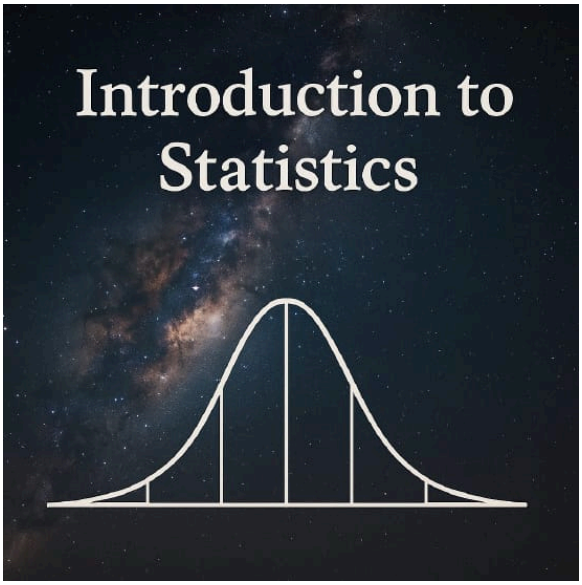
STATISTICS C1000: INTRODUCTION TO STATISTICS



Toros Berberyan, Tracy Nguyen, and Alfie Swan
Citrus College

LibreTexts™

Statistics C1000: Introduction to Statistics



By Toros Berberyan, Tracy Nguyen, and Alfie Swan

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.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 08/26/2026

TABLE OF CONTENTS

Licensing

Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework)

1: Fundamentals of Statistics

1.1: What is Statistics?

1.2: Sampling Methods

1.3: Experimental Design

1.4: How Not to Do Statistics

1.5: Chapter 1 - Key Terms and Symbols

2: Frequency Distributions, and Graphs

2.1: Qualitative Data

2.2: Quantitative Data

2.3: Other Graphical Representations of Data

2.4: Chapter 2 Formulas

2.5: Chapter 2 - Key Terms and Symbols

3: Descriptive Statistics

3.1: Measures of Center

3.2: Measures of Variation Version 1

3.3: Measures of Variation Version 2

3.4: Measures of Position

3.5: Formulas for Chapter 3

3.6: Chapter 3 - Key Terms and Symbols

4: Probability and Combinatorics

4.1: Empirical Probability

4.2: Theoretical Probability

4.3: Addition Rules for Probabilities

4.4: Multiplication Rules and Conditional Probability

4.5: At Least One Rule and Tree Diagrams

4.6: Counting Rules

4.7: Chapter 4 Formulas

4.8: Chapter 4 - Key Terms and Symbols

5: Discrete Probability Distributions

5.1: Basics of Probability Distributions

5.2: Binomial Probability Distribution

5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution

5.4: Poisson Distribution

5.5: Chapter 5 Formulas

5.6: Chapter 5 - Key Terms and Symbols

6: Continuous Probability Distribution

6.1: Introduction to the Normal Distribution

6.2: Finding Probabilities for the Normal Distribution

6.3: The Central Limit Theorem

6.4: Formulas for Chapter 6

6.5: Chapter 6 - Key Terms and Symbols

7: Confidence Interval for One Sample

7.1: Introduction to Confidence Intervals

7.2: Confidence Interval for the Mean Using z-values

7.3: Confidence Interval for the Mean Using t-values

7.4: Confidence Interval and Sample Size for the Proportion

7.5: Formulas for Chapter 7

7.6: Chapter 7 - Key Terms and Symbols

8: Hypothesis Testing for One Sample

8.1: Introduction to Hypothesis Testing

8.2: Type I and II Errors

8.3: z-Test for a Mean

8.4: t-Test for a Mean

8.5: z-Test for a Proportion

8.6: Formulas for Chapter 8

8.7: Chapter 8 Key Terms and Symbols

9: Hypothesis Testing for Two Samples

9.1: z-Test for the Difference Between Two Means

9.2: t-Test for the Difference Between Two Means

9.3: z-Test for the Difference Between Two Proportions

9.4: Confidence Intervals for the Difference Between Two Means and Proportions

9.5: Formulas for Chapter 9

9.6: Chapter 9 - Key Terms and Symbols

10: Correlation and Linear Regression

10.1: Bivariate Data and Scatter Plots

10.2: Correlation Coefficient

10.3: Hypothesis Test for Correlation Using r

10.4: Hypothesis Test for a Correlation Using t-Test

10.5: Linear Regression

10.6: Coefficient of Determination and the Standard Error of the Estimate

10.7: Formulas for Chapter 10

10.8: Chapter 10 - Key Terms and Symbols

11: Chi-Square Test

11.1: Goodness of Fit Test

11.2: Test for Independence

11.3: Chapter 11 Formulas

11.4: Chapter 11 - Key Terms and Symbols

12: Analysis of Variance (ANOVA)

12.1: Analysis of Variance (ANOVA)

12.2: Chapter 12 Formulas

12.3: Chapter 12 - Key Terms and Symbols

13: Statistical Tables

13.1: The Binomial Distribution Table for n = 2 to n = 10

13.2: The Binomial Distribution Table for n = 11 to n = 15

13.3: The Binomial Distribtution Table for n = 16 to n = 18

13.4: The Binomial Distribution Table for n = 19 to n = 20

13.5: The Standard Normal Distriubtion Table for Negative Z-Values

13.6: The Standard Normal Distribution Table for Positive Z-Values

13.7: The t-Distribution Table for Degrees of Freedom = n - 1

13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2

13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1

13.10: The F-Distribution for Alpha = 0.005

13.11: The F-Distribution Table for Alpha = 0.01

13.12: The F-Distribution Table for Alpha = 0.025

13.13: The F-Distribution Table for Alpha = 0.05

13.14: The F-Distribution Table for Alpha = 0.10

Index

Glossary

Detailed Licensing

Detailed Licensing

This page titled Statistics C1000 - Introduction to Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

Licensing

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

This page titled Licensing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework)

Introduction to MyOpenMath

MyOpenMath is a free, online learning platform designed to support math instruction through automated homework, quizzes, and assessments. It offers a wide range of interactive questions and feedback tailored for math courses like algebra, statistics, and calculus. Instructors can assign practice problems, monitor student progress, and customize content to align with course objectives. For students, MyOpenMath provides opportunities to learn from mistakes and practice as many times as needed before the due date.

To access the homework assignments and view the questions in each section, **you must first register for a free account and log in**.

🔗 **Visit MyOpenMath**

👩‍🏫 How to Register for MyOpenMath (Instructor Version)

Step-by-Step Instructions:

1. **Go to** MyOpenMath.com
2. **Click** on the link that says **“Request Instructor Access”**
   - You’ll find it under the login box on the homepage.

3. **Complete the Instructor Access Request Form**
   You’ll be asked to provide:

   - **Your full name**
   - **Your email address**
   - **Your school/institution name**
   - **A description of your teaching role** (such as “I teach Statistics at Citrus College.”)
   - **A link to a school staff directory or department page** where your name appears
   - Optionally, you may include the name of another instructor who can vouch for you
4. **Submit your request**
5. **Wait for approval**
   - Approval usually takes 1–2 business days.
   - You will receive an email once your account is activated.

📌 Important:

- You **must be logged in** to see our homework questions for each section.
- Save your login information in a safe place so you can access it later.

📚 For Instructors: How to Copy the Course

To copy the course for another section or share it with colleagues, follow these steps:

1. **Log into** MyOpenMath with your instructor account.
2. On the homepage, click **“Add New Course.”**
3. Fill in the following:
   - **Course Name** (e.g., "*Introductory Statistics - OER - Fall 2025*")

- **Course Availability Dates** (start and end)

4. In the section titled **“Copy a course,”** enter:
   - **Course ID to copy from**: `224605`
5. Choose one of the following:
   - **Copy entire course** (**recommended**), or
   - **Copy only selected items**
6. Click **“Submit”** to create your new course.

🔐 How to Get the Consumer Key and Secret from MyOpenMath for Canvas Integration *(Needed to Embed to Canvas)*

You need the consumer key and shared secret (also called the LTI key and secret) to connect your MyOpenMath course to Canvas using the **External Tool** option.

✅ Step-by-Step Instructions

1. **Log in** to your MyOpenMath *(opens in new window)* instructor account.
2. From the homepage, **go to the course** you want to embed in Canvas.
3. In the left-hand menu, click **“Course Settings.”**
   1. Open the section labeled **“LTI Integration”** and select **“Show course level key/secret.”**
      - It may also be called **“LTI Key and Secret”** or **“LTI Credentials.”**
4. You will see the following information:
   - **Consumer Key** (sometimes called "*LTI key*")
   - **Shared Secret**

5. **Write down the LTI Key** that ends with a **1** and the **LTI secret.**
   This is the information needed to embed your MyOpenMath course into Canvas.

🔗 How to Embed MyOpenMath Assignments into Canvas

This allows students to access MyOpenMath directly from Canvas, and grades sync automatically.

🌐 Step 1: Enable LTI (Learning Tools Interoperability)

1. **Log in to Canvas.**
2. **Go to the Course Settings.**
3. Click on the **Navigation** tab and ensure that **“MyOpenMath”** or **“External Tool”** is visible.
4. If it’s not installed yet:
   - Go to **Apps** > **View App Configurations** > **+App**
   - Choose **Manual Entry**
5. Add the following information:
   - **Name**: MyOpenMath

   - **Consumer Key (LTI Key)** and **Secret**
     (This is the LTI Key and Secret from Step 6 of the 🔐 How to Get the Consumer Key and Secret from MyOpenMath for Canvas Integration instructions)

   - **Launch URL**:
     `https://www.myopenmath.com/lti/launch.php`

6. **Click Submit.**

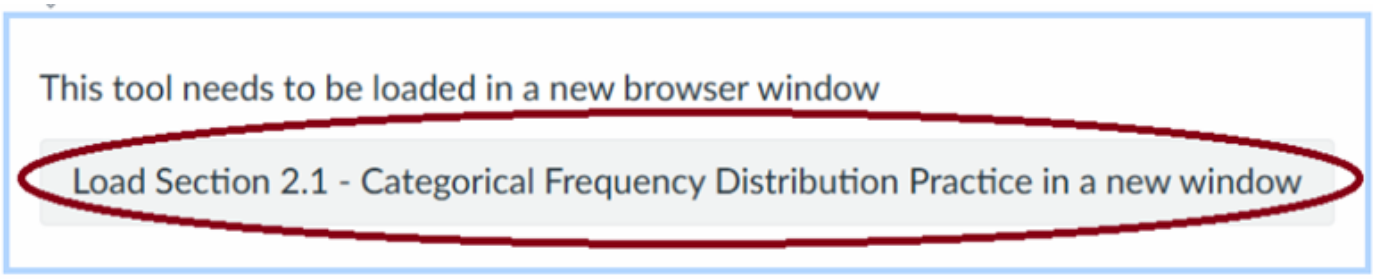
🔄 Step 2: Link Assignments in Canvas

1. In Canvas, go to **Assignments** or **Modules**.

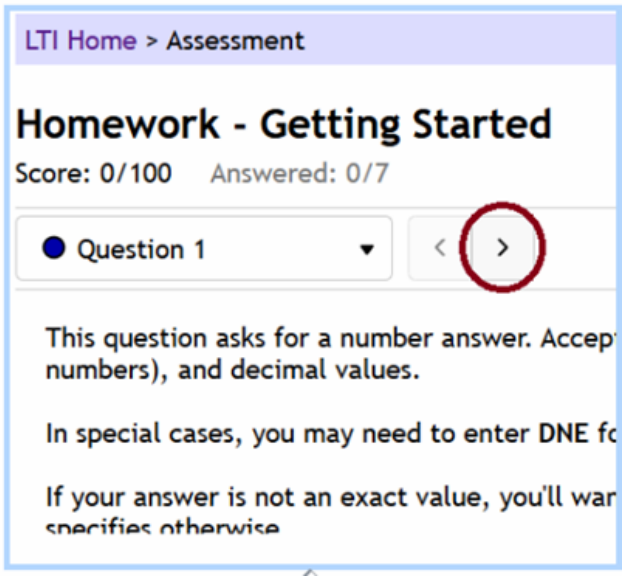
2. Click **+ Add Assignment** *(or* ***+ Add External Tool*** *under Modules).*
3. Select **"External Tool"**, then click **MyOpenMath.**
4. You'll be redirected to your MyOpenMath course.
5. Choose the assignment you want to link.
6. Click **"Use Selected Item."**
7. **Save and publish** the item in Canvas.

Instructions to Students in How to Navigate and Complete MyOpenMath Assignments in Canvas

Click on the link at the bottom of the page to access the homework. It will open in a new tab (**This only applies if you selected "Load This Tool In A New Tab"** under Submission Type)

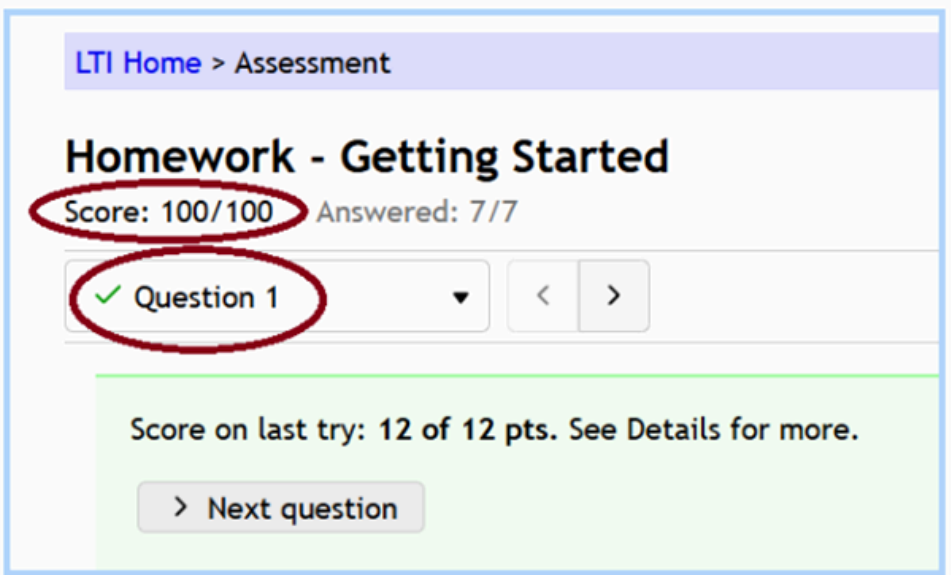


Answer **all questions** in the assignment. After answering a question, press the **">"** arrow to go to the next one.

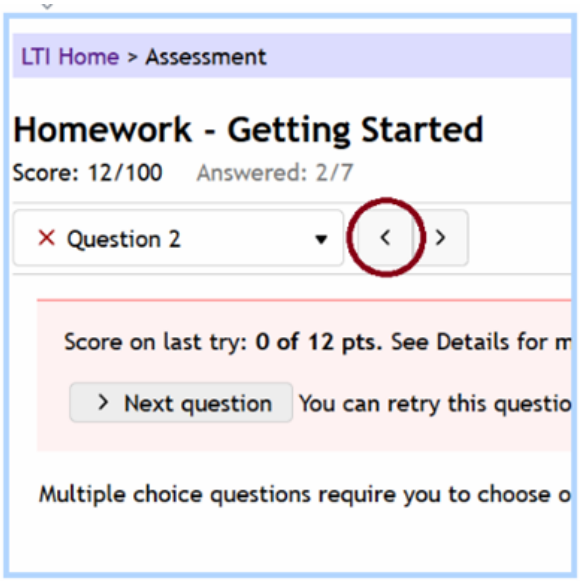


You will know you have completed an assignment when:

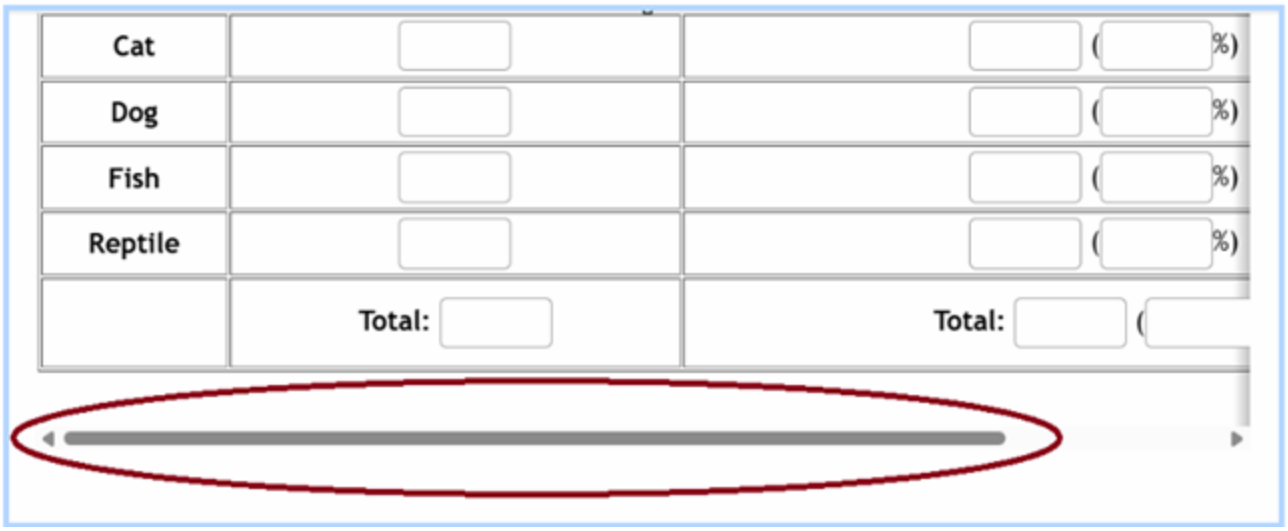
- Each question has a **green check mark** next to it.
- At the top left, it says: **"Score: 100/100."**



If you do not see a green check mark or a full score, click the **"<"** arrow to go back and review your answers. Make sure each question is answered correctly.

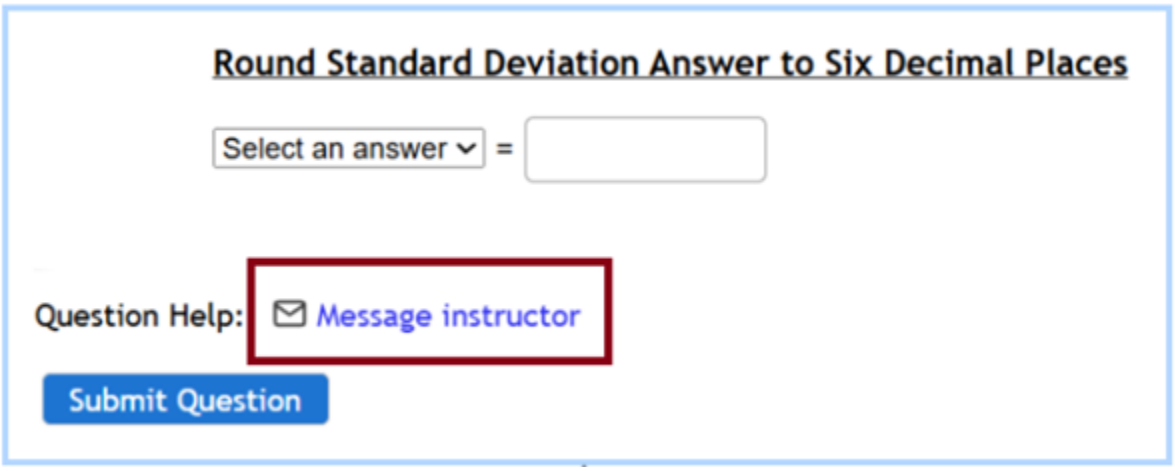


If you cannot see the whole problem on the new tab, use the scroll bar at the bottom of the problem to move side to side.

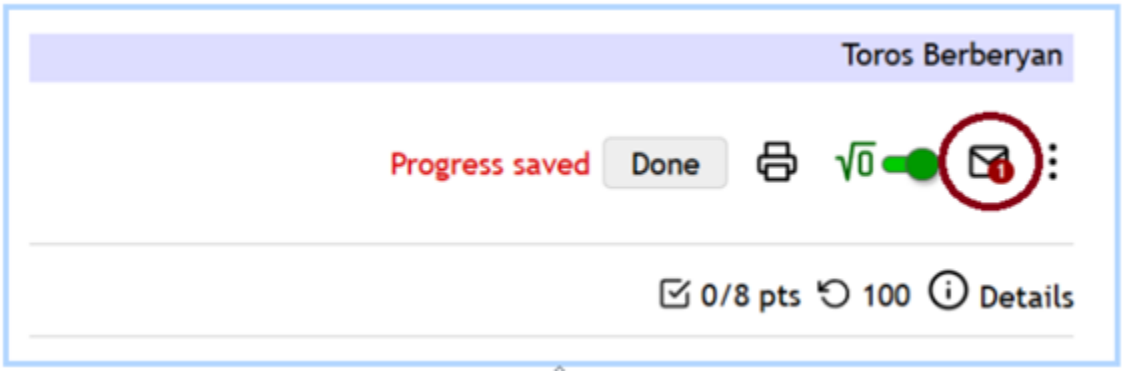


Student Instructions: Additional Support from Instructor

At the **bottom of each homework question**, there's a **"Message Instructor"** link.
Click it if you're stuck.



To view my reply, click the **envelope icon** at the top of any homework assignment.
A **red circle** next to the envelope means you have a new message from me.



How Professors Can Respond to Emails in MyOpenMath

1. **Log in** to MyOpenMath.
2. On the **homepage**, find the course where the student sent the message.
3. In the **left-hand menu**, click on **“Messages.”**
    - *If you don’t see “Messages,” click **“Expand Menu”** first.*
4. You will see your **Inbox**, which shows all messages sent by students.
5. Click on the **message subject** or **student’s name** to open the message.
6. Below the message, click the **“Reply”** button.
7. Type your **response** in the message box.
8. *(Optional)* You can **attach files** or **link to course materials**.
9. When ready, click **“Send Reply.”**

Contact Information if Need Assistance

- Toros Berberyan
- Tracy Nguyen
- Alfie Swan

Authors

"Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework)" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

1: FUNDAMENTALS OF STATISTICS

1.1: What is Statistics?
1.2: Sampling Methods
1.3: Experimental Design
1.4: How Not to Do Statistics
1.5: Chapter 1 - Key Terms and Symbols

This page titled 1: Fundamentals of Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

1.1: What is Statistics?

Learning Objectives

- Define key statistical terms such as population, sample, parameter, and statistic.
- Distinguish between quantitative and qualitative data.
- Differentiate between continuous and discrete variables.
- Understand the four levels of measurement: nominal, ordinal, interval, and ratio.

Introduction

You are exposed to statistics regularly. If you are a sports fan, then you have the statistics for your favorite player. If you are interested in politics, then you look at the polls to see how people feel about certain issues or candidates. If you are an environmentalist, then you research arsenic levels in the water of a town or analyze global temperatures. If you are in the business profession, then you may track the monthly sales of a store or use quality control processes to monitor the number of defective parts manufactured. If you are in the health profession, then you may look at how successful a procedure is or the percentage of people infected with a disease. There are many other examples from other areas. To understand how to collect data and analyze it, you need to understand what the field of statistics is and the basic definitions.

Definition 1.1.1

Statistics is the study of how to collect, organize, analyze, and interpret data collected from a group.

Exploring Descriptive and Inferential Statistics

There are two branches of statistics. One is called descriptive statistics, which is where you collect and organize data. The other is called inferential statistics, which is where you analyze and interpret data. First, you need to look at descriptive statistics since you will use the descriptive statistics when making inferences.

To understand how to create descriptive statistics and then conduct inferences, there are a few definitions that you need to look at. Note, many of the words that are defined have common definitions that are used in non-statistical terminology. In statistics, some have slightly different definitions. You must notice the difference and utilize the statistical definitions.

The first thing to decide in a statistical study is who you want to measure and what you want to measure. You always want to make sure that you can answer the question of whom you measured and what you measured. The who is known as the individual, and the what is the variable.

Definition 1.1.2

Individual – a person or object that you are interested in finding out information about.

Definition 1.1.3

Variable – the measurement or observation of the individual.

If you put the individual and the variable into one statement, then you obtain a population.

Definition 1.1.4

Population – set of all values of the variable for the entire group of individuals.

Notice, the population answers who you want to measure and what you want to measure. Make sure that your population always answers both of these questions. If it doesn't, then you haven't given someone who is reading your study the entire picture. As an example, if you just say that you are going to collect data from the senators in the U.S. Congress, you haven't told your reader what you are going to collect. Do you want to know their income, their highest degree earned, their voting record, their age, their political party, their gender, their marital status, or how they feel about a particular issue? Without telling what you want to measure, your reader has no idea what your study is actually about.

