steroid use for male athletes in NCAA baseball, basketball, and football. We design two different sampling plans.

A: Survey distinct random samples of NCAA athletes from each sport: 500 baseball players, 400 basketball players, 900 football players.
B. Survey a random sample of 1,800 NCAA male athletes and categorize players by sport and admitted steroid use. Responses are anonymous.

https://assessments.lumenlearning.co...sessments/3734

## Let's Summarize

In "Chi-Square Tests for Two-Way Tables," we discussed two different hypothesis tests using the chi-square test statistic:

- Test of independence for a two-way table
- Test of homogeneity for a two-way table

### Test of Independence for a Two-Way Table

- In the test of independence, we consider one population and two categorical variables.
- In *Probability and Probability Distribution*, we learned that two events are independent if $P(A|B) = P(A)$, but we did not pay attention to variability in the sample. With the chi-square test of independence, we have a method for deciding whether our observed $P(A|B)$ is "too far" from our observed $P(A)$ to infer independence in the population.
- The null hypothesis says the two variables are independent (or not associated). The alternative hypothesis says the two variables are dependent (or associated).
- To test our hypotheses, we select a single random sample and gather data for two different categorical variables.
- Example: Do men and women differ in their perception of their weight? Select a random sample of adults. Ask them two questions: (1) Are you male or female? (2) Do you feel that you are overweight, underweight, or about right in weight?

### Test of Homogeneity for a Two-Way Table

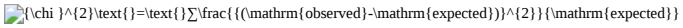
- In the test of homogeneity, we consider two or more populations (or two or more subgroups of a population) and a single categorical variable.
- The test of homogeneity expands on the test for a difference in two population proportions that we learned in *Inference for Two Proportions* by comparing the distribution of the categorical variable across multiple groups or populations.
- The null hypothesis says that the distribution of proportions for all categories is the same in each group or population. The alternative hypothesis says that the distributions differ.
- To test our hypotheses, we select a random sample from each population or subgroup independently. We gather data for one categorical variable.
- Example: Is the rate of steroid use different for different men's collegiate sports (baseball, basketball, football, tennis, track/field)? Randomly select a sample of athletes from each sport and ask them anonymously if they use steroids.

The difference between these two tests is subtle. They differ primarily in study design. In the test of independence, we select individuals at random from a population and record data for two categorical variables. The null hypothesis says that the variables are independent. In the test of homogeneity, we select random samples from each subgroup or population separately and collect data on a single categorical variable. The null hypothesis says that the distribution of the categorical variable is the same for each subgroup or population.

Both tests use the same chi-square test statistic.

### The Chi-Square Test Statistic and Distribution

For all chi-square tests, the chi-square test statistic $\chi^2$ is the same. It measures how far the observed data are from the null hypothesis by comparing observed counts and expected counts. *Expected counts* are the counts we expect to see if the null hypothesis is true.



The chi-square model is a family of curves that depend on degrees of freedom. For a two-way table, the degrees of freedom equals $(r - 1)(c - 1)$. All chi-square curves are skewed to the right with a mean equal to the degrees of freedom.

A chi-square model is a good fit for the distribution of the chi-square test statistic only if the following conditions are met:

- The sample is randomly selected.
- All expected counts are 5 or greater.

If these conditions are met, we use the chi-square distribution to find the P-value. We use the same logic that we have used in all hypothesis tests to draw a conclusion based on the P-value. If the P-value is at least as small as the significance level, we reject the null hypothesis and accept the alternative hypothesis. The P-value is the likelihood that results from random samples have a $\chi^2$ value equal to or greater than that calculated from the data if the null hypothesis is true.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.9: Test of Homogeneity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.10: Putting It Together- Chi-Square Tests

## Let's Summarize

In this module, *Chi-Square Tests*, we discussed three different hypothesis tests using the chi-square test statistic:

- Goodness-of-fit for a one-way table
- Test of independence for a two-way table
- Test of homogeneity for a two-way table

## Goodness-of-Fit test for a One-Way Table

- In a goodness-of-fit test, we consider one population and one categorical variable.
- The goodness-of-fit test expands the z-test for a population proportion that we learned in *Inference for One Proportion* by looking at the distribution of proportions for all categories defined by the categorical variable.
- The goodness-of-fit test determines whether a set of categorical data comes from a claimed distribution. The null hypothesis is that the proportion in each category in the population has a specific distribution. The alternative hypothesis says that the proportions in the population are not distributed as stated in the null hypothesis.
- To test our hypotheses, we select a random sample from the population and gather data for one categorical variable.

## Test of Independence for a Two-Way Table

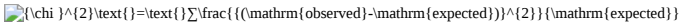
- In the test of independence, we consider one population and two categorical variables.
- In *Probability and Probability Distribution*, we learned that two events are independent if $P(A|B) = P(A)$, but we did not pay attention to variability in the sample. With the chi-square test of independence, we have a method for deciding whether our observed $P(A|B)$ is "too far" from our observed $P(A)$ to infer independence in the population.
- The null hypothesis says the two variables are independent (or not associated). The alternative hypothesis says the two variables are dependent (or associated).
- To test our hypotheses, we select a single random sample and gather data for two different categorical variables.

## Test of Homogeneity for a Two-Way Table

- In the test of homogeneity we consider two or more populations (or two or more subgroups of a population) and a single categorical variable.
- The test of homogeneity expands on the test for a difference in two population proportions that we learned in *Inference for Two Proportions* by comparing the distribution of the categorical variable across multiple groups or populations.
- The null hypothesis says that the distribution of proportions for all categories is the same in each group or population. The alternative hypothesis says that the distributions differ.
- To test our hypotheses, we select a random sample from each population or subgroup independently. We gather data for one categorical variable.