Sometimes the population is very easy to collect. Such as if you are interested in finding the average age of all of the current senators in the U.S. Congress, there are only 100 senators. This wouldn't be hard to find. However, if instead you were interested in knowing the average age that which a senator in the U.S. Congress first took office for all senators that ever served in the U.S. Congress, then this would be a bit more work. It is still doable, but it would take a bit of time to collect. But what if you are interested in finding the average diameter of breast height of all of the Ponderosa Pine trees in the Coconino National Forest? This would be impossible to actually collect. What do you do in these cases? Instead of collecting the entire population, you take a smaller group of the population, kind of a snapshot of the population. This smaller group is called a sample.

Definition 1.1.5

Sample – a subset from the population. It looks just like the population, but contains less data. Also, everybody has the same chance of being picked in the sample.

How you collect your sample can determine how accurate the results of your study are. There are many ways to collect samples. Some of them create better samples than others. No sampling method is perfect, but some are better than others. Sampling techniques will be discussed later. For now, realize that every time you take a sample, you will find different data values. The sample should mirror the population as much as possible. The sample can have different variations based on the sampling process. For example, assume we are trying to measure the average age of students in the classroom. If we collect two different samples of size 5, then most likely the average age will be different for both groups.

Once you have your data, either from a population or a sample, you need to know how you want to summarize the data. As an example, suppose you are interested in finding the proportion of people who like a candidate, the average height a plant grows to using a new fertilizer, or the variability of the test scores. Understanding how you want to summarize the data helps to determine the type of data you want to collect. Since the population is what we are interested in, you want to calculate a number from the population. This is known as a parameter. As mentioned already, you can't collect the entire population. Even though this is the number you are interested in, you can't really calculate it. Instead, you use the number calculated from the sample, called a statistic, to estimate the parameter. Since no sample is exactly the same, the statistical values are going to be different from sample to sample. They estimate the value of the parameter, but again, you do not know for sure if your answer is correct.

Numeric Summaries

Definition 1.1.6

Parameter – a number calculated from the population. Usually denoted with a Greek letter. This number is a fixed, unknown number that you want to find.

Definition 1.1.7

Statistic – a number calculated from the sample. Usually denoted with letters from the Latin alphabet, though sometimes there is a Greek letter with a ^ (called a hat) above it. Since you can find samples, it is readily known, though it changes depending on the sample taken. It is used to estimate the parameter value.

One last concept to mention is that there are two different types of variables – qualitative and quantitative. Each type of variable has different parameters and statistics that you find. It is important to know the difference between them.

Types of Data Variables

Definition 1.1.8

Qualitative or categorical variable – answer is a word or name that describes a quality of the individual.

Definition 1.1.9

Quantitative or numerical variable – answer is a number, something that can be counted or measured from the individual.

There are different types of quantitative variables, called discrete or continuous. The difference is in how many values the data can have. If you can count the number of data values (even if you are counting to infinity), then the variable is called discrete. If it is not possible to count the number of data values, then the variable is called continuous.

Different Types of Quantitative Variables

Definition 1.1.10

Discrete data can only take on particular values, like whole numbers. Discrete data are usually things that can be counted with specific values.

Definition 1.1.11

Continuous data can take on any value. Continuous data are usually things you measure.

Example 1.1.1 discrete or continuous

Classify the quantitative variable as discrete or continuous.

a. The weight of a cat.
b. The number of fleas on a cat.
c. The size of a shoe.

Solution

a. This is continuous since it is something you measure.
b. This is discrete since it is something you count.
c. This is discrete since you can only be certain values, such as $7, 7.5, 8, 8.5, 9$. You can't buy a 9.73 shoe.

There are also are four measurement scales for different types of data, with each building on the ones below it. They are:

Measurement of Scales

Definition 1.1.12

Nominal – data is just a name or category. There is no order to any data, and since there are no numbers, you cannot do any arithmetic on this level of data. Examples of this are gender, car name, ethnicity, and race.

Definition 1.1.13

Ordinal – data that is nominal, but you can now put the data in order, since one value is more or less than another value. You cannot do arithmetic on this data, but you can now put data values in order. Examples of this are grades (A, B, C, D, F), place value in a race (1st place, 2nd place, 3rd place), and size of a drink (small, medium, large). Also, it is not measurable.

Definition 1.1.14

Interval – data that is ordinal, but you can now subtract one value from another, and that subtraction makes sense. You can do arithmetic on this data, but only addition and subtraction. Examples of this are temperature and time on a clock. The zero value is a marker and not an absence of measurement.

Definition 1.1.15

Ratio – data that is interval, but you can now divide one value by another, and that ratio makes sense. You can now do all the arithmetic on this data. Examples of this are height, weight, distance, and time. The zero value is an absence of measurement.

Nominal and ordinal data come from qualitative variables. Interval and ratio data come from quantitative variables.

Most people have a hard time deciding if the data are nominal, ordinal, interval, or ratio. First, if the variable is qualitative (words instead of numbers), then it is either nominal or ordinal. Now, ask yourself if you can put the data in a particular order. If you can, it is ordinal. Otherwise, it is nominal. If the variable is quantitative (numbers), then it is either interval or ratio. For ratio data, a value of 0 means there is no measurement. This is known as the absolute zero. If there is an absolute zero in the data, then it means it is a ratio. If there is no absolute zero, then the data are interval. An example of an absolute zero is if you have $\$0$ in your bank account, then you have without money. The amount of money in your bank account is ratio data. Word of caution, sometimes ordinal data is displayed using numbers, such as 5 being strongly agree, and 1 being strongly disagree. These numbers are not numbers. Instead, they are used to assign numerical values to ordinal data. In reality, you should not perform any computations on this data, though many people do. If there are numbers, make sure the numbers are inherent, and not numbers that were assigned.

Examples of Measurement Scales

Example 1.1.2 measurement scale

State which measurement scale is nominal, ordinal, interval, or ratio.

a. Time of first class
b. Hair color
c. Length of time to take a test
d. Age groupings (baby, toddler, adolescent, teenager, adult, elderly)

Solution

a. This is an interval since it is a number, but 0 o'clock means midnight and not the absence of time.

b. This is nominal since it is not a number, and there is no specific order for hair color.
c. This is a ratio since it is a number, and if you take 0 minutes to take a test, it means you didn't take any time to complete it.
d. This is ordinal since it is not a number, but you could put the data in order from youngest to oldest or the other way around.

Examples of Descriptive Statistics

Example 1.1.3 stating definitions for qualitative variable

In 2010, the Pew Research Center questioned 1500 adults in the U.S. to estimate the proportion of the population favoring marijuana use for medical purposes. It was found that 73% are in favor of using marijuana for medical purposes. State the individual, variable, population, and sample.

Solution

Individual – a U.S. adult

Variable – the response to the question “Should marijuana be used for medical purposes?” This is qualitative data since you are recording a person’s response – yes or no.

Population – set of all responses of adults in the U.S.

Sample – set of 1500 responses of U.S. adults who were questioned.

Parameter – proportion of those who favor marijuana for medical purposes calculated from the population

Statistic– proportion of those who favor marijuana for medical purposes calculated from the sample

Example 1.1.4 stating definitions for qualitative variable

A parking control officer records the manufacturer of every 5th car in the college parking lot to guess the most common manufacturer.

Solution

Individual – a car in the college parking lot.

Variable – the name of the manufacturer. This is qualitative data since you are recording a car type.

Population – All the cars in the parking lot are based on the manufacturer.

Sample – Some of the cars in the parking lot are based on the manufacturer.

Parameter – proportion of each car type calculated from the population.

Statistic – proportion of each car type calculated from the sample.

Example 1.1.5 stating definitions for quantitative variable

A biologist wants to estimate the average height of a plant that is given a new plant food. She gives 10 plants the new plant food. State the individual, variable, population, and sample.

Solution

Individual – a plant given the new plant food

Variable – the height of the plant (Note: it is not the average height since you cannot measure an average – it is calculated from data.) This is quantitative data since you will have a number.

Population – set of all possible heights of plants, provided they were given the new plant food

Sample – set of 10 heights of plants when the new plant food is used

Parameter – average height of all plants

Statistic – average height of 10 plants

Example $1.1.6$ stating definitions for quantitative variable

A doctor wants to see if a new treatment for cancer extends the life expectancy of a patient versus the old treatment. She gives one group of 25 cancer patients the new treatment and another group of 25 the old treatment. She then measures the life expectancy of each of the patients. State the individuals, variables, populations, and samples.

Solution

In this example there are two individuals, two variables, two populations, and two samples.

Individual 1: Cancer patient given new treatment

Individual 2: Cancer patient given old treatment

Variable 1: life expectancy when given a new treatment. This is quantitative data since you will have a number.

Variable 2: life expectancy when given the old treatment. This is quantitative data since you will have a number.

Population 1: set of all life expectancies of cancer patients given a new treatment

Population 2: set of all life expectancies of cancer patients given the old treatment

Sample 1: set of 25 life expectancies of cancer patients given new treatment

Sample 2: set of 25 life expectancies of cancer patients given old treatment

Parameter 1 – average life expectancy of all cancer patients given a new treatment

Parameter 2 – average life expectancy of all cancer patients given the old treatment

Statistic 1 – average life expectancy of 25 cancer patients given new treatment

Statistic 2 – average life expectancy of 25 cancer patients given old treatment

Authors

"1.1: What is Statistics?" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"1.1: What is Statistics?" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 1.1: What is Statistics? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: What is Statistics? by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

1.2: Sampling Methods

Learning Objectives

- Introduction of various sampling methods used for effective data collection.
- Understand and apply simple random, stratified, systematic, cluster, and convenience sampling techniques.
- Select appropriate sampling methods based on population structure and accessibility.

Introduction

As stated before, if you want to know something about a population, it is often impossible or impractical to examine the whole population. It might be too expensive in terms of time or money. It might be impractical – you can't test all batteries for their length of lifetime because there wouldn't be any batteries left to sell. You need to look at a sample. Hopefully, the sample behaves the same as the population.

When you choose a sample, you want it to be as similar to the population as possible. If you want to test a new painkiller for adults, you would want the sample to include people who are fat, skinny, old, young, healthy, not healthy, male, female, etc.

There are many ways to collect a sample. None are perfect, and you are not guaranteed to collect a representative sample. That is, unfortunately, the limitations of sampling. However, several techniques can result in samples that give you a semi-accurate picture of the population. Just remember to be aware that the sample may not be representative. As an example, you can take a random sample of a group of people that are equally males and females, yet by chance, everyone you choose is female. If this happens, it may be a good idea to collect a new sample if you have the time and money.

There are many sampling techniques, though only four will be presented here. The simplest, and the type that is strived for, is a **simple random sample**. This is where you pick the sample such that every sample has the same chance of being chosen. This type of sample is hard to collect, since it is sometimes difficult to obtain a complete list of all individuals. There are many cases where you cannot conduct a truly random sample. However, you can get as close as you can. Now, we suppose you are interested in what type of music people like. It might not make sense to try to find an answer for everyone in the U.S. You probably don't like the same music as your parents. The answers vary so much that you probably couldn't find an answer for everyone all at once. It might make sense to look at people in different age groups or people of different ethnicities. This is called a **stratified sample**. The issue with this sample type is that sometimes people subdivide the population too much. It is best to just have one stratification. Also, a stratified sample has similar problems that a simple random sample has. If your population has some order in it, then you could do a **systematic sample**. This is popular in manufacturing. The problem is that it is possible to miss a manufacturing mistake because of how this sample is taken. If you are collecting polling data based on location, then a **cluster sample** that divides the population based on geographical means would be the easiest sample to conduct. The problem is that if you are looking for the opinions of people, the people who live in the same region may have similar opinions. As you can see, each of the sampling techniques has pluses and minuses.

Simple Random Sample

Definition 1.2.1

A **simple random sample (SRS)** of size n is a sample that is selected from a population in a way that ensures that every different possible sample of size n has the same chance of being selected. Also, every individual associated with the population has the same chance of being selected

Examples of Simple Random Samples:

- Put all names in a hat and draw a certain number of names out.
- Assign each individual a number and use a random number table or a calculator, or a computer to randomly select the

individuals that will be measured.

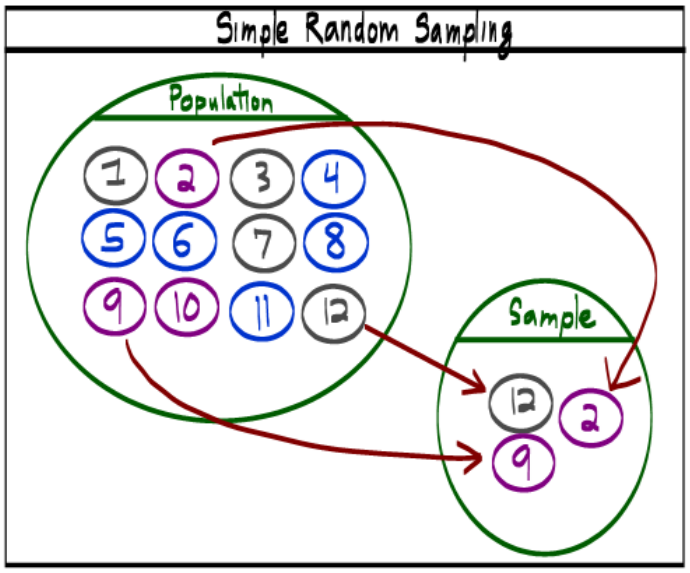


Figure 1.2.1: "Simple Random Sample Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

In the simple random sampling, there is an original population of 12 balls. We conduct a random sample by randomly selecting three balls from the population to be part of the sample. In this case, we have picked the ones that have the numbers 2, 9, and 12.

Example 1.2.1 Choosing a Simple Random Sample

Describe how to take a simple random sample from a classroom.

Solution

Give each student in the class a number. Using a random number generator, you could then pick the number of students you want to pick.

Example 1.2.2 How *Not* to Choose a Simple Random Sample

You want to choose 5 students out of a class of 20. Give some examples of samples that are not simple random samples:

Solution

Choose 5 students from the front row. The people in the last row have no chance of being selected.

Choose the 5 shortest students. The tallest students have no chance of being selected.

Stratified Sampling

Definition 1.2.2

Stratified sampling is where you break the population into groups called strata, and then take a simple random sample from each strata.

Examples of Stratified Sampling:

- If you want to look at musical preference, you could divide the individuals into age groups and then conduct simple random samples inside each group.
- If you want to calculate the average price of textbooks, you could divide the individuals into groups by major and then conduct simple random samples inside each group.

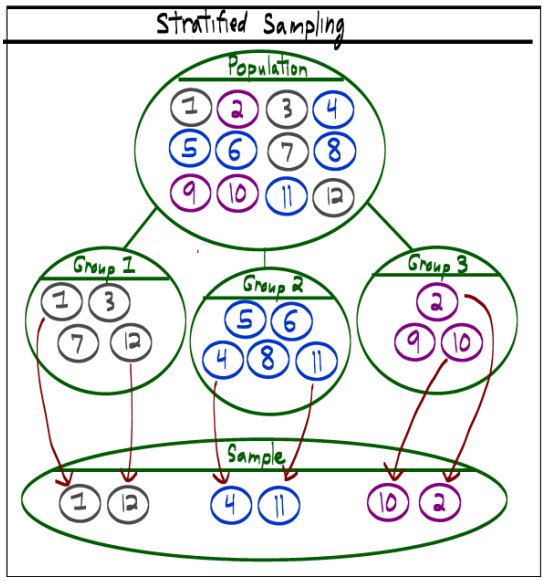


Figure 1.2.2: "Stratified Sampling Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

The illustration above is an example of a stratified sample. There is an original population of 12 balls. They are grouped based on a common characteristic, in this case, color. We have 4 dark grey balls in the first group, 5 blue balls in the second group, and 3 purple balls in the third group. From **all groups**, **some members** are randomly selected to be part of the sample. From the first group, the balls that are numbered 1 and 12 are selected to be part of the sample. From the second group, the balls that are numbered 4 and 11 are selected to be part of the sample. From the third group, the balls that are numbered 2 and 10 are selected to be part of the sample. Based on stratified sampling, we ended up with a sample of 6 balls numbered 1, 2, 4, 10, 11, and 12.

Systematic Sampling

Definition 1.2.3

Systematic sampling is where you randomly choose a starting place and then select every kth individual to measure.

Examples of Systematic Sampling:

- You select every 5th item on an assembly line
- You select every 10th name on the list
- You select every 3rd customer that comes into the store.

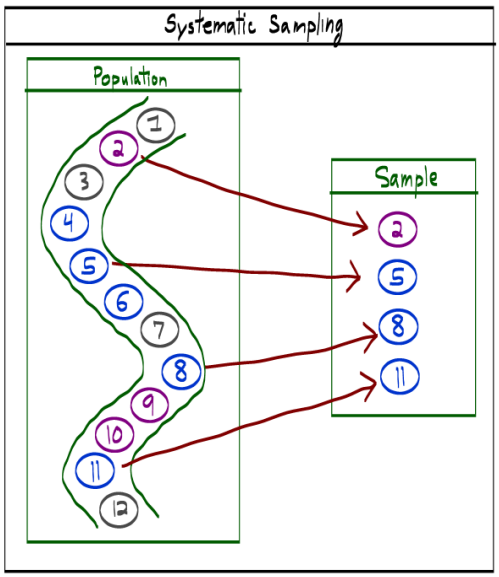


Figure 1.2.3: "Systematic Sampling Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

The diagram above represents a systematic sample. We have a population of 12 balls inside a curvy road. The first one is picked randomly. In this example, it will be the second one from above. Then, the rest of them are selected below the first one in a certain pattern. The diagram shows that every third one is selected to be part of the sample. Based on systematic sampling, the ones selected to be in the sample are the ones numbered 2, 5, 8, and 11.

Cluster Sampling

Definition 1.2.4

Cluster sampling is where you break the population into groups called clusters. Randomly pick some clusters then poll all individuals in those clusters.

Examples of Cluster Sampling:

- A large city wants to poll all businesses in the city. They divide the city into sections (clusters), maybe a square block for each section, and use a random number generator to pick some of the clusters. Then, they poll all businesses in each chosen cluster.
- You want to measure whether a tree in the forest is infected with bark beetles. Instead of having to walk all over the forest, you divide the forest up into sectors, and then randomly pick the sectors that you will travel to. Then record whether a tree is infected or not for every tree in that sector.

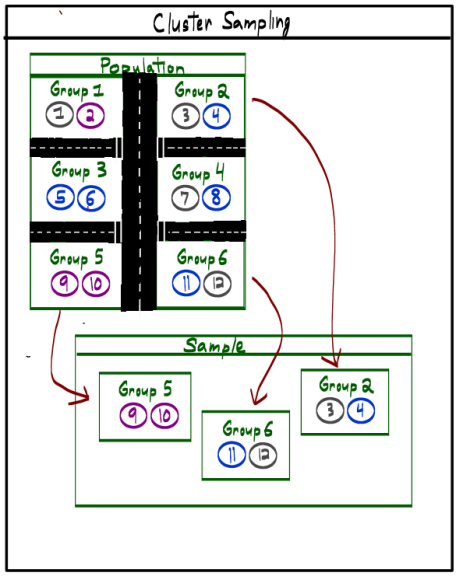


Figure 1.2.4: "Cluster Sampling Diagram" by Toros Berberyan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

The illustration above is an example of a **cluster sample**. There is an original population of 12 balls. They are grouped into clusters (blocks). Each group has two balls. Then, **some groups** are selected, and **all members** from those groups are chosen to be part of the sample. In this example, the group that only had balls numbered 3 and 4 is selected to be part of the sample. Then, a second group with only balls numbered 9 and 10 is selected to be part of the sample. Finally, a third group with only balls numbered 11 and 12 is selected to also be part of the sample. Based on cluster sampling, we ended up with three groups and all its members being part of the sample. The sample included balls with the numbers 3, 4, 9, 10, 11, and 12.

Note

Many people confuse stratified sampling and cluster sampling. In **stratified sampling**, you use **all the groups** and **some of the members** in each group. Cluster sampling is the other way around. **Cluster sampling** uses **some of the groups** and **all the members** in each group.

The four sampling techniques that were presented all have advantages and disadvantages. There is another sampling technique that is sometimes utilized because either the researcher doesn't know better, or it is easier to do. This sampling technique is known as a convenience sample.

Convenience Sampling

Definition 1.2.5

A convenience sample is one where the researcher picks individuals to be included who are easy for the researcher to collect.

Example of a Convenience Sample:

If you want to know the opinion of people about the criminal justice system, you stand on a street corner near the county courthouse, and question the first 10 people who walk by. The people who walk by the county courthouse are most likely involved in some fashion with the criminal justice system, and their opinions would not represent the opinions of all individuals.

Note

The **convenience sample** will not result in a representative sample and **should be avoided**.

On a rare occasion, you do want to collect the entire population. In which case, you conduct a census.

Census

Definition 1.2.6

A **census** is when every individual of interest is measured.

Review

Example 1.2.3 Sampling Type

Banner Health is a multi-state nonprofit chain of hospitals. Management wants to assess the incidence of complications after surgery. They wish to use a sample of surgery patients. Several sampling techniques are described below. Categorize each technique as a simple random sample, stratified sample, systematic sample, cluster sample, or convenience sampling.

a. Obtain a list of patients who had surgery at all Banner Health facilities. Divide the patients according to the type of surgery. Draw simple random samples from each group.
b. Obtain a list of patients who had surgery at all Banner Health facilities. Number these patients, and then use a random number table to obtain the sample.
c. Randomly select some Banner Health facilities from each of the seven states, and then include all the patients on the surgery lists of the states.
d. At the beginning of the year, instruct each Banner Health facility to record any complications from every 100th surgery.
e. Instruct each Banner Health facility to record any complications from 20 surgeries this week and send in the results.

Solution

a. This is a stratified sample since the patients were separated into different strata, and then random samples were taken from each stratum. The problem with this is that some types of surgeries may have a higher chance of complications than others. Of course, the stratified sample would show you this.
b. This is a random sample since each patient has the same chance of being chosen. The problem with this one is that it will take a while to collect the data.
c. This is a cluster sample since all patients are questioned in each of the selected hospitals. The problem with this is that you could have, by chance, selected hospitals that have no complications.
d. This is a systematic sample since they selected every 100th surgery. The problem with this is that if every 90th surgery has complications, you wouldn't see this come up in the data.
e. This is a convenience sample since they left it up to the facility on how to do it. The problem with convenience samples is that the person collecting the data will probably collect data from surgeries that had no complications.

Authors

"1.2: Sampling Methods" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"1.2: Sampling Methods" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 1.2: Sampling Methods is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 1.2: Sampling Methods by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.
- 1.1: What is Statistics? by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

1.3: Experimental Design

Learning Objectives

- Define observational studies and experiments as primary methods of data collection.
- Distinguish between non-interference in observational studies and controlled conditions in experiments.
- Understand common experimental designs, including completely randomized, block, and control designs

The section is an introduction to experimental design. There are two types of research: observational studies and experiments. This section describes how to design an experiment or a survey so that they are statistically sound. Experimental design is a very involved process, so this is just a small introduction.

Observational Studies

Definition 1.3.1

An **observational study** is when the investigator collects data merely by watching or asking questions. He doesn't change anything.

Many observational studies involve surveys. A **survey** uses questions to collect the data and needs to be written so that there is no bias.

Cross-Sectional Study:

Data observed, measured, or collected at one point in time.

Retrospective (or Case-Control) Study:

Data was collected from the past using records, interviews, and other similar artifacts.

Prospective (or Longitudinal or Cohort) Study:

Researchers closely observe and gather data on groups of individuals with common factors over some time to track outcomes and determine potential relationships between factors.

Experiments

Definition 1.3.2

An **experiment** is when the investigator changes a variable or imposes a treatment to determine its effect.

In an experiment, there are various options, such as a randomized two-treatment experiment, a randomized block design, a rigorously controlled design, and more. For this class, we will only focus on the ones mentioned.

Randomized Two-Treatment Experiment:

In this experiment, there are two treatments, and individuals are randomly placed into the two groups. Either both groups get a

treatment, or one group gets a treatment and the other gets either nothing or a placebo. The group getting either no treatment or the placebo is called the control group. The group getting the treatment is called the treatment group. The idea of the placebo is that a person thinks they are receiving a treatment, but in reality, they are receiving a sugar pill or fake treatment. Doing this helps to account for the placebo effect, which is where a person's mind makes their body respond to a treatment because they think they are taking the treatment when they are not taking the treatment. Note that every experiment does not need a placebo, such as when using animals or plants. Also, you can't always use a placebo or no treatment. For example, if you are testing a new blood pressure medication, you can't give a person with high blood pressure a placebo or no treatment for moral reasons.

Randomized Block Design:

A block is a group of subjects that are similar, but the blocks differ from each other. Then randomly assign treatments to subjects inside each block. An example would be separating students into full-time versus part-time, and then randomly picking a certain number of full-time students to receive the treatment and a certain number of part-time students to receive the treatment. This way, each type of student gets the treatment, and some do not.

Rigorously Controlled Design:

Carefully assign subjects to different treatment groups, so those given each treatment are similar in ways that are important to the experiment. An example would be if you want to have a full-time student who is male, takes only night classes, has a full-time job, and has children in one treatment group, then you need to have the same type of student getting the other treatment group. This type of design is hard to implement since you don't know how many differentiations you would use, and should be avoided.

Replication:

Repetition of an experiment on more than one subject ensures that the sample is large enough to distinguish true effects from random effects. It is also the ability for someone else to duplicate the results of the experiment.

Blind Study:

A blind study is where the individual does not know which treatment they are getting or if they are getting the treatment or a placebo.

Double-Blind Study:

A double-blind study is used when neither the individual nor the researcher knows who is getting the treatment or the placebo. This is important so that there can be no bias created by either the individual or the researcher.

One last consideration is the time period during which you are collecting the data. There are three types of periods that you can consider.

Example 1.3.1 observational study or experiment

State if the following is an observational study or an experiment.

a. Poll students to see if they favor increasing tuition.
b. Give some students a tutor to see if their grades improve.

Solution

a. This is an observational study. You are only asking a question.
b. This is an experiment. The tutor is the treatment.

Guidelines for Planning a Statistical Study or Experiment

1. Identify the individuals that you are interested in. Realize that you can only make conclusions for these individuals. For example, if you use a fertilizer on a certain type of plant, you can't say how the fertilizer will work on any other type of plant. However, if you diversify too much, then you may not be able to tell if there really is an improvement since you have too many factors to consider.
2. Specify the variable. You want to make sure this is something you can measure and make sure you control for all other factors, too. As an example, if you are trying to determine if a fertilizer works by measuring the height of the plants on a particular day, you need to make sure you can control how much fertilizer you put on the plants (which would be your treatment), and make sure that all the plants receive the same amount of sunlight, water, and temperature.
3. Specify the population. This is important to know what conclusions you can make and who you are making the conclusions.
4. Specify the method for taking measurements or making observations.
5. Determine if you are taking a census or a sample. If taking a sample, decide on the sampling method.
6. Collect the data.
7. Use appropriate descriptive statistics methods and make decisions using appropriate inferential statistics methods.
8. Note any concerns you might have about your data collection methods and list any recommendations for the future.

Authors

"1.3: Experimental Design" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"1.3: Experimental Design" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 1.3: Experimental Design is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 1.3: Experimental Design by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

1.4: How Not to Do Statistics

Learning Objectives

- Identify common pitfalls in statistical reasoning.
- Recognize the risks of overgeneralizing results and misinterpreting sampling errors.
- Understand the importance of avoiding incorrect assumptions about cause and effect.
- Promote accurate and reliable interpretation of statistical conclusions.

Many studies are conducted, and conclusions are made. However, there are occasions where the study is not conducted correctly or the conclusion is not correctly made based on the data. There are many things that you should question when you read a study. There are many reasons for the study to have a bias. Bias is where a study may have a certain slant or preference for a certain result. The following is a list of some of the questions or issues you should consider to help decide if there is bias in a study.

One of the first issues you should ask is who funded the study. If the entity that sponsored the study stands to gain either profits or notoriety from the results, then you should question the results. It doesn't mean the results are wrong, but you should scrutinize them to ensure they are sound. For example, suppose a study says that genetically modified foods are safe, and the study was funded by a company that sells genetically modified food, then one may question the validity of the study. Since the company funds the study and its profits rely on people buying their food, there may be a bias.

An experiment could have **lurking or confounding variables** when you cannot rule out the possibility that the observed effect is due to some other variable rather than the factor being studied. For example, suppose you give fertilizer to some plants and no fertilizer to others, but the plants are not growing taller due to a lack of sunlight and not due to a lack of fertilizer. You won't know if the plants that received the fertilizer grew taller because of the fertilizer or the sunlight. Make sure you design experiments to eliminate the effects of confounding variables by controlling all the factors.

Overgeneralization

Overgeneralization is a study of one group but makes conclusions for all groups. An example is doing cancer treatments on rats. Just because the treatment works on rats does not mean it will work on humans. Another example is that until recently, most FDA medication testing had been done on white males of a particular age. There is no way to know how the medication affects other genders, ethnic groups, age groups, and races. The new FDA guidelines stress individuals from different groups.

Cause and Effect

Cause and effect is where people assume that one variable causes the other when the variables are only correlated. Unless the study was done as an experiment where a variable was controlled, you cannot say that one variable caused the other. Most likely, there is another variable that caused both. As an example, there is a relationship between the number of drownings at the beach and ice cream sales. This does not mean that purchasing ice cream increases the chance people will drown. Most likely, the cause for both increases is the heat.

Sampling Error

This is the difference between the sample results and the true population results. This is unavoidable and results in the fact that samples are different from each other. For example, if you take a sample of 5 people's heights in your class, you will get 5 numbers. If you take another sample of 5 people's heights in your class, you will likely get 5 different numbers.

Nonsampling Error

This is where the sample is collected poorly, either through a biased sample or an error in measurements. Care should be taken to avoid this error.

Lastly, there should be care taken in considering the difference between **statistical significance versus practical significance**. This is a major issue in statistics. Something could be statistically significant, which means that a statistical test shows there is evidence to support what you are trying to prove. However, in practice, it doesn't mean much, or there are other issues to consider. For example, suppose you find a new drug for high blood pressure that reduces blood pressure. When you look at the improvement, it doesn't amount to a large difference. Even though statistically there is a change, it may not be worth marketing the product because it isn't that big of a change. Another consideration is that the blood pressure medication improves a person's blood pressure, but it has serious side effects, or it costs a great deal for a prescription. In this case, it wouldn't be practical to use it. In both cases, the study is shown to be statistically significant, but practically, you don't want to use the medication. The main thing to remember in a statistical study is that the statistics are only part of the process. You also want to make sure that there is practical significance, too.

Surveys

Surveys have their own areas of bias that can occur. A few of the issues with surveys are in the wording of the questions, the ordering of the questions, the manner the survey is conducted, and the response rate of the survey.

The wording of the questions can cause **hidden bias**, which is where the questions are asked in a way that makes a person respond a certain way. An example is that a poll was done where people were asked if they believed that there should be an amendment to the Constitution protecting a woman's right to choose. About 60% of all people questioned said yes. Another poll was done where people were asked if they believed that there should be an amendment to the Constitution protecting the life of an unborn child. About 60% of all people questioned said yes. These two questions deal with the same issue, though giving opposite results, but how the question was asked affected the outcome.

The ordering of the questions can also cause hidden bias. An example of hidden bias is if you were asked if there should be a fine for texting while driving, but the preceding question asked if you ever texted while driving. By asking a person if they partake in the activity, that person now personalizes the question and may affect how they answer the next question.

Non-response

Non-response is when you send out a survey, but not everyone responds. You can calculate the response rate by dividing the number of responses by the number of surveys sent. Most response rates are around 30-50%. A response rate that is less than 30% is very poo,r and the results of the survey are not valid. To reduce non-response, it is better to conduct the surveys in person, though these are very expensive. Phones are the next best way to conduct surveys, emails can be effective, and physical mailings are the least desirable way to conduct surveys.

Voluntary response

A voluntary response is where people are asked to respond via phone, email, or online and decide to be part of the survey. The problem with these is that only people who care about the topic are more likely to respond. These surveys are not scientific, and the results from these surveys are not valid. Note: all studies involve volunteers. The difference between a voluntary response survey and a scientific study is that in a scientific study, the researchers ask the individuals to be involved, while in a voluntary they choose to be part of the study.

Example 1.4.1: Bias in a Study

Suppose a mathematics department at a community college would like to assess whether computer-based homework improves students' test scores. They use computer-based homework in one classroom with one teacher and traditional paper and pencil homework in a different classroom with a different teacher. The students using the computer-based homework had higher test scores. What is wrong with this experiment?

Solution

Since there were different teachers, you do not know if the better test scores are because of the teacher or the computer-based homework. A better design would be to have the same teacher teach in both classes. The control group would utilize traditional paper and pencil homework, and the treatment group would use computer-based homework. Both classes would

have the same teacher, and the students would be split between the two classes randomly. The only difference between the two groups should be the homework method. Of course, there is still variability between the students, but utilizing the same teacher will reduce any other confounding variables.

Example 1.4.2: Cause and Effect

Determine if a variable causes a change in the other variable.

a. Cinnamon was given to a group of people who had diabetes, and then their blood glucose levels were measured a day later. All other factors for each person were kept the same, and their glucose levels were lower. Did the cinnamon cause the reduction?
b. There is a link between spray-on tanning products and lung cancer. Does that mean that spray-on tanning products cause lung cancer?

Solution

a. Since this was a study where the use of cinnamon was controlled, and all other factors were kept constant from person to person, any changes in glucose levels can be attributed to the use of cinnamon.
b. Since there is only a link, and not a study controlling the use of the spray-on tanning products, you cannot say there is a cause-and-effect relationship between using the product and lung cancer. You can say that there is a link, and that there could be a cause, but you cannot say for sure that the spray causes the cancer.

Example 1.4.3: Generalization

a. A researcher conducts a study on the use of ibuprofen on people and finds that it is safe. Does that mean that all species can use ibuprofen?
b. Aspirin has been used for years to bring down fevers in people. Originally, it was tested on white males between the ages of 25 and 40 and found to be safe. Is it safe to give to everyone?

Solution

a. No, because if a drug is safe to use on one species, it doesn't mean it is safe for all species. Ibuprofen is toxic to cats.
b. No, because if one age group can use it doesn't mean it is safe for all age groups. There has been a link between giving a child under the age of 19 aspirin when they have a fever and Reye's syndrome.

Authors

"1.4: How Not to Do Statistics" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"1.4: How Not to Do Statistics" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 1.4: How Not to Do Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 1.4: How Not to Do Statistics by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

1.5: Chapter 1 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Census: when every individual of interest is measured.

Cluster sampling: The population is divided into groups called clusters. Some clusters are randomly chosen, and all individuals in those clusters are polled.

Continuous data: takes on any value. Continuous data are usually things you measure.

Convenience sample: when the researcher picks individuals to be included who are easy for the researcher to collect.

Discrete: data can only take on particular values, like whole numbers. Discrete data are usually things that can be counted with specific values.

Experiment: when the investigator changes a variable or imposes a treatment to determine its effect.

Individual: a person or object that you are interested in finding out information about.

Interval: data that is ordinal, but you can now subtract one value from another, and that subtraction makes sense. You can do arithmetic on this data, but only addition and subtraction. Examples of this are temperature and time on a clock. The zero value is a marker and not an absence of measurement.

Nominal: data is just a name or category. There is no order to any data, and since there are no numbers, you cannot do any arithmetic on this level of data. Examples of this are gender, car name, ethnicity, and race.

Observational study: when the investigator collects data merely by watching or asking questions. He doesn't change anything.

Ordinal: data that is nominal, but you can now put the data in order since one value is more or less than another value. You cannot do arithmetic on this data, but you can now put data values in order. Examples of this are grades (A, B, C, D, F), place value in a race (1st place, 2nd place, 3rd place), and size of a drink (small, medium, large). Also, it is not measurable.

Parameter: a number calculated from the population. Usually denoted with a Greek letter. This number is a fixed, unknown number that you want to find.

Population: a set of all values of the variable for the entire group of individuals.

Qualitative or categorical variable: the answer is a word or name that describes a quality of the individual.

Quantitative or numerical variable: the answer is a number, something that can be counted or measured from the individual.

Ratio: data that is interval, but you can now divide one value by another, and that ratio makes sense. You can now do all the arithmetic on this data. Examples of this are height, weight, distance, and time. The zero value is an absence of measurement.

Sample: a subset of the population. It looks just like the population, but contains less data. Also, everybody has the same chance of being picked in the sample.

Simple random sample (SRS): is a sample of size n that is selected from a population in a way that ensures that every different possible sample of size n has the same chance of being selected. Also, every individual associated with the population has the same chance of being selected.

Statistic: a number calculated from the sample. Usually denoted with letters from the Latin alphabet, though sometimes there is

a Greek letter with a ^ (called a hat) above it. Since you can find samples, it is readily known, though it changes depending on the sample taken. It is used to estimate the parameter value.

Statistics: the study of how to collect, organize, analyze, and interpret data collected from a group.

Systematic sampling: where you randomly choose a starting place and then select every K^{th} individual to measure.

Variable: the measurement or observation of the individual.

Authors

"1.5: Chapter 1 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 1.5: Chapter 1 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

2: FREQUENCY DISTRIBUTIONS, AND GRAPHS

2.1: Qualitative Data
2.2: Quantitative Data
2.3: Other Graphical Representations of Data
2.4: Chapter 2 Formulas
2.5: Chapter 2 - Key Terms and Symbols

This page titled 2: Frequency Distributions, and Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

2.1: Qualitative Data

Learning Objectives

- Explore qualitative (categorical) data representing non-numerical categories such as colors, brands, or types.
- Use bar graphs, Pareto charts, and pie graphs to visualize categorical data.
- Highlight frequencies, proportions, and key patterns within the categories for clearer analysis

Remember, qualitative data are words describing an individual's characteristics. Several different graphs can be used to represent qualitative data, including bar graphs, Pareto charts, and pie charts.

Pie charts and bar graphs are the most common ways of displaying qualitative data. A spreadsheet program like Microsoft (MS) Excel can create both graphs. The first step for both graphs is to generate a **frequency or relative frequency table**. A frequency table summarizes the data and lists the number of occurrences of each data type.

Categorical Frequency Distribution

Example 2.1.1

Suppose you have the following raw data for different car types for students parked in a college parking lot.

Ford, Chevy, Honda, Toyota, Other, Toyota, Nissan, Nissan, Chevy, Toyota, Honda, Chevy, Toyota, Nissan, Ford, Toyota, Other, Nissan, Chevy, Ford, Nissan, Toyota, Nissan, Ford, Chevy, Toyota, Nissan, Honda, Other, Chevy, Chevy, Honda, Toyota, Chevy, Ford, Nissan, Other, Toyota, Chevy, Honda, Chevy, Toyota, Chevy, Chevy, Nissan, Honda, Toyota, Toyota, Nissan, Other

Solution

A list of data is too hard to look at and analyze, so you need to summarize it. First, you need to decide the categories. Second, list the car types in the first column. Third, for the values in the frequency column, we count the number of cars per car type in the list. For example, there are 5 Fords, 12 Chevys, 6 Hondas, 12 Toyotas, 10 Nissans, and 5 Others. Finally, the total of the frequency column should be the number of observations in the data by adding all the frequency values. The total number of data values is denoted as n. In this example, n = 50. Based on the explanation above, the frequency distribution will look like the table below.

Table 2.1.1: Frequency Table for Type of Car Data

Category	Frequency
Ford	5
Chevy	12
Honda	6
Toyota	12
Nissan	10
Other	5
Total	50

A relative frequency is a percentage that is expressed in decimal form. Since raw data values are not as useful in describing data, it is better to create a third column that lists the relative frequency of each category. The relative frequency per category is the frequency divided by the total. For example, the Ford category is computed as follows:

$$\text{relative frequency} = \frac{5}{50} = 0.10$$

This can be written as a decimal, fraction, or percent. Based on the explanation above, the frequency distribution with a relative frequency column should look like the table below.

Table 2.1.2: Relative Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

The relative frequency column should add up to 1.00. It might be off a little due to rounding.

We can use the frequency distribution to display the data using different types of graphs. These graphs include the bar graph, the pie graph, and the Pareto chart.

Bar Graphs

Bar graphs consist of the frequencies on the y-axis (vertical axis) and the categories on the x-axis (horizontal axis). For each category, draw a bar with a height equal to each frequency. All bars should have the same width, and the spaces between them should be the same.

Example 2.1.2 drawing a bar graph

Draw a bar graph of the data in Example 2.1.1.

Solution

Table 2.1.2: Relative Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

To construct a bar graph using frequencies:

1. Put the frequency scales on the y-axis and the category on the x-axis.
2. Draw a bar above each category with a height equal to its frequency.

3. Label the y-axis, x-axis, and the graph using appropriate titles.

For this example, if the steps are done correctly, you should have a graph similar to the one below.

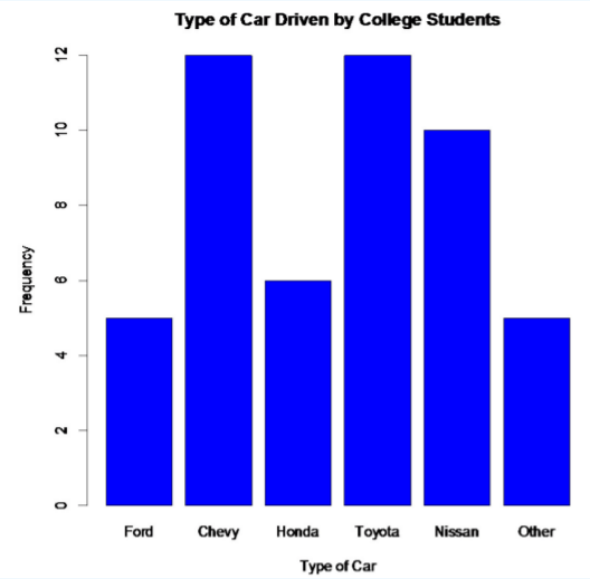


Figure 2.1.1: The Bar Graph for Type of Car Data Using Frequencies.

Notice from the graph, you can see that Toyota and Chevy are the most popular cars, with Nissan not far behind. Ford seems to be the type of car that is least liked; the cars labeled as others would be liked less than a Ford.

Some key features of a bar graph:

- Equal spacing between bars.
- Bars are the same width.
- There should be labels on each axis and a title for the graph.
- There should be a scaling on the y-axis, and the categories should be listed on the x-axis.
- The bars don't touch.

You can also draw a bar graph using relative frequency on the y-axis. This is useful when comparing two samples with different sample sizes because it is better to compare by percentages/relative frequencies to see which categories are the most common. The relative frequency graph and the frequency graph should look the same, except for the scaling on the y-axis.

To construct a bar graph using relative frequencies:

1. Put the relative frequency scales on the y-axis and the category on the x-axis.
2. Draw a bar above each category with a height equal to its frequency.
3. Label the y-axis, x-axis, and the graph using appropriate titles.

For this example, if the steps are done correctly, you should have a graph similar to the one below.

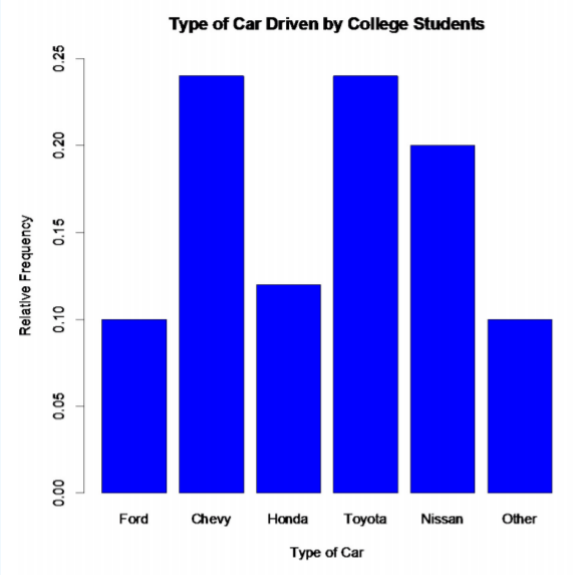


Figure 2.1.2: Bar Graph for Type of Car Data Using Relative Frequencies

Pareto chart

A second type of categorical data graph is a **Pareto chart**, which is a bar graph with the bars sorted from the highest frequency to the lowest frequency, starting from the left. It is used to represent the vital few items. In this case, it will be the most popular car types. These are Chevy and Toyota. This is especially useful in business applications, where you want to know what services your customers like the most, what processes result in more injuries, which issues employees find more important, and other types of questions like these. Here is the Pareto chart for the data in Example 2.1.1.

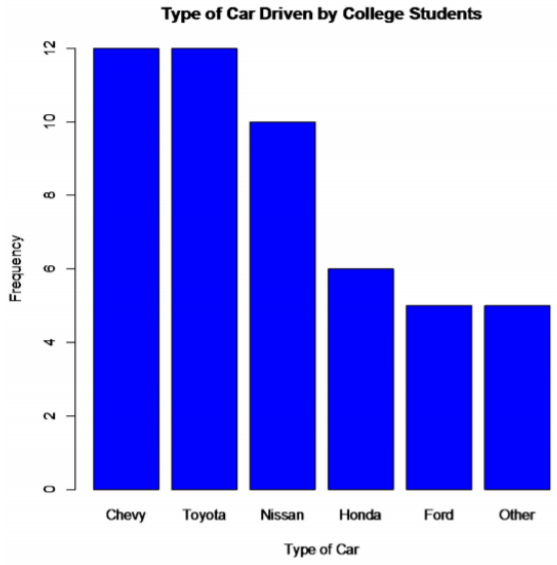


Figure 2.1.3: Pareto Chart for Type of Car Data

Pie Chart

The **pie chart** is a circle divided into sections according to the percentage of frequencies in each category. The formula for these percentages is $P = f / n * 100\%$. Where P stands for percent, f for the frequency of each class, and n for the sum of all the frequencies. To draw each section, use the percentages (recall that a quarter of the circle equals 25%).

Example 2.1.1

A random extended family has been selected, and their family members are documented and listed according to the age group below.

Age of Family Members

Categories	
Children	10
Young Adults	12
Middle Age	19
Elderly	8

Table 2.1.5: Age of Family Members

Solution

Guidelines for Creating a Pie Chart

1. Compute the relative frequencies by dividing each frequency by the total frequency.
2. Multiply each relative frequency by 100 to get the percentages.
3. Draw a circle.
4. Draw and label each section according to the percentage of each category.

Age of Family Members

Categories	Frequency	Relative Frequency (Rounded to Two Decimal Places)	Percent
Children	10	10/49 ≈ 0.20	0.20 · 100 = 20%
Young Adults	12	12/49 ≈ 0.24	0.24 · 100 = 24%
Middle Age	19	19/49 ≈ 0.39	0.39 · 100 = 39%
Elderly	8	8/49 ≈ 0.16	0.16 · 100 = 16%

Table 2.1.6: Age of Family Members With Relative Frequencies and Percents

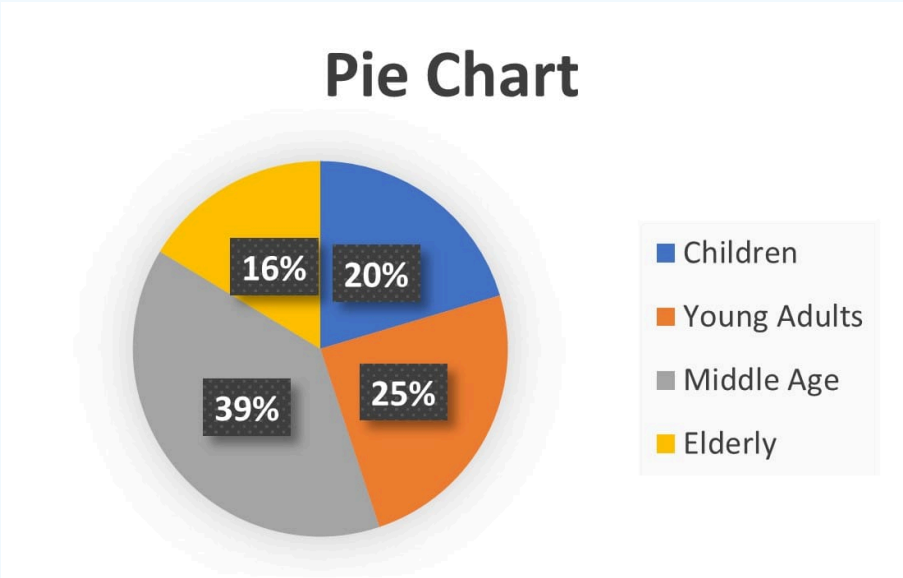


Figure 2.1.10: Pie Chart of Ages of Family Members in Percents.

We have a pie chart with four sections. The children section is represented by 20%. The young adults section is represented by 25%. The middle age section is represented by 39%. Finally, the elderly section is represented by 16%. The size of the slices is based on the percentages.

Example 2.1.3 drawing a pie chart

Draw a pie chart of the data in Example 2.1.1.

First, you need the relative frequencies.

Table 2.1.2: Relative Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

Second, multiply each relative frequency by 360° to obtain the angle measure for each category.

Table 2.1.3: Pie Graph Angles for Type of Car Data

Category	Relative Frequency	Angle (in degrees (°))
Ford	0.10	36.0
Chevy	0.24	86.4
Honda	0.12	43.2
Toyota	0.24	86.4
Nissan	0.20	72.0
Other	0.10	36.0
Total	1.00	360.0

Now draw the pie graph using a compass, protractor, and straight edge. Technology is preferred. If you use technology, there is no need for the relative frequencies or the angles.

The pie graph for this example should be like the one below.