## The Chi-Square Test Statistic and Distribution

For all chi-square tests, the chi-square test statistic $\chi^2$ is the same. It measures how far the observed data are from the null hypothesis by comparing observed counts and expected counts. *Expected counts* are the counts we expect to see if the null hypothesis is true.




The chi-square model is a family of curves that depend on degrees of freedom. For a one-way table the degrees of freedom equals $(r - 1)$. For a two-way table, the degrees of freedom equals $(r - 1)(c - 1)$. All chi-square curves are skewed to the right with a mean equal to the degrees of freedom.

A chi-square model is a good fit for the distribution of the chi-square test statistic only if the following conditions are met:

- The sample is randomly selected.
- All expected counts are 5 or greater.

If these conditions are met, we use the chi-square distribution to find the P-value. We use the same logic that we have used in all hypothesis tests to draw a conclusion based on the P-value. If the P-value is at least as small as the significance level, we reject the null hypothesis and accept the alternative hypothesis. The P-value is the likelihood that results from random samples have a $\chi^2$ value equal to or greater than that calculated from the data if the null hypothesis is true.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 11.10: Putting It Together- Chi-Square Tests is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.11: StatTutor- Risk Factors for Low Birth Weight

You are now ready to practice what you learned in this module by doing a StatTutor exercise. We design StatTutor exercises to help you apply what you have learned to a real life data analysis question.

**Instructions:** One of the first few screens in StatTutor has a link to download the data set for this StatTutor exercise. When you click that link, a pop-up window will appear asking if you want to open or save the file. Make sure you click "Save," which will allow you to save the file to your hard drive. Then find the downloaded file and double-click it to open it if you're using R, Minitab, Excel, or StatCrunch, or transfer it to your calculator if you're using the TI Calculator.

An interactive or media element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/sss/?p=454

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.11: StatTutor- Risk Factors for Low Birth Weight is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.12: Assignment- Test of Independence Using Technology

The purpose of this activity is to gain experience conducting a chi-square test of independence using technology.

Recall the report On the Front Line: *The Work of First Responders in a Post-9/11 World*. We will use data from this report to investigate the question: Are alcohol-related problems among New York firefighters associated with participation in the 9/11 rescue?

Here again are our observed data:

## Question 1:

State the appropriate hypotheses for the chi-square test for independence in this case.

Now you will check whether the conditions for the chi-square test are met. In order to do this, you'll need to first launch the actual research report and read the last paragraph on page iii of the introduction (starting with the "The study was fully funded…")

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 2:

Does the data meet the conditions for the chi-square test?

### Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

## Question 3:

State your conclusion in context. Also explain what the P-value means as a conditional probability based on the null hypothesis.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 11.12: Assignment- Test of Independence Using Technology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 11.13: Assignment- Using Technology with Data to Run a Hypothesis Test

The purpose of this activity is to give you guided practice in carrying out the two-sample t-test, and to show you how to use software to aid in the process.

## Background

Do undergraduates sleep less than graduate students? A student conducted a study of sleep habits at a large state university. His hypothesis is that undergraduates will party more and sleep less than graduate students. He surveyed random samples of 75 undergraduate students and 50 graduate students. Subjects reported the hours they sleep in a typical night.

For this hypothesis test, he defined the population means as follows:

- $\mu_1$ is the mean number of hours undergraduate students sleep in a typical night.
- $\mu_2$ is the mean number of hours graduate students sleep in a typical night.

## Question 1:

State the null and alternative hypotheses that are being tested here.

## Question 2:

Explain why we can safely use the two-sample T-test in this case.

**Comment:** Before we move on to carry out the test, it is important to realize that in the two-sample problem, the data can be provided in three possible ways:

(i) Sample data in one column, and another column that indicates which sample the observation belongs to. Recall that this is the way the data were given in our leading example (looks vs. personality score and gender):

| Score (Y) | Gender (X) |
|---|---|
| 15 | Male |
| 13 | Female |
| 10 | Female |
| 12 | Male |
| 14 | Female |
| 14 | Male |
| 6 | Male |
| 17 | Male |
| ... | ... |

Note that essentially, one column contains the explanatory variable, and one contains the response.

(ii) Sample data in different columns—data from each of the two samples appear in a column dedicated to that category. As you'll see, this is the way the data are provided in this example:

| Undergraduate | Graduate |
|---|---|
| 6 | 8 |
| 5 | 5 |
| 3 | 6 |
| 6 | 6 |
| ... | ... |

(iii) Summarized data—we are not given the actual data, but just the data summaries: sample sizes, sample means and sample standard deviations of both samples. Recall that in our second example, the data were given in this format.

| | $n$ | $\bar{y}$ | $S$ |
|---|---|---|---|
| **20-29 yrs old** | 712 | 83.4 | 18.7 |
| **75+ yrs old** | 1001 | 78.5 | 19.0 |

## Instructions

Click on the link corresponding to your statistical package to see instructions for completing the activity, and then answer the questions below.

R | StatCrunch | Minitab | Excel 2007 | TI Calculator

### Question 3:

Carry out the test and report the test statistic and P-value.

### Question 4:

Draw your conclusions in context.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

---

This page titled 11.13: Assignment- Using Technology with Data to Run a Hypothesis Test is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 12: Main Body

12.1: Chapter 1

This page titled 12: Main Body is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 12.1: Chapter 1

This is the first chapter in the main body of the text. You can change the text, rename the chapter, add new chapters, and add new parts.

This page titled 12.1: Chapter 1 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 13: Course Information


13.1: Welcome to the class!

13.2: Course Learning Activities

13.3: How Your Course Grade is Determined

13.4: Heutagogical Course Objectives

13.5: What is Heutagogy?

13.6: My Expectations…


---


This page titled 13: Course Information is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.1: Welcome to the class!

This course will introduce you to the fascinating study of statistical analysis. As this course progresses you may develop a new way of thinking about the way things work – a perspective called “statistical thinking.” You will learn some new vocabulary, explore some new concepts, work a bit with numbers (data), and discuss what you are learning with your fellow students. Be prepared for a challenging and thought-provoking course. If you are the curious type, and if you like to think, I believe you will enjoy this course.

The “Course Information” documents in this section provide syllabus information which describe the course requirements and policies. Read all of these documents carefully. If you have any questions, contact me in the “Talk With the Professor” forum and I will get right back to you with a clarification. After reading through all of the Course Information documents you should have a clear picture of my expectations for the course. Feel free to print any of the documents if you are more comfortable with hard copies.

The course activities appear in the “Learning Modules” tab at the top of your screen. The Learning Modules contain the graded activities for the course. This is where you will “attend class” on a regular basis. You should logon and participate as often as you can. The Course Schedule document lists the start and end dates for the discussion forums and due dates for all assignments.

**Notice: If something in the course seems odd, or if you are not clear what you are expected to do, post a note in the Bulletin Board.**

http://onlinestatbook.com/2/

http://onlinestatbook.com/2/case_studies/case.html

http://onlinestatbook.com/2/glossary/index.html

http://onlinestatbook.com/2/calculat...lculators.html

CC licensed content, Original

- Welcome to the class!. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.1: Welcome to the class! is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.2: Course Learning Activities

## Course Learning Activities:

**Homework Problems:** There are assigned problems for each chapter.

**Discussion Forums:** There is a discussion forum for each module. This is the way you interact with the other students over the content of the textbook.

**Exams:** There will ba an exam in each module.

**Reflective Blogs:** You are asked to write about and discuss the most important things that you are learning in each module, and the impact the knowledge you are learning will have on your values, attitudes, beliefs behavior and your career.

Talk with the Professor: There is an ungraded "Talk with the Professor" Forum. In this area I may ask discussion questions about issues which I feel haven't been fully explored in the chapter forums. Also, in this area you may ask me questions, which I will respond to. Most often, I expect these questions (mine and yours) will be related to the discussions or the textbook – but no relevant topic is "off-limits." You should check this area each time you log on and participate in these discussion threads if and when they arise.

Extra Credit / Make-up Work / Incomplete Grades:

1. The requirements in this course require frequent participation and cooperation with the other students. There are no substitutes for these requirement, and I do not accept "extra credit" or "alternative credit" assignments.
2. Also, there is no way to "go back" after a discussion forum or blog has ended to "make-up" missed discussion activity. Posting to a discussion forum that has ended is like talking in an empty classroom – there is no benefit to you or to the class.
3. Finally, an incomplete in the course is not appropriate, as there is no way to complete the course once it has ended and all of the other students are gone.

CC licensed content, Original

- Course Learning Activities. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.2: Course Learning Activities is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.3: How Your Course Grade is Determined

**How you are evaluated:**

There are 30 graded activities in this course:

| Category Name | Course Value | Number of Assignments |
|---|---|---|
| Homework Problems | 25% | 15 homework assignments |
| Discussion Forums | 25% | 5 discussion forums |
| Exams | 25% | 5 exams |
| Reflective Blogs | 25% | 5 blogs |

Discussion Forum Grading Rubric:

To earn an "A+" on a discussion forum, you must submit 10 or more acceptable posts on 10 or more different days.
What is an "Acceptable Post?"

1. The subject (post title) must be a complete sentence which summarizes the main point of your post message.
2. Your post message must provide new, accurate, and relevant information. If your post is a question, it must be thoughtful and relevant.

## Course Final Grade Scale: Here are the cutoffs for final course grades:

A+ = 97, A = 94, A- = 90
B+ = 87, B = 84, B- = 80
C+ = 77, C = 74, C- = 70
D+ = 67, D = 64, D- = 60
F = 0-59

Grade Book: Online courses have special features that allow you to see your progress 24/7. Access is through the "MY Grades" link in the sidebar.

CC licensed content, Original

- How Your Course Grade is Determined. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.3: How Your Course Grade is Determined is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.4: Heutagogical Course Objectives

## Heutagogical Course Objectives specific to Prof Pelz' sections of this course:

### This course is designed to:

1. develop proficiency in self-directed learning from the required textbook readings.
2. develop the students' capability for connecting discipline content to personal values and behavior.

### Assessment of Heutagogical Course Objectives:

- Objective 1 is assessed via the textbook chapter discussions
- Objective 2 is assessed via the Knowledge Audit Reflective Blogs

CC licensed content, Original

- Heutagogical Course Objectives. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.4: Heutagogical Course Objectives is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.5: What is Heutagogy?

## What is “Heutagogy”?

The foundation of my educational philosophy (and therefore the basis of my instructional design decisions for this course) is “Heutagogy”. Simply stated, it means that adults learn best when they have a lot of control over what they learn and how they learn it. In addition, heutagogy asserts that the assessment of adult learning should focus on what the learner believes she/he has learned and on the various ways that learning has impacted, or will impact her/his values, ideals, and behavior.

(See this link for an excellent synopsis of heutagogy: http://www.nssa.us/journals/2007-28-1/2007-28-1-04.htm . Additional resources relevant to the theoretical basis of heutagogy are available at http://www.technoheutagogy.com .)