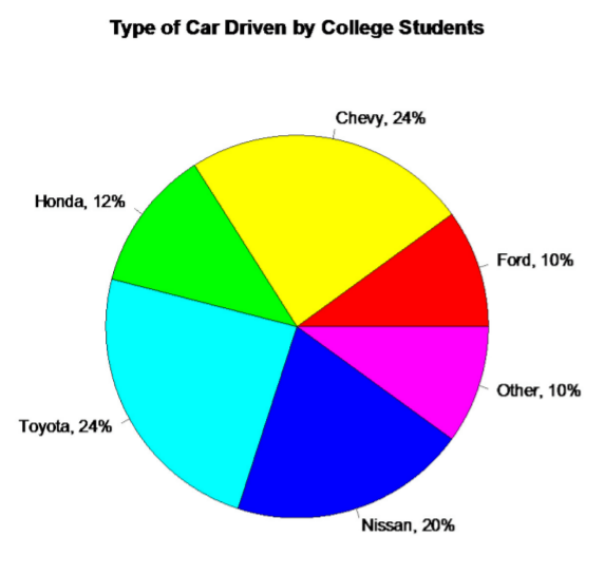


Figure 2.1.4: Pie Chart for Type of Car Data

As you can see from the graph, Toyota and Chevy are the most popular car types, while the cars labeled other are liked the least. Based on the car types that are known, Ford is the least common one in the sample.

Authors

"2.1: Qualitative Data" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"2.1: Qualitative Data" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 2.1: Qualitative Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 2.1: Qualitative Data by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.
- 1.1: What is Statistics? by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

2.2: Quantitative Data

Learning Objectives

- Organize data into intervals using grouped frequency distributions to simplify large data sets.
- Visualize the organized data using histograms, frequency polygons, and ogives.
- Identify patterns, trends, and cumulative totals to support better data interpretation.

There are two ways of organizing quantitative data: group frequency distribution and cumulative frequency distribution. These distributions are represented by the following graphs: histogram, frequency polygon, and ogive. The graphs help describe the shape of the distribution, and the shape determines what type of analysis can be conducted.

The histogram is similar to a bar graph, except for some major differences. The first difference from a bar graph, the categories can be put in any order on the horizontal axis. There is no set order for these data values. You can't say how the data is distributed based on the shape since the shape can change by putting the categories in different orders. With the histogram, the data are in a specific order because the data are expressed with numbers. With quantitative data, you can talk about a distribution since the shape only changes slightly depending on how many categories you set up.

This leads to the second difference from bar graphs. In a bar graph, the categories that were made in the frequency table were determined by the whims of the researcher. In quantitative data, the groups are numerical intervals, and the frequencies are determined by the number of data values in each group (or class).

The third difference in the histogram is that the bars must touch, and there is no gap between them. The reason that bar graphs have gaps is that the data are categorical and not numeric. To create a histogram, you must first create the frequency distribution. The idea of a frequency distribution is to take the interval that the data spans and divide it up into equal subintervals called classes.

Frequency Distribution

Summary of the Steps Involved in Making a Frequency Distribution

1. Find the Range = Largest value – Smallest value
2. Pick the number of classes to use, or the number will be provided. Usually, the number of classes is between five and twenty. Five classes are used if there are a small number of data point,s and twenty classes if there are a large number of data points (over 1000 data points). (Note: categories will now be called classes from now on.)
3. Class width = $\frac{\text{Range}}{\#\text{ Classes}}$. Always round up to the next integer (even if the answer is already a whole number, go to the next integer). If you don't do this, your last class will not contain your largest data value, and the last data value will have no class to go to. If you round up, then your largest data value will fall in the last class.
4. Create the classes. Each class has limits that determine which values fall in each class. To find the class limits, set the smallest value as the lower class limit for the first class. Then add the class width to the lower class limit to get the next lower class limit. Repeat until you get all the classes. The upper-class limit for a class is one less than the lower limit for the next class.
5. For the classes to touch, one class needs to start where the previous one ends. This is known as the class boundary. To find the class boundaries, subtract 0.5 from the lower-class limit and add 0.5 to the upper-class limit.
6. Sometimes it is useful to find the class midpoint. The process is
   Midpoint $= \frac{\text{Lower limit} + \text{Upper limit}}{2}$
7. To determine the number of data points that fall in each class, count how many are within the range, including the lower and upper limits. Utilizing tally marks may help count the data values. The frequency for a class is the number of data values that fall in the class.

Note

The above description is for data values that are whole numbers. If your data value has decimal places, then your class width should be rounded up to the nearest value with the same number of decimal places as the original data. In addition, your class boundaries should have one more decimal place than the original data. For example, if your data has one decimal place, then the class width would have one decimal place, and the class boundaries are formed by adding and subtracting 0.05 from each class limit.

Example 2.2.1 creating a frequency table

The table contains the monthly rent for 24 students from a statistics course. Create a frequency distribution using 7 classes.

Table 2.2.1: Data of Monthly Rent

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

Solution

1. Find the range:
   Largest value - Smallest value $= 2550 - 350 = 2200$
2. Pick the number of classes:
   The direction says to use 7 classes.
3. Find the class width:
   Class Width $= \frac{\text{Range}}{7} = \frac{2200}{7} \approx 314.286$
   Round up to 315
   Always round up to the next integer, even if the width is already an integer.
4. Find the class limits:
   Start at the smallest value. This is the lower class limit for the first class. Add the width to get the lower limit of the next class. Keep adding the width to get all the lower limits.
   First Lower-Limit $= 350$ (Typically the lowest value in the data),
   Second Lower-Limit $= 350 + 315 = 665$,
   Third Lower-Limit $= 665 + 315 = 980$,
   Fourth Lower-Limit $= 980 + 315 = 1295$
   The upper limit is one less than the next lower limit: so for the first class, the upper class limit would be $665 - 1 = 664$. To get the rest of the upper-class limits, repeatedly add the class limit to the previous upper limit.
   First Upper-Limit $= 664$,
   Second Upper-Limit $= 664 + 315 = 979$,
   Third Upper-Limit $= 979 + 315 = 1294$,
   Fourth Upper-Limit $= 1294 + 315 = 1609$
   When you have all 7 classes, make sure the last number, in this case the 2550, is at least as large as the largest value in the data. If not, you made a mistake somewhere.
5. Find the class boundaries: Subtract 0.5 from the lower-class limit to get the lower-class boundary.
   First Lower-Class Boundary $= 350 - 0.5 = 349.5$,
   Second Lower-Class Boundary $= 665 - 0.5 = 664.5$,
   Third Lower-Class Boundary $= 980 - 0.5 = 979.5$,
   Fourth Lower-Class Boundary $= 1295 - 0.5 = 1294.5$
   Add 0.5 to the upper-class limit to get the upper-class boundary.
   First Upper-Class Boundary $= 664 + 0.5 = 664.5$,
   Second Upper-Class Boundary $= 979 + 0.5 = 979.5$,
   Third Upper-Class Boundary $= 1294 + 0.5 = 1294.5$,

Fourth Upper-Class Boundary $= 1609 + 0.5 = 1609.5$
Every value in the data should fall into exactly one of the classes. No data values should fall right on the boundary of two classes.

6. Find the class midpoints:
First Midpoint $= \frac{\text{Lower limit } + \text{ Upper limit}}{2} = \frac{350 + 664}{2} = 507,$
Second Midpoint $= \frac{665 + 979}{2} = 822$
7. Tally and find the frequency of the data:
Go through the data and put a tally mark in the appropriate class for each piece of data by looking to see which class boundaries the data value is between. Fill in the frequency by changing each of the tallies into a number. Be sure the total of the frequencies is the same as the number of data points.

Table 2.2.2: Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency
350-664	349.5-664.5	507	IIII	4
665-979	664.5-979.5	822	𝍸 III	8
980-1294	979.5-1294.5	1137	𝍸	5
1295-1609	1294.5-1609.5	1452	𝍸 I	6
1610-1924	1609.5-1924.5	1767		0
1925-2239	1924.5-2239.5	2082		0
2240-2554	2239.5-2554.5	2397	I	1

Histogram

It is difficult to determine the basic shape of the distribution by looking at the frequency distribution. It would be easier to look at a graph. The graph of a frequency distribution for quantitative data is called a **frequency histogram** or histogram for short.

Definition 2.2.1: Histogram

A Histogram is a graph of the frequencies on the vertical axis and the class boundaries on the horizontal axis. Rectangles where the height is the frequency and the width is the class width are drawn for each class.

Example \({2{.}2{.}\eatSpaces{2} \endSpaces}\: Drawing a Histogram

Draw a histogram for the distribution from Example 2.2.1.

Solution

The class boundaries are plotted on the horizontal axis, and the frequencies are plotted on the vertical axis. You can plot the midpoints of the classes instead of the class boundaries. *Graph 2.2.1* was created using the midpoints because it was easier to do with the software that created the graph.

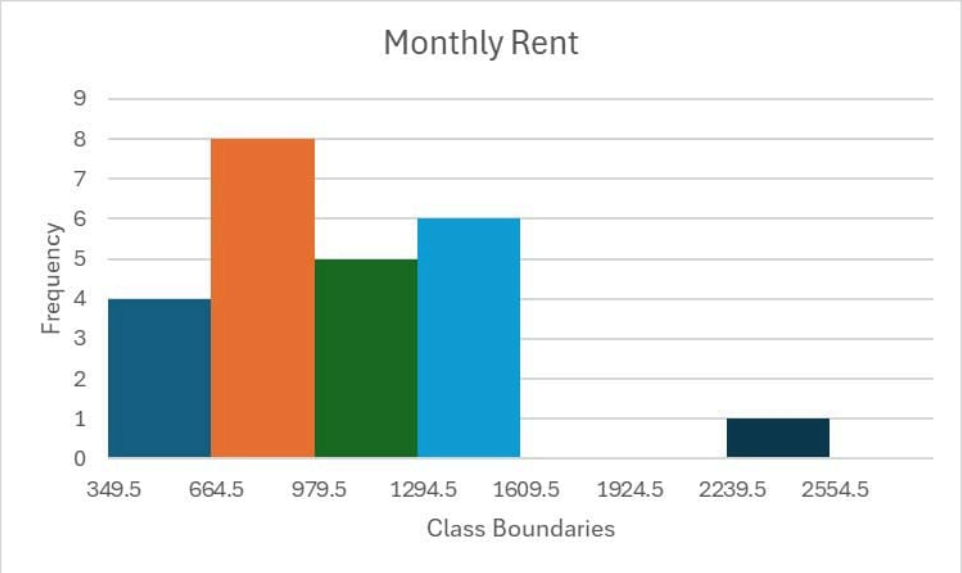


Figure 2.2.1: Histogram for Monthly Rent

Reviewing the graph, you can see that most of the students pay around $750 per month for rent, with about $1500 being the other common value. You can see from the graph that most students pay between $600 and $1600 per month for rent. Of course, these values are just estimates from the graph. There is a large gap between the $1500 class and the highest data value. This seems to say that one student is paying a great deal more than everyone else. This value could be considered an outlier. An **outlier** is a data value that is far from the rest of the values. It may be an unusual value or a mistake. It is a data value that should be investigated. In this case, the student lives in a very expensive part of town, thus, the value is not a mistake and is just very unusual. Other aspects can be discussed, but first, some other concepts need to be introduced.

Frequencies are helpful, but understanding the relative size each class is to the total is also useful. To find this, you can divide the frequency by the total to create a relative frequency. If you have the relative frequencies for all of the classes, then you have a relative frequency distribution.

Definition 2.2.2

Relative Frequency Distribution

A variation on a frequency distribution is a relative frequency distribution. Instead of giving the frequencies for each class, the relative frequencies are calculated.

$$\text{Relative frequency} = \frac{\text{frequency}}{\#\text{ of data points}}$$

This gives you percentages of data that fall in each class.

Example 2.2.3 creating a relative frequency table

Find the relative frequency for the grade data.

Solution

From Example 2.2.1, the frequency distribution is reproduced in Example 2.2.2.

Table 2.2.2: Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency
350-664	349.5-664.5	507	4
665-979	664.5-979.5	822	8
980-1294	979.5-1294.5	1127	5
1295-1609	1294.5-1609.5	1452	6
1610-1924	1609.5-1924.5	1767	0
1925-2239	1924.5-2239.5	2082	0
2240-2554	2239.5-2554.5	2397	1

Divide each frequency by the number of data points.

$$\frac{4}{24} = 0.17, \frac{8}{24} = 0.33, \frac{5}{24} = 0.21, \rightleftharpoons$$

Table 2.2.3: Relative Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency	Relative Frequency
350-664	349.5-664.5	507	4	0.17
665-979	664.5-979.5	822	8	0.33
980-1294	979.5-1294.5	1127	5	0.21
1295-1609	1294.5-1609.5	1452	6	0.25
1610-1924	1609.5-1924.5	1767	0	0
1925-2239	1924.5-2239.5	2082	0	0
2240-2554	2239.5-2554.5	2397	1	0.04
Total			24	1

The relative frequencies should add up to 1 or 100%. (This might be off a little due to rounding errors.)

The graph of the relative frequency is known as a relative frequency histogram. It looks identical to the frequency histogram, but the vertical axis is relative frequency instead of just frequencies.

Example 2.2.4 drawing a relative frequency histogram

Draw a relative frequency histogram for the grade distribution from Example 2.2.1.

Solution

The class boundaries are plotted on the horizontal axis, and the relative frequencies are plotted on the vertical axis. (This is not easy to do in R, so use another technology to graph a relative frequency histogram.)



Figure 2.2.2: Relative Frequency Histogram for Monthly Rent

Notice the shape is the same as the frequency distribution.

Frequency Polygon

The **frequency polygon** is used to help understand the shape of a frequency distribution. It is typically graphed by connecting straight lines to points plotted for the frequencies at the midpoints. The midpoints are computed by adding each pair of class limits (or class boundaries) and then dividing by 2. Also, the first and last points are anchored to the x-axis, where hypothetical midpoints would be located. The first anchor is the first midpoint value minus the class width, and the last anchor is the last midpoint value plus the class width.

Example 2.2.5

Create a frequency polygon for the data in Example 2.2.2. The group frequency distribution of this example is provided below.

Grouped Frequency Distribution.

Class Limits	Class Boundaries	Class Midpoint	Frequency
350-664	349.5-664.5	507	4
665-979	664.5-979.5	822	8
980-1294	979.5-1294.5	1127	5
1295-1609	1294.5-1609.5	1452	6
1610-1924	1609.5-1924.5	1767	0
1925-2239	1924.5-2239.5	2082	0
2240-2554	2239.5-2554.5	2397	1
Total			24

Solution

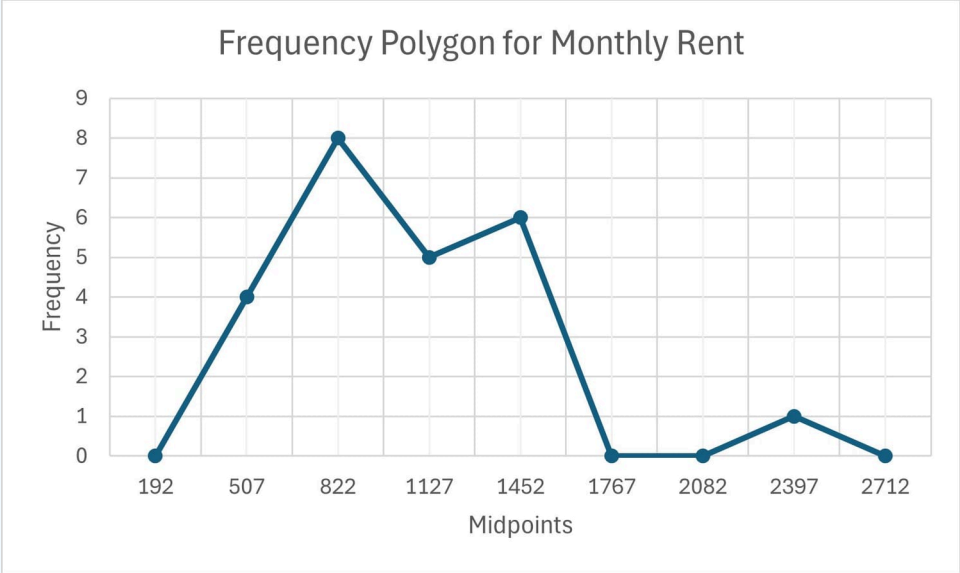


Figure 2.2.3: Frequency Polygon for Monthly Rent

Cumulative Frequency Distribution

Another useful piece of information is how many data points fall below a particular class boundary. As an example, a teacher may want to know how many students received below 80%, a doctor may want to know how many adults have cholesterol below 160, or a manager may want to know how many stores gross less than $2000 per day. This is known as a **cumulative frequency**. If you want to know what percent of the data falls below a certain class boundary, then this would be a **cumulative relative frequency**. For cumulative frequencies, you are finding how many data values fall below the upper-class limit.

To create a **cumulative frequency distribution**, count the number of data points that are below the upper-class boundary, starting with the first class and working up to the top class. The last upper-class boundary should have all of the data points below it. Also include the number of data points below the lowest class boundary, which is zero.

Example 2.2.6 creating a cumulative frequency distribution

Create a cumulative frequency distribution for the data in Example 2.2.1.

Solution

The frequency distribution for the data is in Example 2.2.2.

Table 2.2.2: Frequency Distribution for Monthly Rent

Class Boundaries	Class Boundaries	Class Midpoint	Frequency
350-664	349.5-664.5	507	4
665-979	664.5-979.5	822	8
980-1294	979.5-1294.5	1127	5
1295-1609	1294.5-1609.5	1452	6
1610-1924	1609.5-1924.5	1767	0
1925-2239	1924.5-2239.5	2082	0
2240-2554	2239.5-2554.5	2397	1

Now, ask yourself how many data points fall below each class boundary. Below 349.5, there are 0 data points. Below 664.5, there are 4 data points, below 979.5, there are 4 + 8 = 12 data points, below 1294.5, there are 4 + 8 + 5 = 17 data points, and continue this process until you reach the upper class boundary. To find the cumulative relative frequency values, you divide the cumulative frequency by the total amount of data. For this example, the total is 24. This is summarized in Example 2.2.4.

Table 2.2.4: Cumulative Distribution for Monthly Rent

Class Boundaries	Cumulative Frequency	Cumulative Relative Frequency
< (less than) 349.5	0	0
< 664.5	4	0.17
< 979.5	12	0.5
< 1294.5	17	0.71
< 1609.5	23	0.96
< 1924.5	23	0.96
< 2239.5	23	0.96
< 2554.5	24	1

Ogive

Again, it is hard to look at the data the way it is. A graph would be useful. The graph for cumulative frequency is called an **ogive** (o-jive). To create an ogive, first create a scale on both the x and y axes that will fit the data. Then plot the points of the class boundary values versus the cumulative frequencies from left to right. Finally, connect the plotted points with straight lines.

Example 2.2.7 drawing an ogive

Draw an ogive for the data in Example 2.2.1.

Solution

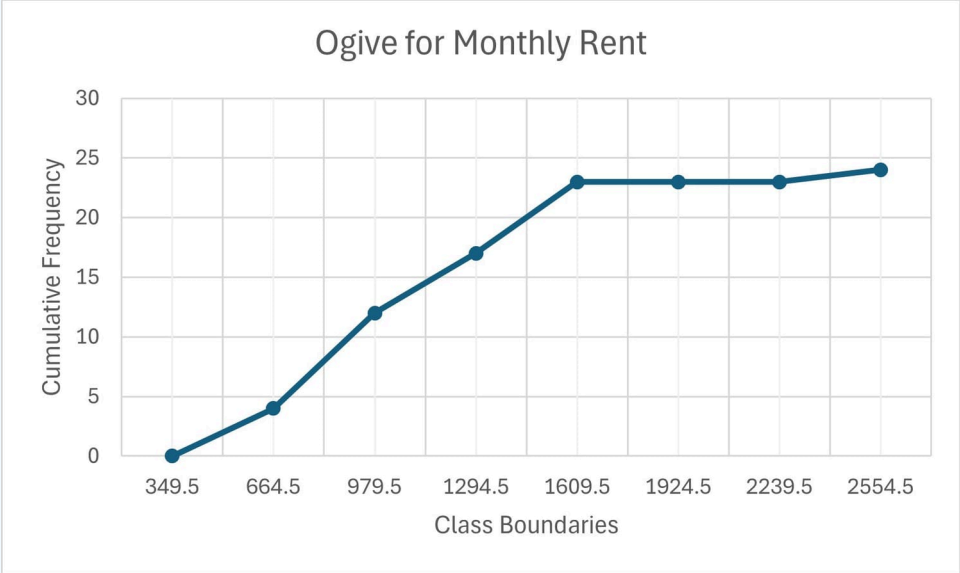


Figure 2.2.4: Ogive for Monthly Rent

Shapes of the Distribution

When you look at a distribution, look at the basic shape. Some basic shapes are seen in histograms. Realize, though that some distributions have no shape. The common shapes are symmetric, skewed, and uniform. Another interest is how many peaks a graph may have. This is known as modal.

Symmetrical means that you can fold the graph in half down the middle, and the two sides will line up. You can think of the two sides as being mirror images of each other. Skewed means one “tail” of the graph is longer than the other. The graph is skewed in the direction of the longer tail (backward from what you would expect). A uniform graph has all the bars at the same height.

Modal refers to the number of peaks. Unimodal has one peak, and bimodal has two peaks. Usually, if a graph has more than two peaks, the modal information is no longer of interest.

Other important features to consider are gaps between bars, a repetitive pattern, how spread out the data is, and where the center of the graph is.