The learning objectives implicit in the heutagogical approach are different than those of traditional pedagogy and contemporary andragogy. I have designed this course to create a learner directed learning environment that will expose each student to the generally accepted concepts, ideas, research methods and research findings which comprise the discipline of Psychology. In addition, the design of this course will provide the opportunity for each student to identify and explore the discipline-specific topics and ideas from the textbook and from the World Wide Web that she/he finds most important and relevant.

---

I recognize that not all students have the same interests and motivations for taking this course. Hence, there are no content-specific learning goals that I have set. In contrast, each student decides which specific content topics to focus on and discuss. A desired outcome of this learner-directed learning environment is that each student will increase her/his capability to identify discipline-salient issues that are personally interesting and relevant and then orchestrate learning activities that result in cognitive growth and behavioral change.

In short – each learner sets her/his own content-specific learning objectives. This is the essence of a heutagogical design.

---

## My role in the course…

**The written assignments and discussion posts that you submit in this class are not for my benefit – they are for the benefit of you and the other students in the class. All of the course rules, policies, and requirements are designed to maximize the teaching / learning value of your coursework.**

**My role in the course begins with the instructional design process – to create a sequence of learning activities that:**

- **engages each learner in a meaningful way with the course content and the other learners in the class.**
- **requires each learner to also be a teacher – to add quality learning opportunities for the other students to benefit from.**
- **allows maximum freedom for self-direction – so that each learner has the responsibility to forge her/his own learning path.**

**The extent to which I am successful in achieving these three goals is the measure of my success in this course.**

CC licensed content, Original

- What is Heutagogy?. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.5: What is Heutagogy? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 13.6: My Expectations…

## My Expectations

The biggest difference between online courses and classroom courses is that Internet study is student-centered rather than teacher centered. This means that you – the student – are responsible for your own learning and success. If you are highly motivated, log on and participate frequently, and produce high quality work – you will be successful. However, if you log on sporadically, participate minimally, or submit poor quality work – you will not. Online higher education is aimed at independent learners. If you require the structure of a classroom, then online courses will not suit you.

In this course, we read and discuss the entire text! If you do not read the text in advance of the discussions, your lack of knowledge will be obvious to me and to the other students. It is not uncommon in these courses for students to provide strong, public criticism of other students who "waste their time" with ill-informed posts.

I expect the successful student to spend an **average** of about ***120-150 total hours*** on this course. Although that may seem like a lot – remember this: A traditional classroom-based course is designed to require 45 hours of "seat time" plus 2-3 hours outside of class for each hour in class. It adds up to around 150 total hours.

**Do you have a plan if you have a technical problem?** If your computer crashes, or if your Internet connection fails – these events do not excuse you from your course responsibilities.You can access this online course from any computer that has an Internet connection. I suggest that you make a plan now for events such as these. If you ever have a technical problem connecting to your course, submitting work to your course, or any other course-related issue, call the SLN Student Helpdesk at (800) 875-6269. If the problem is with the SLN system, you will be granted a time extension for submitting assignments affected by the problem. However, if you do not report the problem to SLN, no time extension will be granted.

**Final note:** You are responsible for keeping up with the requirements of this course. If you do not logon regularly, submit your assignments and discussion posts in a timely fashion, and follow the rules I have posted, it is very likely you will not succeed. However, I will not withdraw you from this course. If you decide not to finish for any reason, you must contact the Registrar's Office and officially withdraw yourself. If you just stop participating, you will receive the final course grade of "F".

CC licensed content, Original

- My Expectations.... **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 13.6: My Expectations… is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 14: Module 1


14.1: Module 1 Learning Activities
14.2: Module 1 Discussion Forum
14.3: Reflective Blog 1
14.4: Blog 1 Forum
14.5: Module 1 Homework
14.6: Module 1 Exam
14.7: Talk with the Professor


---


This page titled 14: Module 1 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.1: Module 1 Learning Activities

1. Read Chapters 1, 2 and 3 in the Open Statistics Education textbook. (go to the online textbook for interactives and videos.)
2. Read Chapters 1 and 2 in the SftSS textbook and complete all exercises.
3. Participate in the discussion forum
4. Submit the homework assignment
5. Submit the exam
6. Complete the Reflective Blog activities.

CC licensed content, Original

- Module 1 Learning Activities. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

---

This page titled 14.1: Module 1 Learning Activities is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.2: Module 1 Discussion Forum

Use this forum to:

- discuss the textbooks
- ask for assistance from your classmates
- introduce Internet resources you find helpful
- discuss any relevant issues

It is important – and required – that you interact with the other students in this class.

CC licensed content, Original

- Module 1 Discussion Forum. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.2: Module 1 Discussion Forum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.3: Reflective Blog 1

Reflect upon what you have learned in this module, and how you might use this knowledge in your day-to-day activities and/or in your career. Minimum requirement: 1000 original words.

CC licensed content, Original

- Reflective Blog 1. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.3: Reflective Blog 1 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.4: Blog 1 Forum

Post your Reflective Blog here, then respond to 5 other blogs. Minimum requirement: 200 original words per reply.

CC licensed content, Original

- Blog 1 Forum. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.4: Blog 1 Forum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.5: Module 1 Homework

Select 10 questions from the “Exercises” section of each chapter in the Online Statistics Education textbook. Do not select questions for which answers have been provided. Create a Word document (or use whatever word processing program you like) showing how you worked out the problems and submit it as a file attachment by the last day of the module.

CC licensed content, Original

- Module 1 Homework. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.5: Module 1 Homework is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.6: Module 1 Exam

Submit ample evidence that you understand the major concepts presented in this module.

CC licensed content, Original

- Module 1 Exam. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.6: Module 1 Exam is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 14.7: Talk with the Professor

This is a public forum where you interact with the course professor. If you wish to discuss a private matter (grades, etc.) use the course **Messages** feature.

CC licensed content, Original

- Talk with the Professor. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 14.7: Talk with the Professor is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# CHAPTER OVERVIEW

## 15: Modules 2, 3, 4 and 5

15.1: Structure of Modules 2 through 5

This page titled 15: Modules 2, 3, 4 and 5 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# 15.1: Structure of Modules 2 through 5

All modules folow the same structure and have the same learning activities and instructions as Module 1:

- Module Learning Activities – a list of required readings and graded assignments
- Folders for the textbook chapters
  - Online Statistics Education
  - Statistics for the Social Sciences
- Module discussion forum
- Reflective Blog written assignment
- Reflective Blog forum
- Module Homework assignment
- Module exam
- Talk with the Professor forum

CC licensed content, Original

- Structure of Modules 2 through 5. **Authored by**: WIlliam Pelz. **Provided by**: Herkimer College / SUNY. **Located at**: http://NA. **Project**: Statistics for the Social Sciences - Achieving the Dream course. **License**: *CC BY: Attribution*

This page titled 15.1: Structure of Modules 2 through 5 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

# Index

## D

dire

# Glossary

**Sample Word 1** | Sample Definition 1

# Detailed Licensing

## Overview

**Title:** Statistics for the Social Sciences (Pelz)

**Webpages:** 249

**All licenses found:**

- CC BY 4.0: 99.6% (248 pages)
- Undeclared: 0.4% (1 page)

## By Page


- Statistics for the Social Sciences (Pelz) - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY 4.0*
  - 1: Types of Statistical Studies and Producing Data - *CC BY 4.0*
    - 1.1: Why It Matters- Types of Statistical Studies and Producing Data - *CC BY 4.0*
    - 1.2: Introduction- Types of Statistical Studies - *CC BY 4.0*
    - 1.3: Types of Statistical Studies (1 of 4) - *CC BY 4.0*
    - 1.4: Types of Statistical Studies (2 of 4) - *CC BY 4.0*
    - 1.5: Types of Statistical Studies (4 of 4) - *CC BY 4.0*
    - 1.6: Types of Statistical Studies (3 of 4) - *CC BY 4.0*
    - 1.7: Introduction- Sampling - *CC BY 4.0*
    - 1.8: Sampling (1 of 2) - *CC BY 4.0*
    - 1.9: Sampling (2 of 2) - *CC BY 4.0*
    - 1.10: Introduction- Conducting Experiments - *CC BY 4.0*
    - 1.11: Conducting Experiments (1 of 2) - *CC BY 4.0*
    - 1.12: Conducting Experiments (2 of 2) - *CC BY 4.0*
    - 1.13: Putting It Together- Types of Statistical Studies and Producing Data - *CC BY 4.0*
  - 2: Summarizing Data Graphically and Numerically - *CC BY 4.0*
    - 2.1: Why It Matters- Summarizing Data Graphically and Numerically - *CC BY 4.0*
    - 2.2: Introduction- Categorical vs. Quantitative Data - *CC BY 4.0*
    - 2.3: Categorical vs. Quantitative Data - *CC BY 4.0*
    - 2.4: Introduction- Dotplots - *CC BY 4.0*
    - 2.5: Dotplots (1 of 2) - *CC BY 4.0*
    - 2.6: Dotplots (2 of 2) - *CC BY 4.0*
    - 2.7: Introduction- Histograms - *CC BY 4.0*
    - 2.8: Histograms (1 of 4) - *CC BY 4.0*
    - 2.9: Histograms (2 of 4) - *CC BY 4.0*
    - 2.10: Histograms (3 of 4) - *CC BY 4.0*
    - 2.11: Histograms (4 of 4) - *CC BY 4.0*
    - 2.12: Introduction- Measures of Center - *CC BY 4.0*
    - 2.13: Mean and Median (1 of 2) - *CC BY 4.0*
    - 2.14: Mean and Median (2 of 2) - *CC BY 4.0*
    - 2.15: Introduction- Measures of Spread - *CC BY 4.0*
    - 2.16: Interquartile Range and Boxplots (1 of 3) - *CC BY 4.0*
    - 2.17: Interquartile Range and Boxplots (2 of 3) - *CC BY 4.0*
    - 2.18: Interquartile Range and Boxplots (3 of 3) - *CC BY 4.0*
    - 2.19: Introduction- Describing a Distribution - *CC BY 4.0*
    - 2.20: Standard Deviation (1 of 4) - *CC BY 4.0*
    - 2.21: Standard Deviation (2 of 4) - *CC BY 4.0*
    - 2.22: Standard Deviation (3 of 4) - *CC BY 4.0*
    - 2.23: Standard Deviation (4 of 4) - *CC BY 4.0*
    - 2.24: Putting It Together- Summarizing Data Graphically and Numerically - *CC BY 4.0*
    - 2.25: StatTutor- Drinking Habits of College Students - *CC BY 4.0*
    - 2.26: Assignment- Histogram - *CC BY 4.0*
    - 2.27: Assignment- Five-Number Summary - *CC BY 4.0*
    - 2.28: Assignment- Boxplot - *CC BY 4.0*
    - 2.29: Assignment- Standard Deviation - *CC BY 4.0*
  - 3: Examining Relationships- Quantitative Data - *CC BY 4.0*
    - 3.1: Why It Matters- Examining Relationships- Quantitative Data - *CC BY 4.0*
    - 3.2: Introduction- Scatterplots - *CC BY 4.0*
    - 3.3: Scatterplots (1 of 5) - *CC BY 4.0*
    - 3.4: Scatterplots (2 of 5) - *CC BY 4.0*
    - 3.5: Scatterplots (3 of 5) - *CC BY 4.0*