Examples of Graphs:

This graph is roughly symmetric and unimodal:

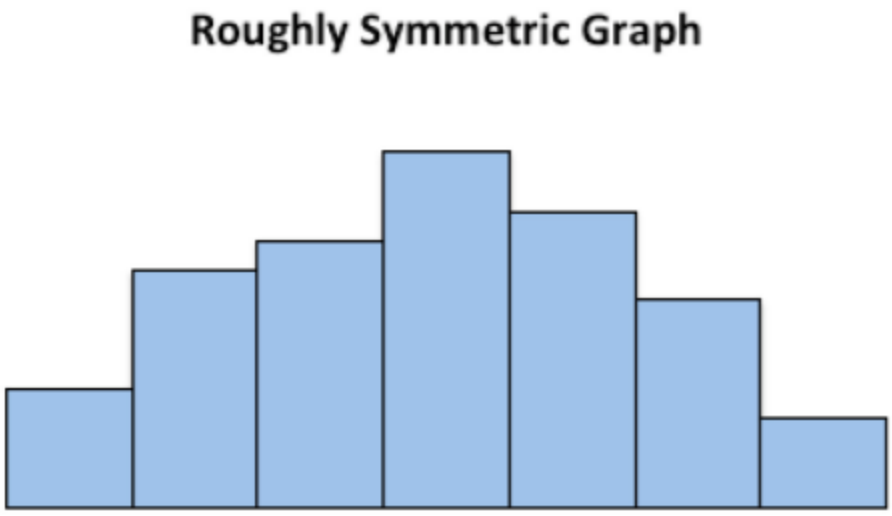


Figure 2.2.5: Graph of a Unimodal and Symmetrical Distribution

This graph is symmetric and bimodal:

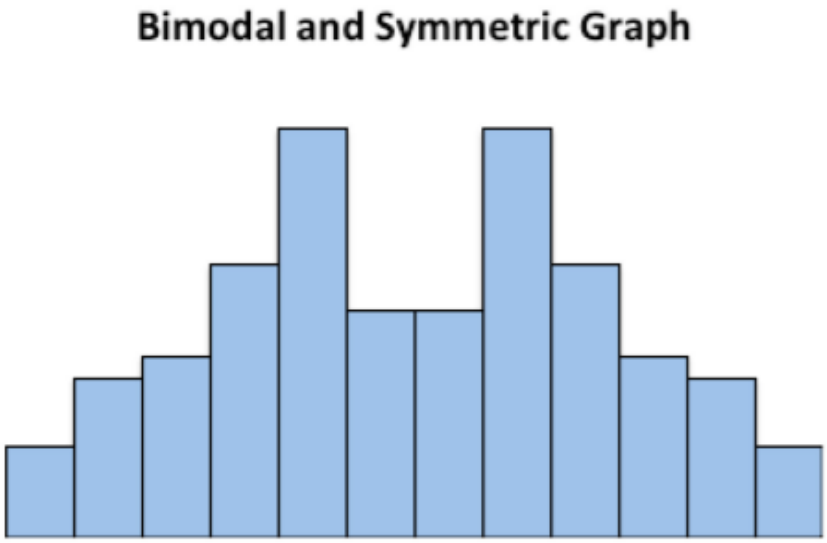


Figure 2.2.6: Graph of a Bimodal and Symmetrical Distribution

This graph is skewed to the right:

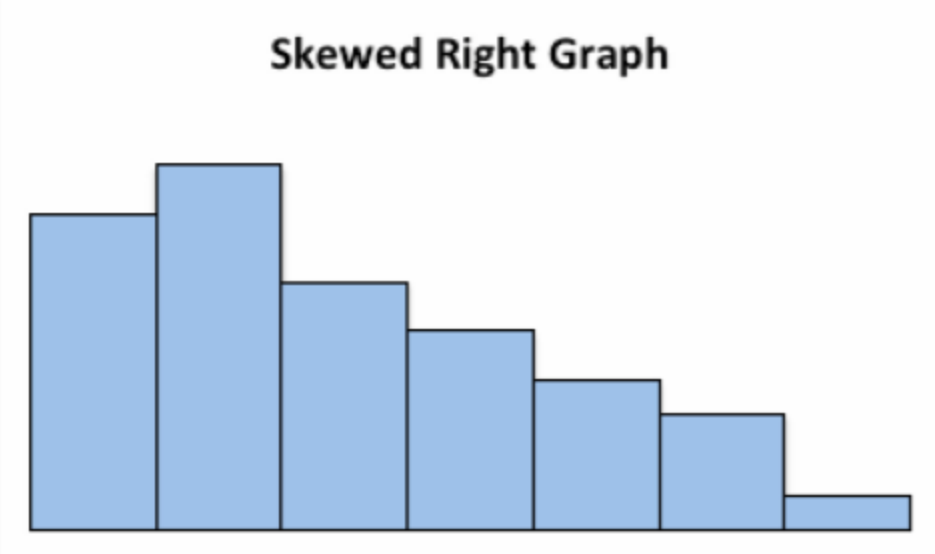


Figure 2.2.7: Graph of a Right (Positive) Skewed Distribution

This graph is skewed to the left and has a gap:

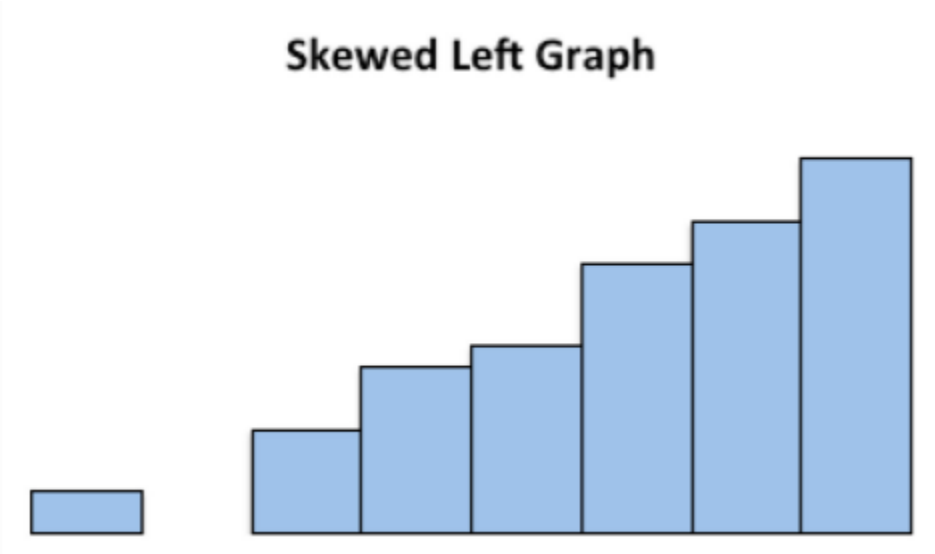


Figure 2.2.8: Graph of a Left (Negative) Skewed Distribution

This graph is uniform since all the bars are the same height:

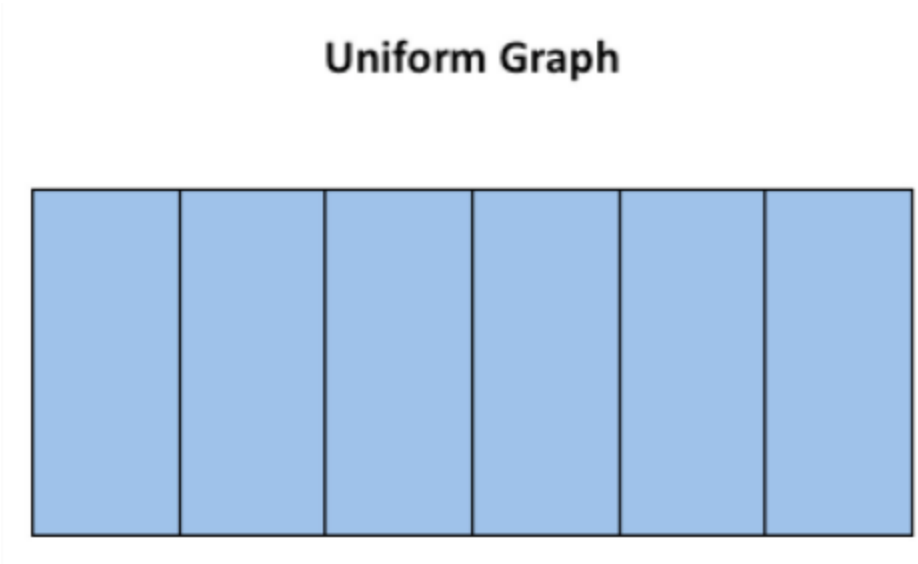


Figure 2.2.9: Graph of a Uniform Distribution

There are other types of graphs for quantitative data. They will be explored in the next section.

Authors

"2.2: Quantitative Data" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"2.2: Quantitative Data" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 2.2: Quantitative Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 2.2: Quantitative Data by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.
- 1.1: What is Statistics? by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

2.3: Other Graphical Representations of Data

Learning Objectives

- Introduction of additional statistical graphs to illustrate data structure and relationships.
- Utilize stem-and-leaf plots, scatter plots, dot plots, and time series graphs for data visualization
- Reveal patterns, trends, and distributions across different types of data

There are many other types of graphs. Some of the more common ones are the dot plot, the stem-and-leaf plot, the scatter plot, and the time-series plot. Many different graphs have emerged lately for qualitative data. Many are found in publications and websites. The following is a description of the stem-and-leaf plot, the scatter plot, and the time-series plot.

Stem-and-Leaf Plots

Stem-and-leaf plots are a quick and easy way to look at small samples of numerical data. You can look for any patterns or any strange data values. It is easy to compare two samples using stem-and-leaf plots.

The first step is to divide each number into 2 parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit). There are no set rules, you just have to look at the data and see what makes sense.

Example 2.3.1 stem plot for grade distribution

The following are the percentage grades of 25 students from a statistics course. Draw a stem-and-leaf plot of the data.

Table 2.3.1: Data of Test Grades

62	87	81	69	87	62	45	95	76	76
62	71	65	67	72	80	40	77	87	58
84	73	93	64	89					

Solution

Divide each number so that the tens digit is the stem and the ones digit is the leaf. 62 becomes 6|2.

Make a vertical chart with the stems on the left of a vertical bar. Be sure to fill in any missing stems. In other words, the stems should have equal spacing (for example, count by ones or count by tens). *Graph 2.3.1* shows the stems for this example.

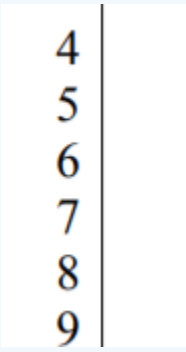


Figure 2.3.1: Stem Plot for Test Grades Step 1

Now go through the list of data and add the leaves. Put each leaf next to its corresponding stem. Don't worry about the order yet just get all the leaves down.

When the data value 62 is placed on the plot, it looks like the plot in *Graph 2.3.2*.

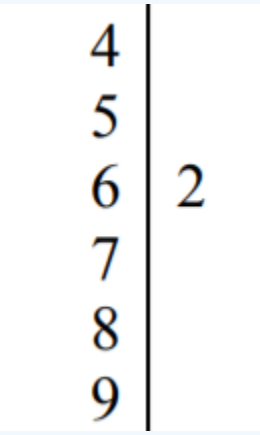


Figure 2.3.2: Stem Plot for Test Grades Step 2

When the data value 87 is placed on the plot, it looks like the plot in *Graph 2.3.3*.

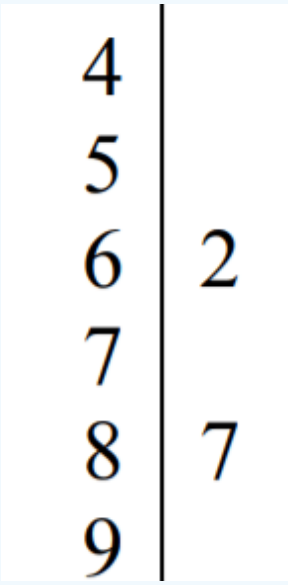


Figure 2.3.3: Stem Plot for Test Grades Step 3

Filling in the rest of the leaves to obtain the plot in *Graph 2.3.4*.

4	5 0
5	8
6	2 9 2 2 5 7 4
7	6 6 1 2 7 3
8	7 1 7 0 7 4 9
9	5 3

Figure 2.3.4: Stem Plot for Test Grades Step 4

Now you have to add labels and make the graph look pretty. You need to add a label and sort the leaves into increasing order. You also need to tell people what the stems and leaves mean by inserting a legend. **Be careful to line the leaves up in columns**. You need to be able to compare the lengths of the rows when you interpret the graph. The final stem plot for the test grade data is in *Graph 2.3.5*.

Test Scores

4	0 5
5	8
6	2 2 2 4 5 7 9
7	1 2 3 6 6 7
8	0 1 4 7 7 7 9
9	3 5

Figure 2.3.5: Stem Plot for Test Grades

Now you can interpret the stem-and-leaf display. The data is bimodal and somewhat symmetric. There are no gaps in the data. The center of the distribution is around 70.

Scatter Plot

Sometimes you have two different variables and you want to see if they are related in any way. A scatter plot helps you to see what the relationship would look like. A scatter plot is just a plotting of the ordered pairs.

Example 2.3.2 scatter plot

Is there any relationship between elevation and high temperature on a given day? The following data are the high temperatures at various cities on a single day and the elevation of the city.

Table 2.3.2: Data of Temperature versus Elevation

Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (°F)	50	60	48	70	55	55	60

Solution

Preliminary: State the random variables

Let x = altitude

y = high temperature

Now plot the x values on the horizontal axis, and the y values on the vertical axis. Then set up a scale that fits the data on each axis. Once that is done, then just plot the x and y values as an ordered pair.

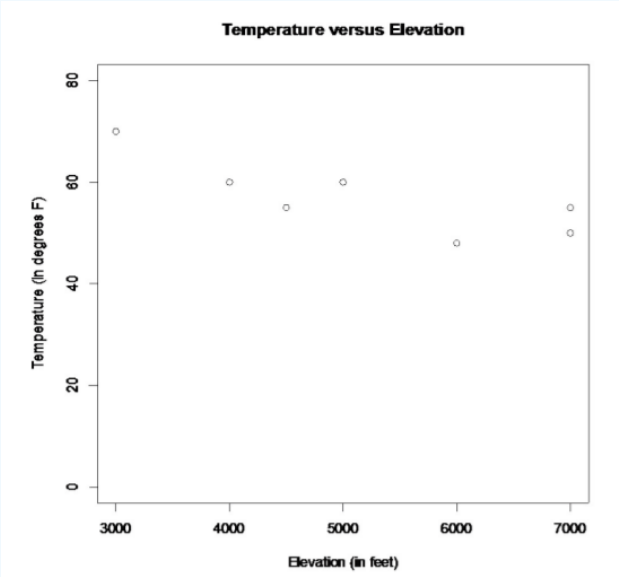


Figure 2.3.6: Scatter Plot of Temperature versus Elevation

Looking at the graph, it appears that there is a linear relationship between temperature and elevation. It also appears to be a negative relationship; thus, as elevation increases, the temperature decreases.

Time-Series Graph

A time-series plot is a graph showing the data measurements in chronological order, the data being quantitative. For example, a time-series plot is used to show profits over the last 5 years. To create a time-series plot, the time always goes on the horizontal axis, and the other variable goes on the vertical axis. Then plot the ordered pairs and connect the dots. The purpose of a time-series graph is to look for trends over time. Caution, you must realize that the trend may not continue. Just because you see an increase, doesn't mean the increase will continue forever. As an example, before 2007, many people noticed that housing prices were increasing. The belief at the time was that housing prices would continue to increase. However, the housing bubble burst in 2007, and many houses lost value and haven't recovered.

Example 2.3.3 Time-series plot

The following table tracks the weight of a dieter, where the time in months is measuring how long since the person started the diet

Table 2.3.3: Data of Weights versus Time

Time (months)	0	1	2	3	4	5
Weight (pounds)	200	195	192	193	190	187

Make a time-series plot of this data

Solution

Guidelines for creating a time-series graph

1. Put the title.
2. Create and label the x and y-axes.
3. Plot the points using time as the x-coordinate, and weight as the y-coordinate.
4. Connect the points with straight lines.

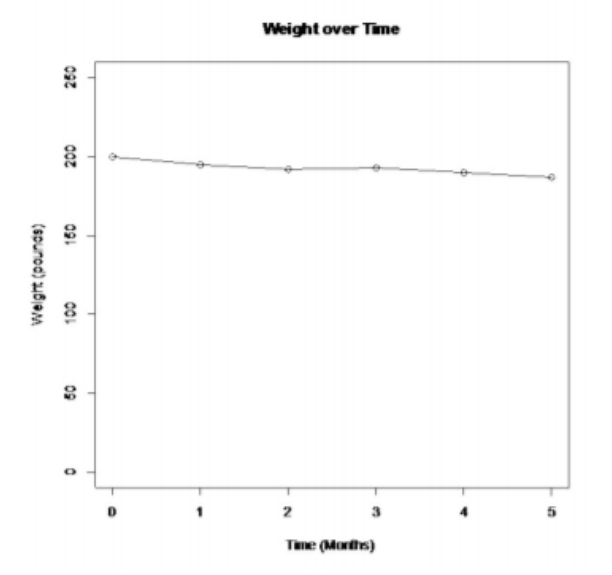


Figure 2.3.7: Time Series Graph Weights Versus Time

Notice, that over the 5 months, the weight appears to be decreasing. However, it doesn't look like there is a large decrease.

Be careful when making a graph. If you don't start the vertical axis at 0, then the change can look much more dramatic than it really is. As an example, *Graph 2.3.8* shows *Graph 2.3.7* with a different scaling on the vertical axis. Notice the decrease in weight looks much larger than it really is.

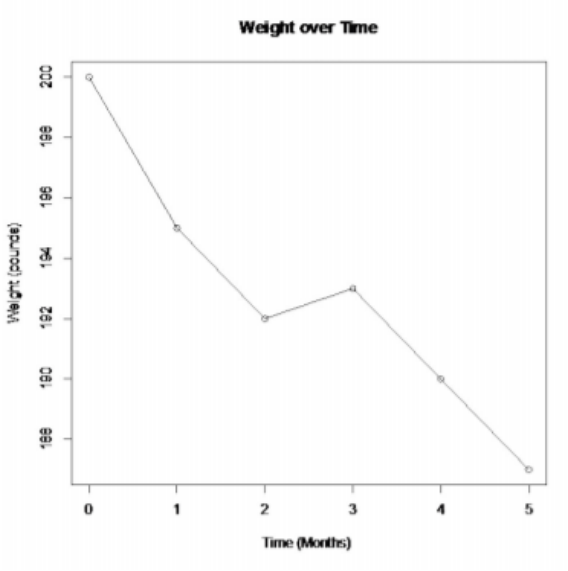


Figure 2.3.8: Time Series Graph Weights Versus Time With Error

Dot Plot

A **dot plot** is used to analyze the frequencies of a data set in a simplified manner. Each data value is plotted on the number line, and dots are stacked above the value to represent the frequencies.

Example 2.3.1

Use the data below to construct a dot plot.

Number of Pets for Each Family

5	4	2	0	4	1	2
2	0	3	0	1	2	1
2	1	1	3	4	4	5
3	2	5	2	1	1	0
4	5	3	5	5	3	4
1	1	3	2	3	4	1
3	2	5	5	2	2	2

Table 2.3.4: Number of Pets

Solution

Guidelines for Creating a Dot Plot

1. Create a title.
2. Draw a number line.
3. List the numbers from lowest to highest on the number line.
4. For each data value above, place a dot above its number on the number line. If the numbers repeat, stack them with more dots.

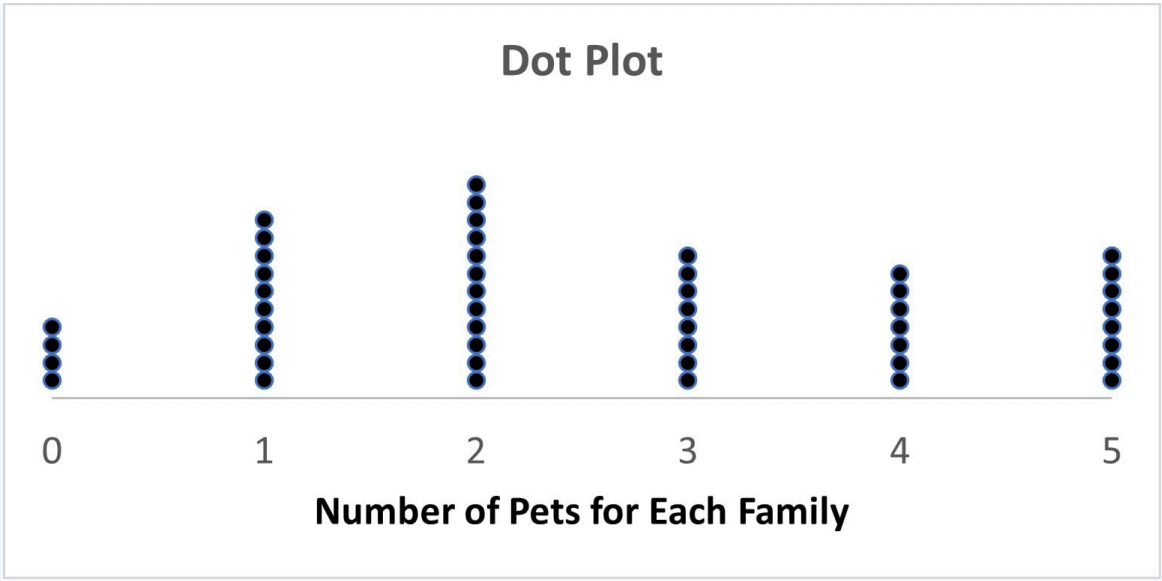


Figure 2.3.9: Dot Plot of the Number of Pets for Each Family

Authors

"2.3: Other Graphical Representations of Data" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"2.3: Other Graphical Representations of Data" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 2.3: Other Graphical Representations of Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 2.3: Other Graphical Representations of Data by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.
- 1.1: What is Statistics? by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

2.4: Chapter 2 Formulas

Chapter 2 Formulas

Class Width

The class width is the difference between two consecutive classes' lower boundaries (or lower limits) in a grouped frequency distribution. It indicates how wide each class interval is. Round up the result to the next whole number or the next odd whole number to ensure all data values are included. The class width helps ensure that the classes are evenly spaced in the frequency table.

$$\text{Class Width} = \frac{\text{Maximum Value} - \text{Minimum Value}}{\text{Number of Classes}}$$

Degrees of Sections of a Circle

The degrees formula represents how much of the full circle (360°) is allocated to each category based on its proportion of the total data. It is commonly used to compute the degrees of each section in a pie graph. This allows the pie chart to visually show how large each part is compared to the whole. A larger slice (more degrees) means a larger share of the total.

$$\text{Degrees} = \left(\frac{\text{Category Frequency}}{\text{Total Frequency}}\right) \times 360^\circ$$

Midpoints of a Class in a Frequency Distribution

In a frequency distribution, the midpoint of a class represents the central value of that class interval. It is used to summarize data in a way that reflects the center of each class. This value is used in graphing frequency polygons and calculating estimates like the mean of grouped data.

$$\text{Midpoint} = \frac{\text{Lower Class Limit} + \text{Upper Class Limit}}{2}$$

Percentage of Frequency

In categorical frequency distributions and pie graphs, percents represent the proportion of each category relative to the total number of observations, expressed as a percentage. This helps to easily compare the size of each category and is especially useful in pie charts where each slice reflects a category's percentage of the whole.

$$\text{Percent} = \left(\frac{\text{Category Frequency}}{\text{Total Frequency}}\right) \times 100\%$$

Range

In a grouped frequency distribution, the range is the difference between the highest and lowest data values. It gives a basic measure of how spread out the data is. The range helps to determine the overall span of the data set and is often used when deciding on class limits or class width.

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Relative Frequency

In a grouped frequency distribution, relative frequency is the proportion of data values that fall within a specific class interval compared to the total number of data values. It helps describe how the data is distributed across different class intervals

and allows for comparison between classes, especially when the total number of observations is large. This value can also be expressed as a decimal or a percentage, and is often used to construct relative frequency histograms.

$$\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Frequency}}$$

Relative Cumulative Frequency

Relative cumulative frequency is the running total of relative frequencies up to and including a specific class in a grouped frequency distribution. It shows the proportion of data values that fall at or below a particular class boundary. This helps in understanding how data accumulates across intervals and is useful for constructing ogives and analyzing percentiles.

$$\text{Relative Cumulative Frequency} = \frac{\text{Cumulative Frequency}}{\text{Total Frequency}}$$

Authors

"2.4: Chapter 2 Formulas" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 2.4: Chapter 2 Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

2.5: Chapter 2 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Bar graph: a chart that uses rectangular bars to show the size or amount of different categories. The length or height of each bar represents the value or frequency of that category, making it easy to compare different groups.

Dot plot: a graph that uses dots to represent individual data values along a number line. Each dot represents one occurrence of a value. If a value appears more than once, the dots are stacked vertically above that number.

Frequency polygon: a line graph that shows the distribution of a dataset. It is created by plotting points at the midpoints of each class interval and connecting them with straight lines. This graph helps to visualize the shape of the data.

Histogram: a type of bar graph that shows how data is grouped into intervals. Each bar represents the frequency of data within a specific range, and the bars touch each other to show that the data is continuous.

Ogive: a line graph of the cumulative frequency distribution.

O**utlier:** a data value far from the rest of the values.

Pareto chart: a bar graph combined with a line graph that represents categorical data in descending order of frequency or impact.

Pie chart: a circle divided into sections according to the percentage of frequencies in each category. The formula for these percentages is *P = f / n * 100%*. Where *P* stands for percent, *f* for the frequency of each class, and *n* for the sum of all the frequencies. To draw each section, use the percentages (recall that a quarter of the circle equals 25%).

Relative frequency: the proportion of times a value or category occurs compared to the total number of observations.

Scatter plot: a graph used in statistics to display the relationship between two numerical variables.

Skewness: describes the direction in which the data tails off. A distribution is **skewed right** (positively skewed) if it has a long tail on the right side, and **skewed left** (negatively skewed) if it has a long tail on the left.

Stem-and-leaf plot: a type of graph where each number is split into a **"stem"** (the leading digit or digits) and a **"leaf"** (the last digit). The stems are listed in a vertical column, and the leaves that correspond to each stem are listed in rows to the right.

Symmetry: a balanced shape where the left and right sides of the distribution are mirror images. A perfectly symmetrical distribution, like the normal distribution, has values evenly distributed around the center.

Key Symbols

f - frequency (the count of how often a value appears).

n - total number of observations (sum of all frequencies).

$\%$ - percent (used to express a part out of 100).

$\frac{f}{n} \times 100$ - formula to calculate percent.

θ - degrees (angle measure for a pie chart).

$\frac{f}{n} \times 360^\circ$ - formula to convert frequency to degrees in a pie chart.

$\frac{f}{n}$ - formula for relative frequency.

$\sum$ - summation symbol (used to indicate the sum of frequencies or other values).

Authors

"2.5: Chapter 2 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 2.5: Chapter 2 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

3: DESCRIPTIVE STATISTICS

3.1: Measures of Center

3.2: Measures of Variation Version 1

3.3: Measures of Variation Version 2

3.4: Measures of Position

3.5: Formulas for Chapter 3

3.6: Chapter 3 - Key Terms and Symbols

This page titled 3: Descriptive Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

3.1: Measures of Center

Learning Objectives

- Describe measures of center to identify the typical or central value in a data set.
- Understand and calculate the mean, median, mode, and midrange.
- Explore variations such as the grouped mean for frequency tables and the weighted mean for data with varying importance.

Mean, Median, Mode, and Midrange for Raw Data

This section focuses on measures of central tendency. Many times you are asking what to expect on average. Such as when you pick a major, you would probably ask how much you expect to earn in that field. If you are thinking of relocating to a new town, you might ask how much you can expect to pay for housing. If you are planting vegetables in the spring, you might want to know how long it will be until you can harvest. These questions, and many more, can be answered by knowing the center of the data set. There are three measures of the "center" of the data. They are the mode, median, and mean. Any of the values can be referred to as the "average."