- 3.6: Scatterplots (4 of 5) - *CC BY 4.0*
- 3.7: Scatterplots (5 of 5) - *CC BY 4.0*
- 3.8: Introduction- Linear Relationships - *CC BY 4.0*
- 3.9: Linear Relationships (1 of 4) - *CC BY 4.0*
- 3.10: Linear Relationships (2 of 4) - *CC BY 4.0*
- 3.11: Linear Relationships (3 of 4) - *CC BY 4.0*
- 3.12: Linear Relationships (4 of 4) - *CC BY 4.0*
- 3.13: Introduction- Association vs Causation - *CC BY 4.0*
- 3.14: Causation and Lurking Variables (1 of 2) - *CC BY 4.0*
- 3.15: Causation and Lurking Variables (2 of 2) - *CC BY 4.0*
- 3.16: Introduction- Linear Regression - *CC BY 4.0*
- 3.17: Linear Regression (1 of 4) - *CC BY 4.0*
- 3.18: Linear Regression (2 of 4) - *CC BY 4.0*
- 3.19: Linear Regression (3 of 4) - *CC BY 4.0*
- 3.20: Linear Regression (4 of 4) - *CC BY 4.0*
- 3.21: Introduction- Assessing the Fit of a Line - *CC BY 4.0*
- 3.22: Assessing the Fit of a Line (1 of 4) - *CC BY 4.0*
- 3.23: Assessing the Fit of a Line (2 of 4) - *CC BY 4.0*
- 3.24: Assessing the Fit of a Line (3 of 4) - *CC BY 4.0*
- 3.25: Assessing the Fit of a Line (4 of 4) - *CC BY 4.0*
- 3.26: Putting It Together- Examining Relationships- Quantitative Data - *CC BY 4.0*
- 3.27: StatTutor- Academic Performance - *CC BY 4.0*
- 3.28: Assignment- Scatterplot - *CC BY 4.0*
- 3.29: Assignment- Linear Relationships - *CC BY 4.0*
- 3.30: Assignment- Linear Regression - *CC BY 4.0*

4: Nonlinear Models - *CC BY 4.0*

- 4.1: Why It Matters- Nonlinear Models - *CC BY 4.0*
- 4.2: Introduction- Exponential Relationships - *CC BY 4.0*
- 4.3: Exponential Relationships (1 of 6) - *CC BY 4.0*
- 4.4: Exponential Relationships (2 of 6) - *CC BY 4.0*
- 4.5: Exponential Relationships (3 of 6) - *CC BY 4.0*
- 4.6: Exponential Relationships (4 of 6) - *CC BY 4.0*
- 4.7: Exponential Relationships (5 of 6) - *CC BY 4.0*
- 4.8: Exponential Relationships (6 of 6) - *CC BY 4.0*
- 4.9: Putting It Together- Nonlinear Models - *CC BY 4.0*

5: Relationships in Categorical Data with Intro to Probability - *CC BY 4.0*

- 5.1: Why It Matters- Relationships in Categorical Data with Intro to Probability - *CC BY 4.0*
- 5.2: Introduction- Two-Way Tables - *CC BY 4.0*
- 5.3: Two-Way Tables (1 of 5) - *CC BY 4.0*
- 5.4: Two-Way Tables (2 of 5) - *CC BY 4.0*
- 5.5: Two-Way Tables (3 of 5) - *CC BY 4.0*
- 5.6: Two-Way Tables (4 of 5) - *CC BY 4.0*
- 5.7: Two-Way Tables (5 of 5) - *CC BY 4.0*
- 5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability - *CC BY 4.0*
- 5.9: StatTutor- Treating Depression- A Randomized Clinical Trial - *CC BY 4.0*

6: Probability and Probability Distributions - *CC BY 4.0*

- 6.1: Why It Matters- Probability and Probability Distributions - *CC BY 4.0*
- 6.2: Introduction- Another Look at Probability - *CC BY 4.0*
- 6.3: Another Look at Probability (1 of 2) - *CC BY 4.0*
- 6.4: Another Look at Probability (2 of 2) - *CC BY 4.0*
- 6.5: Introduction- Probability Rules - *CC BY 4.0*
- 6.6: Probability Rules (1 of 3) - *CC BY 4.0*
- 6.7: Probability Rules (2 of 3) - *CC BY 4.0*
- 6.8: Probability Rules (3 of 3) - *CC BY 4.0*
- 6.9: Introduction- Discrete Probability Distribution - *CC BY 4.0*
- 6.10: Discrete Random Variables (1 of 5) - *CC BY 4.0*
- 6.11: Discrete Random Variables (2 of 5) - *CC BY 4.0*
- 6.12: Discrete Random Variables (3 of 5) - *CC BY 4.0*
- 6.13: Discrete Random Variables (4 of 5) - *CC BY 4.0*
- 6.14: Discrete Random Variables (5 of 5) - *CC BY 4.0*
- 6.15: Introduction- Continuous Probability Distribution - *CC BY 4.0*
- 6.16: Continuous Probability Distribution (1 of 2) - *CC BY 4.0*
- 6.17: Continuous Probability Distribution (2 of 2) - *CC BY 4.0*
- 6.18: Introduction- Normal Random Variables - *CC BY 4.0*
- 6.19: Normal Random Variables (1 of 6) - *CC BY 4.0*
- 6.20: Normal Random Variables (2 of 6) - *CC BY 4.0*
- 6.21: Normal Random Variables (3 of 6) - *CC BY 4.0*
- 6.22: Normal Random Variables (4 of 6) - *CC BY 4.0*
- 6.23: Normal Random Variables (5 of 6) - *CC BY 4.0*
- 6.24: Normal Random Variables (6 of 6) - *CC BY 4.0*
- 6.25: Putting It Together- Probability and Probability Distribution - *CC BY 4.0*