- The **mean** is the arithmetic average of the numbers. This is the center that most people call the average, though all three – mean, median, and mode – really are averages.
- The **median** is the data value in the middle of a sorted list of data. To find it, you put the data in order, and then determine which data value is in the middle of the data set.
- The **mode** is the data value that occurs the most frequently in the data. To find it, you count how often each data value occurs, and then determine which data value occurs most often.
- The **midrange** is the sum of the least and the greatest value in the data set, divided by 2.

There are no symbols for the mode and the median, but the mean is used a great deal, and statisticians gave it a symbol. There are two symbols, one for the population parameter and one for the sample statistic. In most cases, you cannot find the population parameter, so you use the sample statistic to estimate the population parameter.

Definition 3.1.1: Population Mean

The population mean is given by

$\mu = \frac{\sum x}{N}$, pronounced mu

where

- N is the size of the population.
- x represents a data value.
- $\sum x$ means to add up all of the data values.

Definition 3.1.2: Sample Mean

Sample Mean:

$\overline{x} = \frac{\sum x}{n}$, pronounced x bar, where

- n is the size of the sample.
- x represents a data value.
- $\sum x$ means to add up all of the data values.

The value for $\overline{x}$ is used to estimate μ since μ can't be calculated in most situations.

Example 3.1.1 Finding the mean, median, mode, and midrange

Suppose a vet wants to find the average weight of cats. The weights (in pounds) of five cats are in Example 3.1.1.

Table 3.1.1: Finding the Mean, Median, Mode, and Midrange

6.8	8.2	7.5	9.4	8.2

Find the mean, median, mode, and midrange of the weight of a cat.

Solution

Before starting any mathematics problem, it is always a good idea to define the unknown in the problem. In this case, you want to define the variable. The symbol for the variable is x.

The variable is $x =$ weight of a cat

Mean:

$$\overline{x} = \frac{6.8 + 8.2 + 7.5 + 9.4 + 8.2}{5} = \frac{40.1}{5} = 8.02 \text{ pounds}$$

Median:

You need to sort the list for both the median and the mode. The sorted list is in Example 3.1.2.

Table 3.1.2: Sorted List of Cat's Weights

6.8	7.5	8.2	8.2	9.4

There are 5 data points so the middle of the list would be the 3rd number. (Just put a finger at each end of the list and move them toward the center one number at a time. Where your fingers meet is the median.)

Table 3.1.3: Sorted List of Cats' Weights with Median Marked

6.8	7.5	8.2	8.2	9.4

The median is therefore 8.2 pounds.

Mode:

This is easiest to do from the sorted list that is in Example 3.1.2. Which value appears the most number of times? The number 8.2 appears twice, while all other numbers appear once.

Mode = 8.2 pounds.

Midrange:

To find the midrange we first need to know the lowest value, which is 6.8, and the highest value, which is 9.4. We add them and divide by 2.

Midrange = MR $= \frac{Lowest + Highest}{2} = \frac{6.8 + 9.4}{2} = \frac{16.2}{2} = 8.1$

A data set can have more than one mode. If there is a tie between two values for the most number of times, then both values are the mode, and the data is called bimodal (two modes). If every data point occurs the same number of times, there is no mode. If there are more than two numbers that appear the most times, then usually there is no mode.

In Example 3.1.1, there were an odd number of data points. In that case, the median was just the middle number. What happens if there is an even number of data points? What would you do?

Example 3.1.2 finding the median with an even number of data points

Suppose a vet wants to find the median weight of cats. The weights (in pounds) of six cats are in Example 3.1.4. Find the median.

Table 3.1.4: Weights of Six Cats

6.8	**8.2**	**7.5**	**9.4**	**8.2**	**6.3**

Solution

Variable: $x =$ weight of a cat

First, sort the list if it is not already sorted.

There are 6 numbers in the list so the number in the middle is between the 3rd and 4th numbers. Use your fingers starting at each end of the list in Example 3.1.5 and move toward the center until they meet. There are two numbers there.

Table 3.1.5: Sorted List of Weights of Six Cats

6.3	**6.8**	**7.5**	**8.2**	**8.2**	**9.4**

To find the median, just average the two numbers.

$$\text{Median } = MD = \frac{7.5 + 8.2}{2} = 7.85 \text{ pounds}$$

The median is 7.85 pounds.

Example 3.1.3 finding mean and median using technology

Suppose a vet wants to find the median weight of cats. The weights (in pounds) of six cats are in Example 3.1.4. Find the median

Solution

Variable: $x =$ weight of a cat

You can do the calculations for the mean and median using the technology.

The procedure for calculating the sample mean ($\overline{x}$) and the sample median (Med) on the TI-83/84 is in *Figures 3.1.1* through *3.1.4*. First, you need to go into the STAT menu and then Edit. This will allow you to type in your data (see *Figure* 3.1.1).

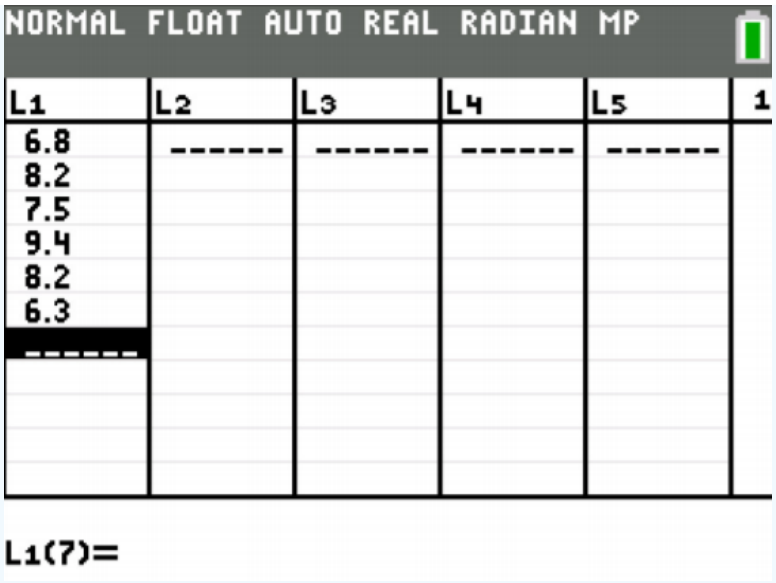


Figure 3.1.1: TI-83/84 Calculator Edit Setup

Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see *Figure* 3.1.2). The calculator will now put 1-Var Stats on the main screen. Now type in L1 (2nd button and 1) and then press ENTER. (Note if you have the newer operating system on the TI-84, then the procedure is slightly different.) If you press the down arrow, you will see the rest of the output from the calculator. The results from the calculator are in *Figure* 3.1.3 .

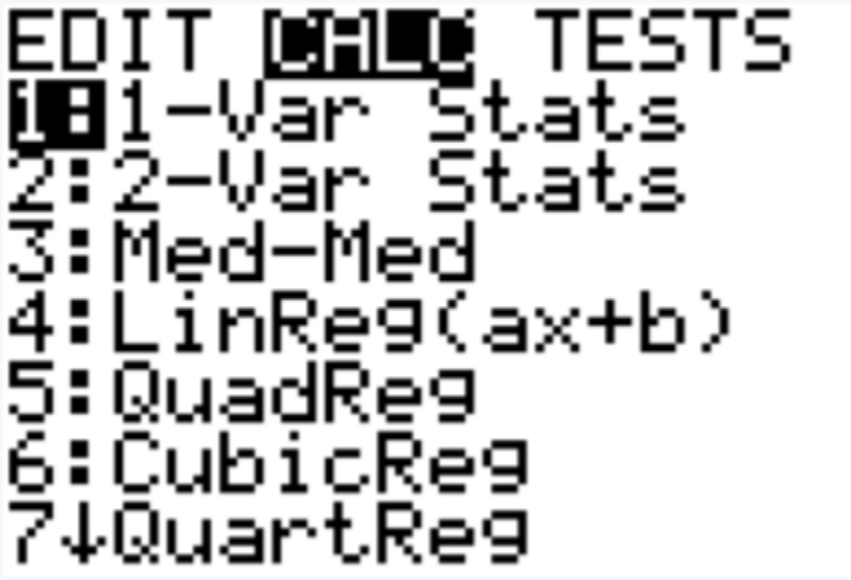


Figure 3.1.2: TI-83/84 Calculator CALC Menu

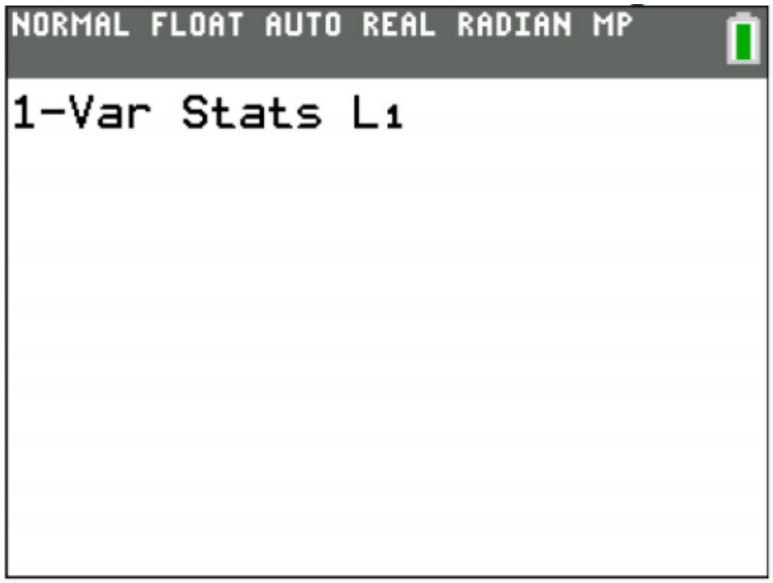


Figure 3.1.3: TI-83/84 Calculator Input for Example 3.1.3 Variable

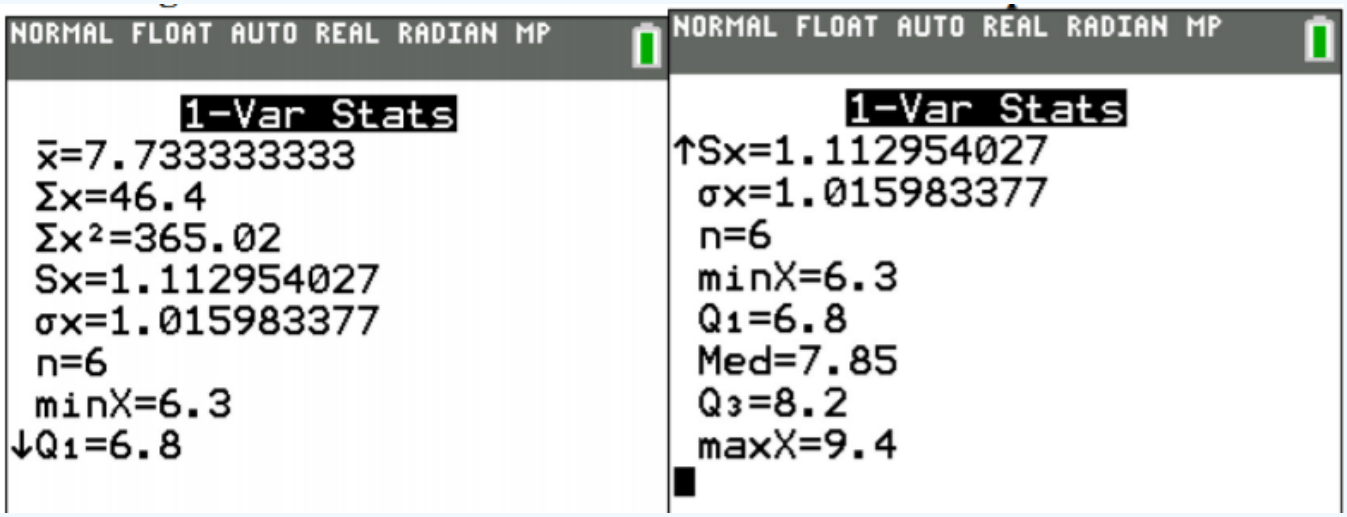


Figure 3.1.4: TI-83/84 Calculator Results for Example 3.1.3 Variable

Example 3.1.4 affect of extreme values on mean and median

Suppose you have the same set of cats from Example 3.1.1, but one additional cat was added to the data set. Example 3.1.6 contains the six cats' weights, in pounds.

Table 3.1.6: Weights of Six Cats

6.8	7.5	8.2	8.2	9.4	22.1

Find the mean and the median.

Solution

Variable: $x =$ weight of a cat

$$\text{Mean } = \overline{x} = \frac{6.8 + 7.5 + 8.2 + 8.2 + 9.4 + 22.1}{6} = 10.37 \text{ pounds}$$

The data is already in order, thus the median is between 8.2 and 8.2.

$$\text{Median } = MD = \frac{8.2 + 8.2}{2} = 8.2 \text{ pounds}$$

The mean is much higher than the median. Why is this? Notice that when the value of 22.1 was added, the mean went from 8.02 to 10.37, but the median did not change at all. This is because the mean is affected by extreme values, while the median is not. The hefty cat brought the mean weight up. In this case, the median is a much better measure of the center.

An outlier is a data value that is very different from the rest of the data. It can be really high or really low. Extreme values may be outliers if the extreme value is far enough from the center. In the example above, the data value 22.1 pounds is an extreme value, and it may be an outlier.

If there are extreme values in the data, the median is a better measure of the center than the mean. If there are no extreme values, the mean and the median will be similar, so most people use the mean.

The mean is not a resistant measure because it is affected by extreme values. The median and mode are resistant measures because extreme values do not affect them.

Determine Best Representation for Center: Mean or Median

As a consumer, you need to be aware that people choose the measure of center that best supports their claim. When you read an article in the newspaper that talks about the “average,” it usually means the mea,n but sometimes it refers to the median. Some articles will use the word “median” instead of “average” to be more specific. If you need to make an important decision and the information says “average”, it would be wise to ask if the “average” is the mean or the median before you decide.

As an example, suppose that a company wants to use the mean salary as the average salary for the company. This is because the high salaries of the administration will pull the mean higher. The company can say that the employees are paid well because the average is high. However, the employees want to use the median since it discounts the extreme values of the administration and will give a lower value of the average. This will make the salaries seem lower, and that a raise is in order.

Why use the mean instead of the median? The reason is that when multiple samples are taken from the same population, the sample means tend to be more consistent than other measures of the center. The sample mean is the more reliable measure of center.

To understand how the different measures of center relate to skewed or symmetric distributions, see the figure below. As you can see, sometimes the mean is smaller than the median and mode, sometimes the mean is larger than the median and mode, and sometimes they are the same value.

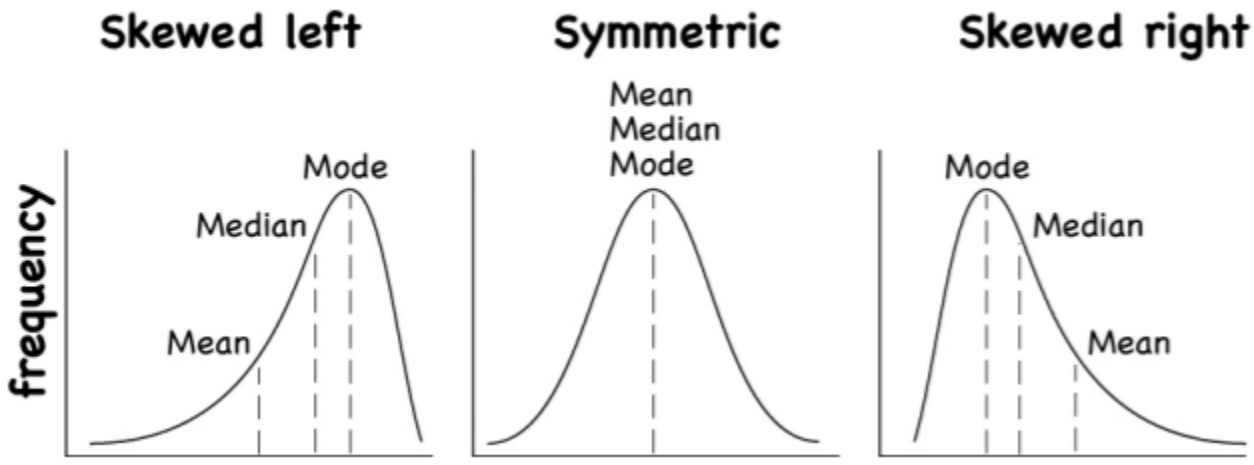


Figure 3.1.5: Mean, Median, Mode as Related to a Distribution

One last type of average is a weighted average. Weighted averages are used quite often in real life. Some teachers use them in calculating your grade in the course or your grade on a project. Some employers use them in employee evaluations. The idea is

that some activities are more important than others. As an example, a full-time teacher at a community college may be evaluated on their service to the college, their service to the community, whether their paperwork is turned in on time, and their teaching. However, teaching is much more important than whether their paperwork is turned in on time. When the evaluation is completed, more weight needs to be given to the teaching and less to the paperwork. This is a weighted average.

Weighted Mean

Definition 3.1.3

Weighted Mean

$\overline{x} = \frac{\sum x \cdot w}{\sum w}$ where w is the weight of the data value, x.

Example 3.1.5 weighted mean

In your biology class, your final grade is based on several things: a lab score, scores on two major tests, and your score on the final exam. There are 100 points available for each score. The lab score is worth 15% of the course, the two exams are worth 25% of the course each, and the final exam is worth 35% of the course. Suppose you earned scores of 95 on the labs, 83 and 76 on the two exams, and 84 on the final exam. Compute your weighted mean for the course.

Solution

Variable: $x =$ score

(Weighted Mean) $\overline{x} = \frac{\sum x \cdot w}{\sum w} = \frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$

$$\overline{x} = \frac{95(0.15) + 83(0.25) + 76(0.25) + 84(0.35)}{0.15 + 0.25 + 0.25 + 0.35} = \frac{83.4}{1.00} = 83.4\%$$

A weighted mean can be found using technology.

The procedure for calculating the weighted average on the TI-83/84 is in *Figures 3.1.6* through *3.1.9*. First, you need to go into the **[STAT]** menu, and then **[EDIT]**. This will allow you to type in the scores into **L1** and the weights into **L2** (see *Figure* 3.1.6).

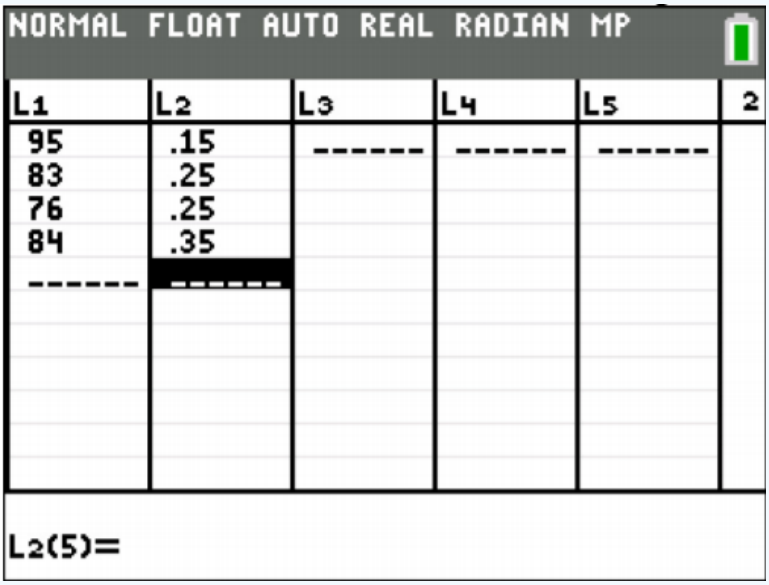


Figure 3.1.6: TI-3/84 Calculator Edit Setup

Once you have the data in the calculator, you then go back to the **[STAT]** menu, move over to **[CALC]**, and then choose **[1-Var Stats]** (see *Figure* 3.1.7). The calculator will now put **[1-Var Stats]** on the main screen. Now type in **L1** (**[2ND]** button and **[1]**), then a comma (button above the 7 button), and then **L2** (**[2ND]** button and **[2]**), and then press **[ENTER]**. (Note: if you have the newer operating system on the TI-84, then the procedure is slightly different.) The results from the calculator are in *Figure* 3.1.9. The $\overline{x}$ is the weighted mean.

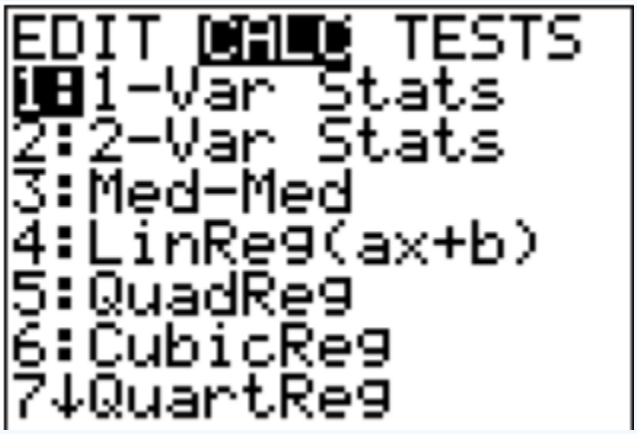


Figure 3.1.7: TI-83/84 Calculator CALC Menu

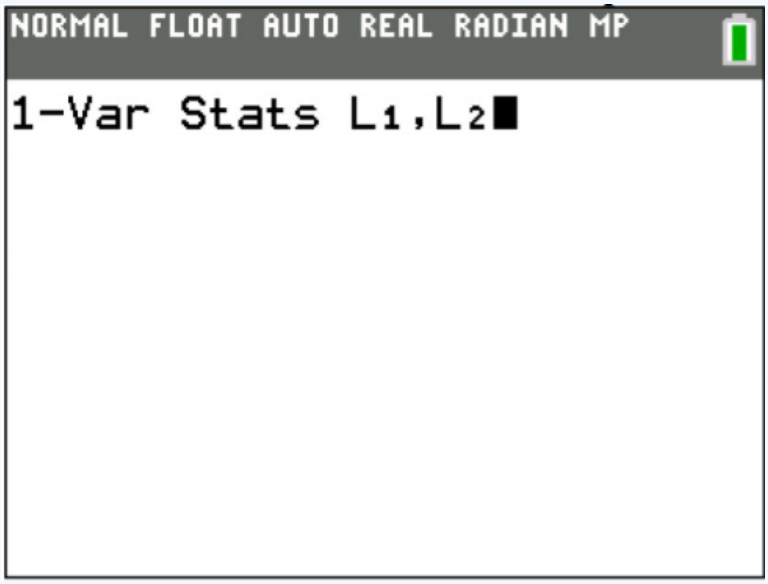


Figure 3.1.8: TI-83/84 Calculator Input for Weighted Mean

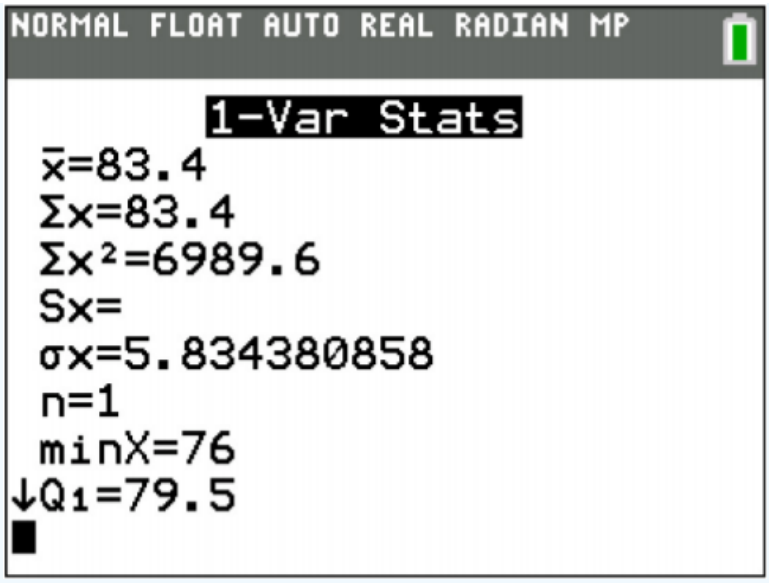


Figure 3.1.9: TI-83/84 Calculator Results for Weighted Mean

Example 3.1.6 weighted mean

The faculty evaluation process at John Jingle University rates a faculty member on the following activities: teaching, publishing, committee service, community service, and submitting paperwork promptly. The process involves reviewing student evaluations, peer evaluations, and supervisor evaluations for each teacher and awarding them a score on a scale from 1 to 10 (with 10 being the best). The weights for each activity are 20 for teaching, 18 for publishing, 6 for committee service, 4 for community service, and 2 for paperwork.

a. One faculty member had the following ratings: 8 for teaching, 9 for publishing, 2 for committee work, 1 for community service, and 8 for paperwork. Compute the weighted mean.
b. Another faculty member had ratings of 6 for teaching, 8 for publishing, 9 for committee work, 10 for community service, and 10 for paperwork. Compute the weighted mean.
c. Which faculty member had the higher mean?

Solution

a. Variable: $x =$ rating

(Weighted Mean) $\overline{x} = \dfrac{\sum x \cdot w}{\sum w} = \dfrac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$

$$\overline{x} = \frac{8(20) + 9(18) + 2(6) + 1(4) + 8(2)}{20 + 18 + 6 + 4 + 2} = \frac{354}{50} = 7.08$$

b. $\overline{x} = \dfrac{6(20) + 8(18) + 9(6) + 10(4) + 10(2)}{20 + 18 + 6 + 4 + 2} = \dfrac{378}{50} = 7.56$

c. The second faculty member has a higher mean.

You can find a weighted mean using technology. The last thing to mention is which average is used on which type of data.

The mode can be found on nominal, ordinal, interval, and ratio data since the mode is just the data value that occurs most often. You are just counting the data values. Median can be found on ordinal, interval, and ratio data since you need to put the data in order. As long as there is order to the data, you can find the median. Mean can be found on interval and ratio data since you must have numbers to add together.

Grouped Data

Sometimes the mean is computed from grouped data rather than the raw data set. The mean for grouped data is computed using the following steps. Also, the modal class is the class with the highest frequency.

Perform the following steps:

1. Add two columns to the table with headings "Midpoint (X_M)" and "f $\cdot X_M$."
2. Fill in the columns. The first of these columns consists of midpoints and the second column is each frequency multiplied with the corresponding midpoint.
3. Find the sum of the frequency columns. This sum is $n = \sum f$.
4. Find the sum of the last column, which is denoted as $\sum f \bullet X_M$.
5. Compute the mean.

$$\overline{x} = \frac{\sum f \cdot X_M}{\sum f}$$

Example 3.1.7 Grouped Mean and Modal Class

Find the grouped mean and modal class from the grouped data.

Grouped Frequency Distribution

Class Limits	Frequency
5 - 11	5
12 - 18	9
19 - 25	18
26 - 32	6
33 - 39	8

Table 3.1.7: Grouped Frequency Distribution

Solution

Grouped Frequency Distribution

Class Limits	Frequency	X_M	$f \cdot X_M$
5 - 11	5	8	40
12 - 18	9	15	135
19 - 25	18	22	396
26 - 32	6	29	174

Class Limits	Frequency	X_M	$f \cdot X_M$
33 - 39	8	36	288

Table 3.1.8: Grouped Frequency Distribution With Midpoint X_M and $f \cdot X_M$ Columns

n = Σf = 46

$\Sigma f \cdot X_M = 1033$

$$\overline{x} = \frac{1033}{46} \approx 22.46$$

Modal class: 19 - 25

Authors

"3.1: Measures of Center" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"3.1: Measures of Center" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 3.1: Measures of Center is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: Measures of Center by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

3.2: Measures of Variation Version 1

Learning Objectives

- Explain measures of variation to assess how spread-out data values are.
- Calculate the range, variance, and standard deviation for grouped and ungrouped data.
- Understand the importance of variation in evaluating consistency and comparing data sets.

Measures of variation describe how data values are spread out or dispersed within a dataset. Three key measures of variation are **range, variance, and standard deviation**. The **range** is the simplest measure, calculated as the difference between the maximum and minimum values, providing a rough estimate of variability. **Variance** measures the average squared deviation from the mean, capturing how much data points differ from the center. The **standard deviation**, the square root of variance, expresses variability in the same units as the original data, making it easier to interpret. Together, these measures help in understanding the consistency and distribution of data in statistical analysis.

Range

Definition: Range

The **range** of a data set is a measure of variation that represents the difference between the **maximum** and **minimum** values in the set. It is calculated using the formula:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

The range provides a simple way to understand the spread of data, but it is highly sensitive to outliers since it only considers the two extreme values and ignores the distribution of the other data points.

Example 3.2.1

A survey was conducted on how much money 8 university students spent on books in a semester. The recorded amounts (in dollars) were as follows:

250, 320, 275, 400, 150, 380, 290, 310

Find the range of the amount spent on books by the students.

Solution

Range = 400 - 150 = 250

Sample Variance

Sample variance is a measure of how much the values in a sample differ from the sample mean. It quantifies the spread or dispersion of the data points in a sample.

Definition: Sample Variance

The sample variance can be computed using the formula below.

$$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)}$$

where:

- s^2 = sample variance
- n = number of observations in the sample
- x = individual data points
- $\sum x$ = sum of all data points
- $\sum x^2$ = sum of the squares of all data points

Example 3.2.2

A survey was conducted on how much money 8 university students spent on books in a semester. The recorded amounts (in dollars) were as follows:

250, 320, 275, 400, 150, 380, 290, 310

Find the variance of the amount spent on books by the students. Round to three decimal places.

Solution

1. Create a table with two columns. The first column lists the data values, and the second column lists the squares of the data values.

Data Values and Their Squares

x	x^2
250	62500
320	102400
275	75625
400	160000
150	22500
380	144400
290	84100
310	96100

Table 3.2.1: The data values and Their Squares

2. Find the sum of each column and the number of data values.

$\sum x = 2,375$, $\sum x^2 = 747,625$, and $n = 8$

3. Substitute the results of step 2 into the formula.

$$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)}$$

$$s^2 = \frac{8 \cdot 747,625 - (2375)^2}{8 \cdot 7}$$

$$s^2 = \frac{5,981,000 - 5,640,625}{56}$$

$s^2 = \frac{340,375}{56}$

$s^2 = 6078.125$

Standard Deviation

Standard deviation measures how spread out the values in a dataset are from the mean. It indicates the average distance of data points from the mean, providing insight into the dataset's variability. A higher standard deviation means the data points are more dispersed, while a lower standard deviation indicates they are closer to the mean. It is the square root of variance and is commonly used in statistics to assess consistency, risk, or volatility in various fields.

Definition: Standard Deviation

s $= \sqrt{s^2} = \sqrt{Variance}$

Example 3.2.3

A survey was conducted on how much money 8 university students spent on books in a semester. The recorded amounts (in dollars) were as follows:

250, 320, 275, 400, 150, 380, 290, 310

Find the standard deviation of the amount spent on books by the students, round to three decimal places.

Solution

Use the variance from example 2 to compute the standard deviation.

s $= \sqrt{6078.125} = 77.962$

Example 3.2.4

Studies indicate that the height of 12-year-old students in the U.S. varies based on genetics, nutrition, and other environmental factors. A classroom's height distribution can help researchers understand if the students' growth aligns with national trends.

Given the heights of 11 students (in inches):

58, 62, 65, 60, 67, 70, 64, 66, 59, 68, 63
Find the range, variance, and standard deviation of the data set. Round the answers to three decimal places.

Solution

1. Range = 70 - 58 = 12
2. Compute the variance as follows.

Create a table with two columns. The first column lists the data values, and the second column lists the squares of the data values.

Data Values and Their Squares

x	x^2
58	3,364
62	3,844
65	4,225
60	3,600
67	4,489
70	4,900
64	4,096
66	4,356
59	3,481
68	4,624
63	3,969

Table 3.2.2: The data values and Their Squares

Find the sum of each column and the number of data values.

$\sum x = 702$, $\sum x^2 = 44,948$, and $n = 11$

Substitute the results above into the formula for the variance.

$$s^2 = \frac{n \cdot \sum x^2 - (\sum x)^2}{n \cdot (n-1)}$$

$$s^2 = \frac{11 \cdot 44,948 - (702)^2}{11 \cdot 10}$$

$$s^2 = \frac{494,428 - 492,804}{110}$$

$$s^2 = \frac{1624}{110}$$

$$s^2 = 14.764$$

3. Compute the standard deviation.

$s = \sqrt{14.764} = 3.842$

Instructions to Compute Variance and Standard Deviation Using TI-84+ Calculator

Step 1: Enter the Data

1. Press the [**STAT**] button.
2. Select [**1: Edit...**] and press [**ENTER**].
3. In **L1** (List 1), enter each data value one by one, pressing [**ENTER**] after each.

Step 2: Compute the Statistics

1. Press [**STAT**] again.
2. Use the right arrow key to highlight [**CALC**].
3. Select [**1: 1-Var Stats**] and press [**ENTER**].
4. Type **L1** (if your data is stored in List 1).
   - If L1 is not selected, press [**2nd**] and [**1**] to insert **L1**.
5. Press [**ENTER**] to compute the statistics.

Step 3: Locate the Standard Deviation and Variance

1. Find **σx** (sigma x), which represents the population standard deviation.
2. Find **Sx**, which represents the sample standard deviation.
   - Use the sample standard deviation **Sx** for our results.
   - Do not use the population standard deviation **σx**.
3. To find the variance, square the standard deviation.
   - Sample variance = $(S_x)^2$

Variance and Standard Deviation for Grouped Data

The steps to find the sample variance (s^2) and standard deviation (s) for grouped data are provided below.

Step 1: Make a table as shown and find the midpoint of each class.

Titles for Grouped Data

Class	Frequency: f	Midpoint: X_M	$f \cdot X_M$	$f \cdot (X_M)^2$

Table 3.2.3: Headers of Grouped Frequency Distribution

Step 2: Find the sums of the second and last two columns. Substitute the sums into the formula below and solve to get the variance.

$$s^2 = \frac{n(\sum f \cdot x_m^2) - (\sum f \cdot x_m)^2}{n(n-1)}$$

Step 3: Take the square root to get the standard deviation.

$$s = \sqrt{\text{variance}}$$

Example 3.2.5 Variance and Standard Deviation for Grouped Data

Find the variance and standard deviation for the frequency distribution of the data.

Grouped Frequency Distribution

Class Boundaries	Frequency
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5

Class Boundaries	Frequency
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2

Table 3.2.4: Grouped Frequency Distribution

Solution

Grouped Frequency Distribution

Class Boundaries	Frequency	X_M	$f \cdot X_M$	$f \cdot (X_M)^2$
5.5 – 10.5	1	8	8	64
10.5 – 15.5	2	13	26	338
15.5 – 20.5	3	18	54	972
20.5 – 25.5	5	23	115	2645
25.5 – 30.5	4	28	112	3136
30.5 – 35.5	3	33	99	3267
35.5 – 40.5	2	38	76	2888
	20		490	13310

Table 3.2.5:Grouped Frequency Distribution with Added Columns

Compute the grouped variance (s^2) and standard deviation (s):

$$s^2 = \frac{n(\sum f \cdot x_m^2) - (\sum f \cdot x_m)^2}{n(n-1)}$$

$$s^2 = \frac{20\,(13310) - (490)^2}{20\,(19)}$$

$$s^2 = \frac{26100}{380} \approx 68.68421053$$

$$s \approx \sqrt{68.68421053}$$

$$s \approx 8.287593772$$

If we round to two decimal places the answers for the sample variance (s^2) and standard deviation (s) for the grouped data, we will get the following:

$$s^2 \approx 68.68$$

$$s \approx 8.29$$

Instructions to Compute Grouped Variance and Standard Deviation Using TI-84+ Calculator

Step 1: Enter Data into the Calculator

1. Press the [**STAT]** button.
2. Select [**1: Edit...]** and press [**ENTER]**.
3. Enter the midpoints into **L1**.
4. Enter the frequencies into **L2**.

Step 2: Calculate the Grouped Standard Deviation

1. Press [**STAT]**.
2. Use the right arrow key to select [**CALC]**.
3. Select [**1: 1-Var Stats]** and press [**ENTER]**.
4. Make sure the screen has [**List:L1]** and [**FreqList:L2]**, and then use the down arrow to select [**Calculate]**. (Note: Press [**2nd]** and **1** for **L1** and Press [**2nd]** and **2** for **L2**.
5. Press [**ENTER]**.

Step 3: Calculate the Grouped Variance

1. Find **σx** (sigma x), which represents the grouped population standard deviation.
2. Find **Sx**, which represents the grouped sample standard deviation.
   - Use the grouped sample standard deviation **Sx** for our results.
   - Do not use the grouped population standard deviation **σx**.
3. To find the grouped variance, square the standard deviation.
   - grouped variance = $(S_x)^2$

Author

"3.2: Measures of Variation Version 1" by Alfie Swan is licensed under CC BY 4.0

Attributions

"3.2: Measures of Spread" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 3.2: Measures of Variation Version 1 is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

- 3.2: Measures of Spread by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

3.3: Measures of Variation Version 2

Learning Objectives

- Explain measures of variation to assess how spread-out data values are.
- Calculate the range, variance, and standard deviation for grouped and ungrouped data.
- Understand the importance of variation in evaluating consistency and comparing data sets.

Range

Variability is an important idea in statistics. If you were to measure the height of everyone in your classroom, every observation gives you a different value. That means not every student has the same height. Thus, there is variability in people's heights. If you were to take a sample of the income level of people in a town, every sample would give you different information. There is variability between samples, too. Variability describes how the data are spread out. If the data are very close to each other, then there is low variability. If the data are very spread out, then there is high variability. How do you measure variability? It would be good to have a number that measures it. This section will describe some of the different measures of variability, also known as variation.

In Example 3.3.1, the average weight of a cat was calculated to be 8.02 pounds. How much does this tell you about the weight of all cats? Can you tell if most of the weights were close to 8.02, or were the weights spread out? What are the highest and the lowest weights? All you know is that the center of the weights is 8.02 pounds. You need more information.

Definition 3.3.1

The **range (R)** of a set of data is the difference between the highest and the lowest data values.

$$\mathrm{R} = \text{highest value} - \text{lowest value}$$

Example 3.3.1: Finding the Range

Look at the following three sets of data. Find the range of each of these.

a. $10, 20, 30, 40, 50$
b. $10, 29, 30, 31, 50$
c. $28, 29, 30, 31, 32$

Solution

a.

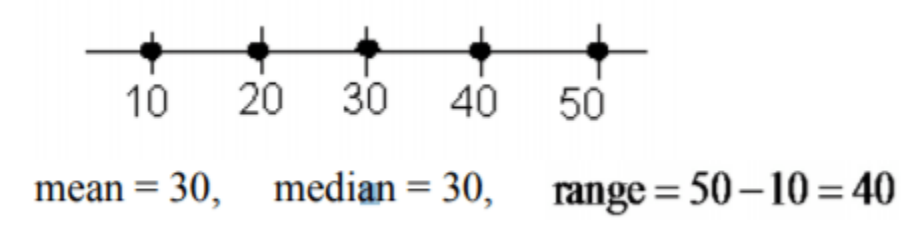


Figure 3.3.1: Dot Plot for Example 3.3.1a

b.

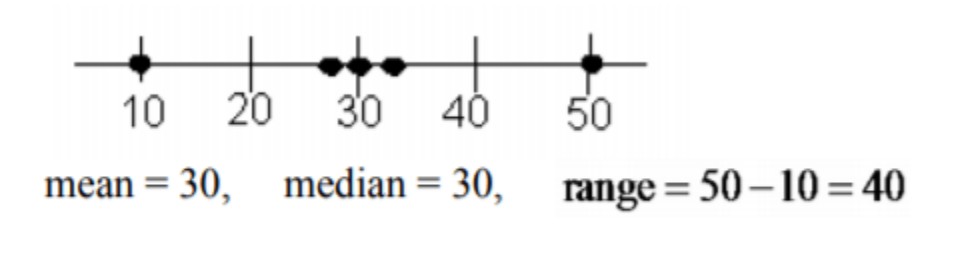


Figure 3.3.2: Dot Plot for Example 3.3.1b

c.

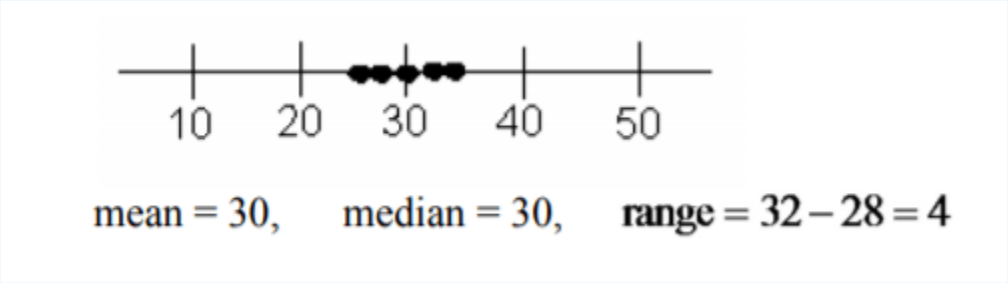


Figure 3.3.3: Dot Plot for Example 3.3.1

Based on the mean, median, and range in Example 3.3.1, the first two distributions are the same, but you can see from the graphs that they are different. In Example 3.3.1a the data are spread out equally. In Example 3.3.1b, the data has a clump in the middle and a single value at each end. The mean and median are the same for Example 3.3.1c, but the range is very different. All the data is clumped together in the middle.

The range doesn't provide a very accurate picture of the variability. A better way to describe how the data is spread out is needed. Instead of looking at the distance, the highest value is from the lowest How about looking at the distance each value is from the mean? This distance is called the **deviation**.

Variance and Standard Deviation of Raw Data

Example 3.3.2: Finding the Deviations

Suppose a vet wants to analyze the weights of cats. The weights (in pounds) of five cats are 6.8, 8.2, 7.5, 9.4, and 8.2. Find the standard deviation for each of the data values.

Solution

Variable: $x =$ weight of a cat

The mean for this data set is $\overline{x} = 8.02$ pounds.

Table 3.3.1: Deviations of Weights of Cats

x	$x - \overline{x}$
6.8	6.8-8.02 = -1.22
8.2	8.2-8.02=0.18
7.5	7.5-8.02=-0.52

x	$x-\overline{x}$
9.4	9.4-8.02=1.38
8.2	8.2-8.02=0.18

Now you might want to average the deviation, so you need to add the deviations together.

Table 3.3.2: Sum of Deviations of Weights of Cats

x	$x-\overline{x}$
6.8	6.8-8.02 = -1.22
8.2	8.2-8.02=0.18
7.5	7.5-8.02=-0.52
9.4	9.4-8.02=1.38
8.2	8.2-8.02=0.18
Total	0

This can't be right. The average distance from the mean cannot be 0. The reason it adds to 0 is that there are some positive and negative values. You need to get rid of the negative signs. How can you do that? You could square each deviation.

Table 3.3.3: Squared Deviations of Weights of Cats

x	$x-\overline{x}$	$(x-\overline{x})^2$
6.8	6.8-8.02 = -1.22	1.4884
8.2	8.2-8.02=0.18	0.0324
7.5	7.5-8.02=-0.52	0.2704
9.4	9.4-8.02=1.38	1.9044
8.2	8.2-8.02=0.18	0.0324
Total	0	3.728

Now, average the total of the squared deviations. The only thing is that in statistics, there is a strange average here. Instead of dividing by the number of data value,s you divide by the number of data values minus 1 because all the values in the data are free to be any value except for one. It has to be a specific value to calculate the sample mean. The number of numbers that are free is called the degrees of freedom. Since the sample mean is part of the variance formula, the denominator will be the total number of data values minus one. In this case,e you would have

$$s^2 = \frac{3.728}{5-1} = \frac{3.728}{4} = 0.932 \text{ pounds}^2$$

Notice that this is denoted as s^2. This is called the variance, and it is a measure of the average squared distance from the mean. If you now take the square root, you will get the average distance from the mean. This is called the standard deviation, and is denoted with the letter s.

$$s = \sqrt{.932} \approx 0.965 \text{ pounds}$$

The standard deviation is the average (mean) distance from a data point to the mean. It can be thought of as how much a typical data point differs from the mean.

Definition 3.3.2: Sample Variance

The **sample variance** formula:

$$s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

where $\overline{x}$ is the sample mean, n is the sample size, and $\sum$ means to find the sum.

Definition 3.3.3: Sample Standard Deviation

The **sample standard deviation** formula:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$$

The $n - 1$ on the bottom has to do with a concept called degrees of freedom. It makes the sample standard deviation a better approximation of the population standard deviation.

Definition 3.3.4: Population Variance

The **population variance** formula:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

where σ is the Greek letter sigma and σ^2 represents the population variance, μ is the population mean, and N is the size of the population.

Definition 3.3.5: Population Standard Deviation

The **population standard deviation** formula:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Note

The sum of the deviations should always be 0. If it isn't, then it is because you rounded, you used the median instead of the mean, or you made an error. Try not to round too much in the calculations for standard deviation, since each rounding causes a slight error

Example 3.3.3: Finding the Standard Deviation

Suppose that a manager wants to test two new training programs. He had 5 people for each training type and measured the time it took to complete a task after the training. Both training groups include all the recruits. These two groups represent two different populations. The times for both trainings are in Example 3.3.4. Which training method is better?

Table 3.3.4: Time to Finish Task in Minutes

Training 1	56	75	48	63	59
Training 2	60	58	66	59	58

Solution

It is important that you define what each variable is, since there are two of them.

Variable 1: $X_1 =$ productivity from training 1

Variable 2: $X_2 =$ productivity from training 2

To answer which training method is better, first you need some descriptive statistics. Start with the mean for each sample.

$$\mu_1 = \frac{56 + 75 + 48 + 63 + 59}{5} = 60.2 \text{ minutes}$$

$$\mu_2 = \frac{60 + 58 + 66 + 59 + 58}{5} = 60.2 \text{ minutes}$$

Since both means are the same values, you cannot answer the question about which is better. Now, calculate the standard deviation for each population.

Table 3.3.5: Squared Deviations for Training 1

x_1	$x_1 - \mu_1$	$(x_1 - \mu_1)^2$
56	-4.2	17.64
75	14.8	219.04
48	-12.2	148.84
63	2.8	7.84
59	-1.2	1.44
Total	0	394.8

Table 3.3.6: Squared Deviations for Training 2

x_2	$x_2 - \mu_2$	$(x_2 - \mu_2)^2$
60	-0.2	0.04
58	-2.2	4.84
66	5.8	33.64
59	-1.2	1.44
58	-2.2	4.84
Total	0	44.8

The variance for each population is:

$$\sigma_1^2 = \frac{394.8}{5 - 1} = 98.7 \text{ minutes}^2$$

$$\sigma_2^2 = \frac{44.8}{5 - 1} = 11.2 \text{ minutes}^2$$

The standard deviations are:

$$\sigma_1 = \sqrt{98.7} \approx 9.93 \text{ minutes}$$

$$\sigma_2 = \sqrt{11.2} \approx 3.35 \text{ minutes}$$

From the standard deviations, the second training seemed to be the better training since the data is less spread out. This means it is more consistent. It would be better for the managers in this case to have a training program that produces more consistent results so they know what to expect for the time it takes to complete the task.

You can do the calculations for the descriptive statistics using the technology. The procedure for calculating the sample mean ($\overline{x}$) and the sample standard deviation (s_x) for X_2 in Example 3.3.3 on the TI-83/84 is in *Figures 3.2.1* through *3.2.4* (the procedure is the same for X_1). Note that the calculator gives you the population standard deviation (σ_x) because it doesn't know whether the data you input is a population or a sample. You need to decide which value you need to use, based on whether you have a population or a sample. In almost all cases, you have a sample and will be using s_x. Also, the calculator uses the notation s_x instead of just s. It is just a way for it to denote the information. First, you need to go into the STAT menu and then Edit. This will allow you to type in your data (see *Figure* 3.3.1).

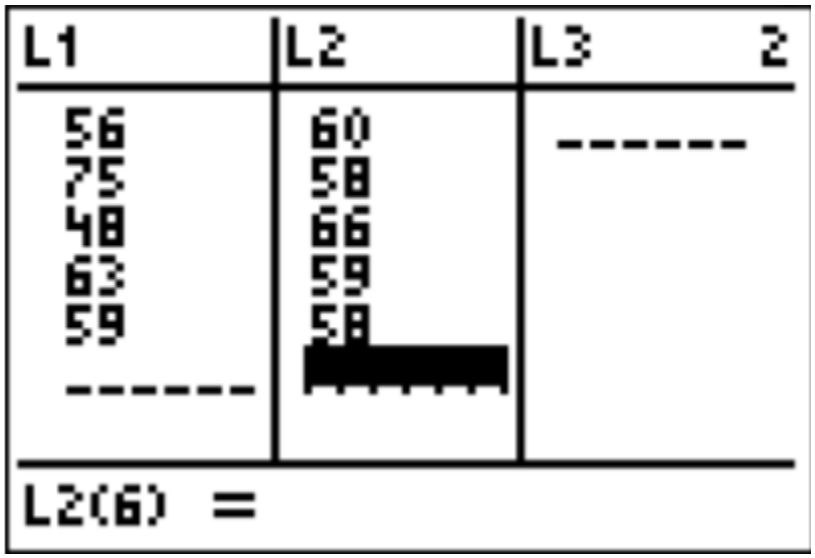


Figure 3.3.4: TI-83/84 Calculator Edit Setup

Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see *Figure* 3.3.2). The calculator will now put 1-Var Stats on the main screen. Now type in L2 (2nd button and 2) and then press ENTER. (Note: if you have the newer operating system on the TI-84, then the procedure is slightly different.) The results from the calculator are in *Figure* 3.3.4.