7: Linking Probability to Statistical Inference - *CC BY 4.0*

- 7.1: Why It Matters- Linking Probability to Statistical Inference - *CC BY 4.0*
- 7.2: Parameters vs. Statistics - *CC BY 4.0*
- 7.3: Introduction- Statistical Inference - *CC BY 4.0*
- 7.4: Distribution of Sample Proportions (1 of 6) - *CC BY 4.0*

- 7.5: Distribution of Sample Proportions (2 of 6) - *CC BY 4.0*
- 7.6: Distribution of Sample Proportions (3 of 6) - *CC BY 4.0*
- 7.7: Distribution of Sample Proportions (4 of 6) - *CC BY 4.0*
- 7.8: Distribution of Sample Proportions (5 of 6) - *CC BY 4.0*
- 7.9: Distribution of Sample Proportions (6 of 6) - *CC BY 4.0*
- 7.10: Introduction- Distribution of Sample Proportions - *CC BY 4.0*
- 7.11: Statistical Inference (1 of 3) - *CC BY 4.0*
- 7.12: Statistical Inference (2 of 3) - *CC BY 4.0*
- 7.13: Statistical Inference (3 of 3) - *CC BY 4.0*
- 7.14: Putting It Together- Linking Probability to Statistical Inference - *CC BY 4.0*

8: Inference for One Proportion - *CC BY 4.0*

- 8.1: Why It Matters- Inference for One Proportion - *CC BY 4.0*
- 8.2: Introduction- Estimating a Population Proportion - *CC BY 4.0*
- 8.3: Estimating a Population Proportion (1 of 3) - *CC BY 4.0*
- 8.4: Estimating a Population Proportion (2 of 3) - *CC BY 4.0*
- 8.5: Estimating a Population Proportion (3 of 3) - *CC BY 4.0*
- 8.6: Introduction- Hypothesis Testing - *CC BY 4.0*
- 8.7: Hypothesis Testing (1 of 5) - *CC BY 4.0*
- 8.8: Hypothesis Testing (2 of 5) - *CC BY 4.0*
- 8.9: Hypothesis Testing (4 of 5) - *CC BY 4.0*
- 8.10: Hypothesis Testing (3 of 5) - *CC BY 4.0*
- 8.11: Hypothesis Testing (5 of 5) - *CC BY 4.0*
- 8.12: Introduction- Hypothesis Test for a Population Proportion - *CC BY 4.0*
- 8.13: Hypothesis Test for a Population Proportion (1 of 3) - *CC BY 4.0*
- 8.14: Hypothesis Test for a Population Proportion (2 of 3) - *CC BY 4.0*
- 8.15: Hypothesis Test for a Population Proportion (3 of 3) - *CC BY 4.0*
- 8.16: Putting It Together- Inference for One Proportion - *CC BY 4.0*
- 8.17: StatTutor- Cell Phones - *CC BY 4.0*
- 8.18: Assignment- Hypothesis Testing for the Population Proportion p - *CC BY 4.0*

9: Inference for Two Proportions - *CC BY 4.0*

- 9.1: Why It Matters- Inference for Two Proportions - *CC BY 4.0*
- 9.2: Introduction- Distribution of Differences in Sample Proportions - *CC BY 4.0*
- 9.3: Distribution of Differences in Sample Proportions (1 of 5) - *CC BY 4.0*
- 9.4: Distribution of Differences in Sample Proportions (2 of 5) - *CC BY 4.0*
- 9.5: Distribution of Differences in Sample Proportions (3 of 5) - *CC BY 4.0*
- 9.6: Distribution of Differences in Sample Proportions (4 of 5) - *CC BY 4.0*
- 9.7: Distribution of Differences in Sample Proportions (5 of 5) - *CC BY 4.0*
- 9.8: Introduction- Estimate the Difference Between Population Proportions - *CC BY 4.0*
- 9.9: Estimate the Difference between Population Proportions (1 of 3) - *CC BY 4.0*
- 9.10: Estimate the Difference between Population Proportions (2 of 3) - *CC BY 4.0*
- 9.11: Estimate the Difference between Population Proportions (3 of 3) - *CC BY 4.0*
- 9.12: Introduction- Hypothesis Test for Difference in Two Population Proportions - *CC BY 4.0*
- 9.13: Hypothesis Test for Difference in Two Population Proportions (1 of 6) - *CC BY 4.0*
- 9.14: Hypothesis Test for Difference in Two Population Proportions (2 of 6) - *CC BY 4.0*
- 9.15: Hypothesis Test for Difference in Two Population Proportions (3 of 6) - *CC BY 4.0*
- 9.16: Hypothesis Test for Difference in Two Population Proportions (4 of 6) - *CC BY 4.0*
- 9.17: Hypothesis Test for Difference in Two Population Proportions (5 of 6) - *CC BY 4.0*
- 9.18: Hypothesis Test for Difference in Two Population Proportions (6 of 6) - *CC BY 4.0*
- 9.19: Putting It Together- Inference for Two Proportions - *CC BY 4.0*
- 9.20: Assignment- A Statistical Investigation using Software - *CC BY 4.0*

10: Inference for Means - *CC BY 4.0*

- 10.1: Why It Matters- Inference for Means - *CC BY 4.0*
- 10.2: Introduction- Distribution of Sample Means - *CC BY 4.0*
- 10.3: Distribution of Sample Means (1 of 4) - *CC BY 4.0*
- 10.4: Distribution of Sample Means (2 of 4) - *CC BY 4.0*

- 10.5: Distribution of Sample Means (3 of 4) - *CC BY 4.0*
- 10.6: Distribution of Sample Means (4 of 4) - *CC BY 4.0*
- 10.7: Introduction- Estimating a Population Mean - *CC BY 4.0*
- 10.8: Estimating a Population Mean (1 of 3) - *CC BY 4.0*
- 10.9: Estimating a Population Mean (2 of 3) - *CC BY 4.0*
- 10.10: Estimating a Population Mean (3 of 3) - *CC BY 4.0*
- 10.11: Introduction- Hypothesis Test for a Population Mean - *CC BY 4.0*
- 10.12: Hypothesis Test for a Population Mean (1 of 5) - *CC BY 4.0*
- 10.13: Hypothesis Test for a Population Mean (2 of 5) - *CC BY 4.0*
- 10.14: Hypothesis Test for a Population Mean (3 of 5) - *CC BY 4.0*
- 10.15: Hypothesis Test for a Population Mean (4 of 5) - *CC BY 4.0*
- 10.16: Hypothesis Test for a Population Mean (5 of 5) - *CC BY 4.0*
- 10.17: Introduction- Inference for a Difference in Two Population Means - *CC BY 4.0*
- 10.18: Inference for a Difference in Two Population Means - *CC BY 4.0*
- 10.19: Hypothesis Test for a Difference in Two Population Means (1 of 2) - *CC BY 4.0*
- 10.20: Hypothesis Test for a Difference in Two Population Means (2 of 2) - *CC BY 4.0*
- 10.21: Estimating the Difference in Two Population Means - *CC BY 4.0*
- 10.22: Putting It Together- Inference for Means - *CC BY 4.0*
- 10.23: StatTutor- Analyzing Data From a Course's Grade Book - *CC BY 4.0*
- 10.24: Assignment- Distribution of Sample Means - *CC BY 4.0*
- 10.25: Assignment- Connection between Confidence Intervals and Sampling Distributions - *CC BY 4.0*
- 10.26: Assignment- Hypothesis Testing for the Population Mean - *CC BY 4.0*
- 10.27: Assignment- Matched Pairs - *CC BY 4.0*
- 10.28: Assignment- Checking Conditions - *CC BY 4.0*
- 10.29: Assignment- Two Independent Samples - *CC BY 4.0*

- 11: 11- Chi-Square Tests - *CC BY 4.0*
  - 11.1: Why It Matters- Chi-Square Tests - *CC BY 4.0*
  - 11.2: Introduction- Chi-Square Test for One-Way Tables - *CC BY 4.0*
  - 11.3: Goodness-of-Fit (1 of 2) - *CC BY 4.0*
  - 11.4: Goodness-of-Fit (2 of 2) - *CC BY 4.0*
  - 11.5: Introduction- Chi-Square Tests for Two-Way Tables - *CC BY 4.0*
  - 11.6: Test of Independence (1 of 3) - *CC BY 4.0*
  - 11.7: Test of Independence (2 of 3) - *CC BY 4.0*
  - 11.8: Test of Independence (3 of 3) - *CC BY 4.0*
  - 11.9: Test of Homogeneity - *CC BY 4.0*
  - 11.10: Putting It Together- Chi-Square Tests - *CC BY 4.0*
  - 11.11: StatTutor- Risk Factors for Low Birth Weight - *CC BY 4.0*
  - 11.12: Assignment- Test of Independence Using Technology - *CC BY 4.0*
  - 11.13: Assignment- Using Technology with Data to Run a Hypothesis Test - *CC BY 4.0*
- 12: Main Body - *CC BY 4.0*
  - 12.1: Chapter 1 - *CC BY 4.0*
- 13: Course Information - *CC BY 4.0*
  - 13.1: Welcome to the class! - *CC BY 4.0*
  - 13.2: Course Learning Activities - *CC BY 4.0*
  - 13.3: How Your Course Grade is Determined - *CC BY 4.0*
  - 13.4: Heutagogical Course Objectives - *CC BY 4.0*
  - 13.5: What is Heutagogy? - *CC BY 4.0*
  - 13.6: My Expectations… - *CC BY 4.0*
- 14: Module 1 - *CC BY 4.0*
  - 14.1: Module 1 Learning Activities - *CC BY 4.0*
  - 14.2: Module 1 Discussion Forum - *CC BY 4.0*
  - 14.3: Reflective Blog 1 - *CC BY 4.0*
  - 14.4: Blog 1 Forum - *CC BY 4.0*
  - 14.5: Module 1 Homework - *CC BY 4.0*
  - 14.6: Module 1 Exam - *CC BY 4.0*
  - 14.7: Talk with the Professor - *CC BY 4.0*
- 15: Modules 2, 3, 4 and 5 - *CC BY 4.0*
  - 15.1: Structure of Modules 2 through 5 - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
  - Index - *CC BY 4.0*
  - Glossary - *CC BY 4.0*
  - Detailed Licensing - *CC BY 4.0*