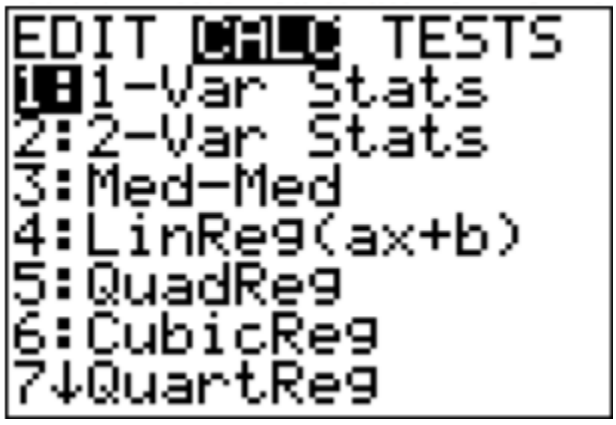


Figure 3.3.5: TI-83/84 Calculator CALC Menu



Figure 3.3.3: TI-83/84 Calculator Input for Example 3.3.6 Variable X_2

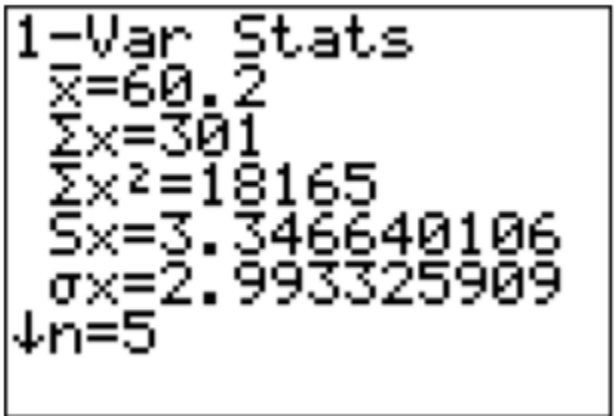


Figure 3.3.4: TI-83/84 Calculator Results for Example 3.3.7 Variable X_2

In general, a “small” standard deviation means the data is close together (more consistent), and a “large” standard deviation means the data is spread out (less consistent). Sometimes you want consistent data, and sometimes you don’t. As an example,e if you are making bolts, you want to lengths to be very consistent, so you want a small standard deviation. If you are administering a test to see who can be a pilot, you want a large standard deviation so you can tell who the good pilots are and who the bad ones.

What do “small” and “large” mean? To a bicyclist whose average speed is 20 mph, s = 20 mph is huge. To an airplane whose average speed is 500 mph, s = 20 mph is nothing. The “size” of the variation depends on the size of the numbers in the problem and the mean. Another situation where you can determine whether a standard deviation is small or large is when you are comparing two different samples, such as in example #3.2.3. A sample with a smaller standard deviation is more consistent than a sample with a larger standard deviation.

Many other books and authors stress that there is a computational formula for calculating the standard deviation. However, this formula doesn’t give you an idea of what standard deviation is and what you are doing. It is only good for doing the calculations quickly. It goes back to the days when standard deviations were calculated by hand, and the person needed a quick way to calculate the standard deviation. It is an archaic formula that this author is trying to eradicate it. It is not necessary anymore, since most calculators and computers will do the calculations for you with as much meaning as this formula gives. It is suggested that you never use it. If you want to understand what the standard deviation is doing, then you should use the definition formula. If you want an answer quickly, use a computer or calculator.

Use of Standard Deviation

One of the uses of the standard deviation is to describe how a population is distributed by using Chebyshev’s Theorem. This theorem works for any distribution, whether it is skewed, symmetric, bimodal, or any other shape. It gives you an idea of how much data is a certain distance on either side of the mean.

Definition 3.3.6: Chebyshev's Theorem

For any set of data:

- At least 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- At least 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- At least 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

Example 3.3.4: Using Chebyshev's Theorem

The U.S. Weather Bureau has provided the information in Example 3.3.7 about the total annual number of reported strong to violent (F3+) tornados in the United States for the years 1954 to 2012. ("U.S. tornado climatology," 17).

Table 3.3.7: Annual Number of Violent Tornadoes in the U.S.

46	47	31	41	24	56	56	23	31	59
39	70	73	85	33	38	45	39	35	22
51	39	51	131	37	24	57	42	28	45
98	35	54	45	30	15	35	64	21	84
40	51	44	62	65	27	34	23	32	28
41	98	82	47	62	21	31	29	32	

a. Use Chebyshev's theorem to find an interval centered on the mean annual number of strong to violent (F3+) tornadoes in which you would expect at least 75% of the years to fall.
b. Use Chebyshev's theorem to find an interval centered on the mean annual number of strong to violent (F3+) tornadoes in which you would expect at least 88.9% of the years to fall.

Solution

a. Variable: $x =$ number of strong or violent (F3+) tornadoes Chebyshev's theorem says that at least 75% of the data will fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.

You do not have the population, so you need to estimate the population mean and standard deviation using the sample mean and standard deviation. You can find the sample mean and standard deviation using technology:

$$\overline{x} \approx 46.24, s \approx 22.18$$

So,

$$\mu \approx 46.24, \sigma \approx 22.18$$

$$\mu - 2\sigma \text{ to } \mu + 2\sigma$$

$$46.24 - 2(22.18) \text{ to } 46.24 + 2(22.18)$$

$$46.24 - 44.36 \text{ to } 46.24 + 44.36$$

$$1.88 \text{ to } 90.60$$

Since you can't have a fractional number of tornadoes, round to the nearest whole number.

At least 75% of the years have between 2 and 91 strong to violent (F3+) tornadoes. (Actually, all but three years' values fall in this interval, which means that $\frac{56}{59} \approx 94.9\%$ fall in the interval.)

b. Variable: $x =$ number of strong or violent (F3+) tornadoes Chebyshev's theorem says that at least 88.9% of the data will fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.

$$\mu - 3\sigma \text{ to } \mu + 3\sigma$$

$$46.24 - 3(22.18) \text{ to } 46.24 + 3(22.18)$$

$$46.24 - 66.54 \text{ to } 46.24 + 66.54$$

$$-20.30 \text{ to } 112.78$$

Since you can't have a negative number of tornadoes, the lower limit is 0. Since you can't have a fractional number of tornadoes, round to the nearest whole number.

At least 88.9% of the years have between 0 and 113 strong to violent (F3+) tornadoes.

(Actually, all but one year fall in this interval, which means that $\frac{58}{59} \approx 98.3\%$ fall in the interval.)

Chebyshev's Theorem says that at least 75% of the data is within two standard deviations of the mean. That percentage is fairly high. There isn't much data outside two standard deviations. A rule that can be followed is that if a data value is within two standard deviations, then that value is a common data value. If the data value is outside two standard deviations of the mean, either above or below, then the number is uncommon. It could even be called unusual. An easy calculation that you can do to figure it out is to find the difference between the data point and the mean, and then divide that answer by the standard deviation. As a formula, this would be

$$\frac{x - \mu}{\sigma}.$$

If you don't know the population mean, μ, and the population standard deviation, σ, then use the sample mean, $\overline{x}$, and the sample standard deviation, s, to estimate the population parameter values. However, realize that using the sample standard deviation may not be very accurate.

Example 3.3.5 determining if a value is unusual

a. In 1974, there were 131 strong or violent (F3+) tornadoes in the United States. Is this value unusual? Why or why not?
b. In 1987, there were 15 strong or violent (F3+) tornadoes in the United States. Is this value unusual? Why or why not?

Solution

a. Variable: $x =$ number of strong or violent (F3+) tornadoes

To answer this question, first find how many standard deviations 131 is from the mean. From Example 3.3.4, we know $\mu \approx 46.24$ and $\sigma \approx 22.18$. For $x = 131$,

$$\frac{x - \mu}{\sigma} = \frac{131 - 46.24}{22.18} \approx 3.82$$

Since this value is more than 2, it is unusual to have 131 strong or violent (F3+) tornadoes in a year.

b. Variable: $x =$ number of strong or violent (F3+) tornadoes For this question the $x = 15$,

$$\frac{x - \mu}{\sigma} = \frac{15 - 46.24}{22.18} \approx -1.41$$

Since this value is between -2 and 2, it is not unusual to have only 15 strong or violent (F3+) tornadoes in a year.

Variance and Standard Deviation for Grouped Data

The steps to find the sample variance (s^2) and standard deviation (s) for grouped data is provided below.

Step 1: Make a table as shown and find the midpoint of each class.

Titles for Grouped Data

Class	Frequency: f	Midpoint: X_M	$f \cdot X_M$	$f \cdot (X_M)^2$

Table 3.3.8: Headers of Grouped Frequency Distribution

Step 2: Find the sum of the second and last two columns. Substitute the sums into the formula below and solve to get the variance.

$$s^2 = \frac{n(\sum f \cdot x_m^2) - (\sum f \cdot x_m)^2}{n(n-1)}$$

Step 3: Take the square root to get the standard deviation.

$$s = \sqrt{\text{variance}}$$

Example 3.3.6 Variance and Standard Deviation for Grouped Data

Find the variance and standard deviation for the frequency distribution of the data.

Grouped Frequency Distribution

Class Boundaries	Class Boundaries
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2

Table 3.3.9: Grouped Frequency Distribution

Solution

Grouped Frequency Distribution

Class Boundaries	Frequency	X_M	$f \cdot X_M$	$f \cdot (X_M)^2$
5.5 – 10.5	1	8	8	64

Class Boundaries	Frequency	X_M	$f \cdot X_M$	$f \cdot (X_M)^2$
10.5 – 15.5	2	13	26	338
15.5 – 20.5	3	18	54	972
20.5 – 25.5	5	23	115	2645
25.5 – 30.5	4	28	112	3136
30.5 – 35.5	3	33	99	3267
35.5 – 40.5	2	38	76	2888
	20		490	13310

Table 3.3.10:Grouped Frequency Distribution with Added Columns

Compute the grouped variance (s^2) and standard deviation (s):

$$s^2 = \frac{n(\sum f \cdot x_m^2) - (\sum f \cdot x_m)^2}{n(n-1)}$$

$$s^2 = \frac{20\,(13310) - (490)^2}{20\,(19)}$$

$$s^2 = \frac{26100}{380} \approx 68.68421053$$

$$s \approx \sqrt{68.68421053}$$

$$s \approx 8.287593772$$

If we round to two decimal places the answers for the sample variance (s^2) and standard deviation (s) for the grouped data, we will get the following:

$$s^2 \approx 68.68$$

$$s \approx 8.29$$

Authors

"3.3: Measures of Variation Version 2" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"3.2: Measures of Spread" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 3.3: Measures of Variation Version 2 is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Measures of Spread by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

3.4: Measures of Position

Learning Objectives

- Identify the position of data values within a data set using measures of position.
- Calculate z-scores to determine how many standard deviations a value is from the mean.
- Use percentiles and quartiles to divide data into ranked sections.

Measures of position help describe where a specific data point stands about the entire dataset. Z-scores measure how many standard deviations a value is from the mean, helping identify outliers and standardize comparisons across different datasets. Percentiles indicate the percentage of values below a given data point, commonly used in standardized testing and growth charts. Quartiles divide data into four equal parts, with Q1 (25th percentile) representing the lower quarter, Q2 (50th percentile or median) marking the middle, and Q3 (75th percentile) showing the upper quarter. These measures are essential for understanding data distribution, detecting outliers, and making informed statistical decisions.

Z-scores

A **z-score** tells you how far a data point is from the average (mean) of a data set, measured in standard deviations. It helps compare values from different distributions by showing if a number is above or below the average and by how much. Z-scores help standardize values so they can be compared, even if they come from different groups or scales.

Definition: Z-Score Formula

$$Z = \frac{X - \bar{X}}{s}$$

Meaning of Each Variable:

- Z represents the z-score (standardized value), which tells how many standard deviations a data point is from the sample mean
- X represents the individual data point (observed value)
- $\bar{X}$ represents the sample mean (average of the sample data)
- s represents the sample standard deviation

Example 3.4.1

A professor in a Psychology 101 class gives a test, and the students' scores are recorded. Suppose the sample mean test score is 78 with a sample standard deviation of 10.

Two students, Alex and Jordan, received the following scores:

a. Alex's score: 85
b. Jordan's score: 65

Solution

Using the z-score formula:

$$Z = \frac{X - \bar{X}}{s}$$

where:

- X is the individual test score
- $\bar{X} = 78$ (sample mean)
- $s = 10$ (sample standard deviation)

Now, let's compute their z-scores.

Alex's Z-Score:

$$Z = \frac{85 - 78}{10} = \frac{7}{10} = 0.7$$

Jordan's Z-Score:

$$Z = \frac{65 - 78}{10} = \frac{-13}{10} = -1.3$$

Example 3.4.2

A server at a restaurant tracks their weekly tips earned and hours worked. Over several weeks, the average weekly tips earned is $500 with a standard deviation of $100. This week, the server earned $650 in tips. Meanwhile, the average number of hours worked per week is 30 hours with a standard deviation of 5 hours, and this week, the server worked 40 hours.

Using z-scores, determine whether the server's weekly tips or weekly hours worked deviate more from the average, indicating where their performance was stronger relative to past trends.

Solution

Using the z-score formula:

1. Z-score for weekly tips: $Z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$
2. Z-score for weekly hours worked: $Z = \frac{40 - 30}{5} = \frac{10}{5} = 2.0$

Conclusion:

The z-score for hours worked 2 is higher than the z-score for tips earned 1.5. This means that relative to their average, the server worked significantly more hours than usual compared to how much more they earned in tips. Therefore, their performance in terms of hours worked deviated more from the average than their tips.

Percentiles

A **percentile** is a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors also use percentiles to track a child's growth.

The **kth percentile** is the data value that has k% of the data at or below that value.

Example 3.4.3 interpreting percentile

a. What does a score of the 90th percentile mean?
b. What does a score of the 70th percentile mean?

Solution

a. This means that 90% of the scores were at or below this score. (A person did the same as or better than 90% of the test takers.)
b. This means that 70% of the scores were at or below this score.

Example 3.4.4 percentile versus score

If the test was out of 100 points and you scored at the 80th percentile, what was your score on the test?

Solution

You don't know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% or so.

Quartiles

Quartiles are a way of splitting a group of numbers into four equal parts to help understand how data is spread out. Imagine you line up all your test scores from lowest to highest. Quartiles break this list into four sections:

- First quartile (Q_1): The number that separates the lowest 25% of scores from the rest.
- Second quartile (Q_2 or median): The middle number that splits the data into two halves.
- Third quartile (Q_3): The number that separates the highest 25% of scores from the rest.

Why are quartiles useful?

1. Understanding data spread – They help show if data is grouped closely or spread out.
2. Finding outliers – If a value is far beyond Q1 or Q3, it might be an unusual (outlier) data point.
3. Comparing groups – They help compare data sets, like test scores of different classes.
4. Making box plots – Quartiles are used to create box-and-whisker plots, a great way to visualize data distribution.

Simply put, quartiles help break large amounts of information into smaller, meaningful parts, making it easier to analyze and compare.

Definition 3.4.1

To find the quartiles:

1. Sort the data in increasing order.
2. Find the median, this divides the data list into 2 halves.
3. Find the median of the data below the median. This value is Q_1.
4. Find the median of the data above the median. This value is Q_3.
   Ignore the median in both calculations for Q_1 and Q_3.

If you record the quartiles together with the maximum and minimum you have five numbers. This is known as the five-number summary. The five-number summary consists of the minimum, the first quartile (Q_1), the median, the third quartile (Q_3), and the maximum (in that order).

The interquartile range, *IQR*, is the difference between the first and third quartiles, Q_1 and Q_3. Half of the data (50%) falls in the interquartile range. If the *IQR* is "large" the data is spread out and if the *IQR* is "small" the data is closer together.

Definition 3.4.2

Interquartile Range (*IQR*)

***IQR* = Q_3 – Q_1**

Determining probable outliers from IQR: *fences*

A value that is less than $Q_1 - 1.5 \cdot IQR$ (this value is often referred to as a ***lower fence***) is considered an outlier.

Similarly, a value that is more than $Q_3 + 1.5 \cdot IQR$ (the ***upper fence***) is considered an outlier.

A box plot (or box-and-whisker plot) is a graphical display of the five-number summary. It can be drawn vertically or horizontally. The basic format is a box from *Q1* to *Q3*, a vertical line across the box for the median and horizontal lines as whiskers extending out each end to the minimum and maximum. The minimum and maximum can be represented with dots. Don't forget to label the tick marks on the number line and give the graph a title.

If the data are *symmetrical*, then the box plot will be visibly symmetrical. If the data distribution has a left skew or a right skew, the line on that side of the box plot will be visibly long. If the plot is symmetrical, and the four quartiles are all about the same length, then the data are likely a near-*uniform* distribution. If a box plot is symmetrical, and both outside lines are noticeably longer than the Q1 to median and median to Q3 distance, the distribution is then probably *bell-shaped*.

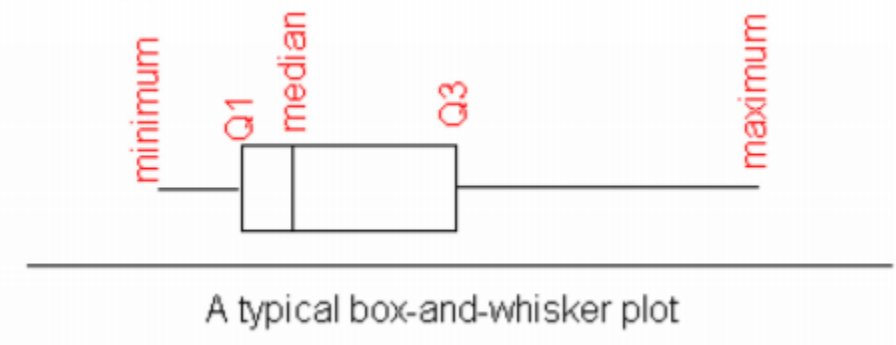


Figure 3.4.1: Typical Box Plot

Example 3.4.5 five-number summary for an even number of data points

The total assets in billions of Australian dollars (AUD) of Australian banks for the year 2012 are given in Example 3.4.1 ("Reserve Bank of," 2013).

1. Find the five-number summary, and draw a box-and-whiskers plot.
2. Find the interquartile range (IQR) and determine if there are any outliers in the raw data.

Table 3.4.1: Total Assets (in billions of AUD) of Australian Banks

2855	2862	2861	2884	3014	2965
2971	3002	3032	2950	2967	2964

Solution

1. Find the five-number summary, and draw a box-and-whiskers plot.

Variable: $x =$ total assets of Australian banks

First sort the data.

Table 3.4.2: Sorted Data for Total Assets

2855	2861	2862	2884	2950	2964	2965	2967	2971	3002	3014	3032

The minimum is 2855 billion AUD and the maximum is 3032 billion AUD.

There are 12 data points so the median is the average of the 6th and 7th numbers.

2855	2861	2862	2884	2950	2964	2965	2967	2971	3002	3014	3032

$$\frac{2964+2965}{2}=2964.5 \text{ billion AUD}$$

Figure 3.4.2: *Sorted Data for Total Assets with Median*

To find Q_1, find the median of the first half of the list.

2855	2861	2862	2884	2950	2964

$Q1$

$$Q1=\frac{2862+2884}{2}=2873 \text{ billion AUD}$$

Figure 3.4.3:: *Finding Q_1*

To find Q_3, find the median of the second half of the list.

2965	2967	2971	3002	3014	3032

$Q3$

$$Q3=\frac{2971+3002}{2}=2986.5 \text{ billion AUD}$$

Figure 3.4.4:: *Finding Q_3*

The five-number summary is (all numbers in billion AUD)

Minimum: 2855

Q_1: 2873

Median: 2964.5

Q_3: 2986.5

Maximum: 3032

You can use the five-number summary to draw the box-and-whiskers plot.

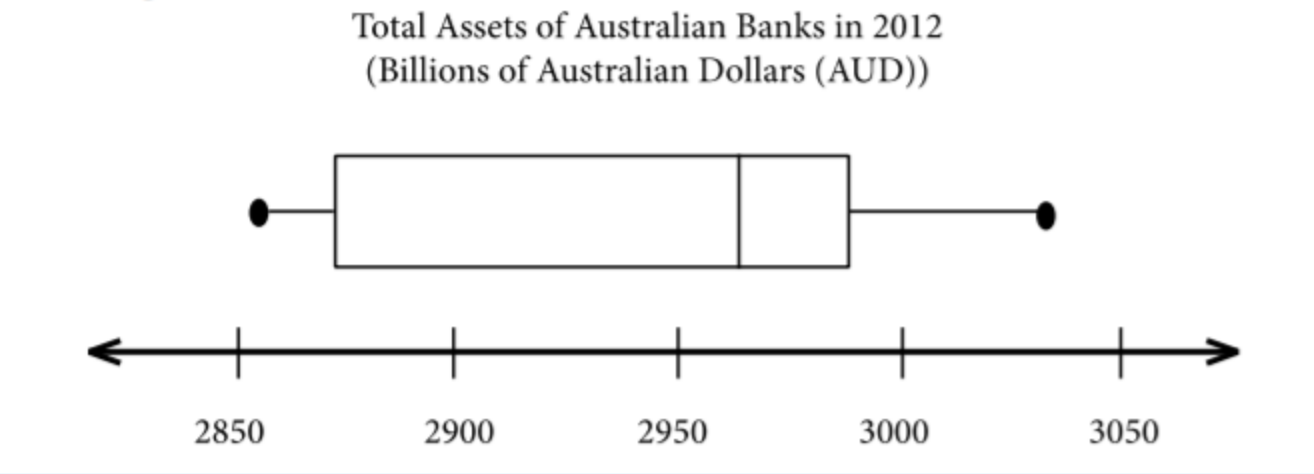


Figure 3.4.5: Box Plot of Total Assets of Australian Banks

The distribution is skewed right because the right tail is longer.

2. Find the interquartile range (IQR) and determine if there are any outliers in the raw data.

   To find the interquartile range, $IQR = Q_3 - Q_1$

   $$IQR = 2986.5 - 2873 = 113.5 \text{ billion AUD}$$

   This tells you the middle 50% of assets were within 113.5 billion AUD of each other.

   $$\text{Lower Fence} = Q_1 - 1.5 \cdot IQR = 2873 - 1.5 \cdot 113.5 = 2873 - 170.25 = 2702.75$$

   $$\text{Upper Fence} = Q_3 + 1.5 \cdot IQR = 2986.5 + 1.5 \cdot 113.5 = 2986.5 + 170.25 = 3156.75$$

2855	2861	2862	2884	2950	2964	2965	2967	2971	3002	3014	3032

   From our arranged data, there is no outlier because there is no number in the raw data below the lower fence, 2702.75, and higher than the upper fence, 3156.75.

Example 3.4.6 five-number summary for an odd number of data points

The life expectancy for a person living in one of 11 countries in the region of South East Asia in 2012 is given below ("Life expectancy in," 2013).

1. Find the five-number summary, and draw a box-and-whiskers plot.
2. Find the interquartile range (IQR) and determine if there are any outliers in the raw data.

Table 3.4.3: Life Expectancy of a Person Living in South-East Asia

70	67	69	65	69	77
65	68	75	74	64	

Solution

1. Find the five-number summary, and draw a box-and-whiskers plot.

Variable: $x =$ life expectancy of a person.

Sort the data first.

Table 3.4.7: Sorted Life Expectancies

64	65	65	67	68	69	69	70	74	75	77

The minimum is 64 years and the maximum is 77 years.

There are 11 data points so the median is the 6th number in the list.

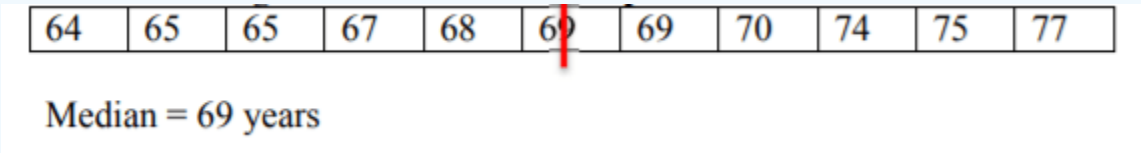


Figure 3.4.6:: *Finding the Median of Life Expectancies*

To find the Q_1 and Q_3 you need to find the median of the numbers below the median and above the median. The median is not included in either calculation.

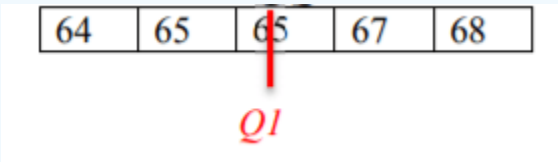


Figure 3.4.7: *Finding Q_1*

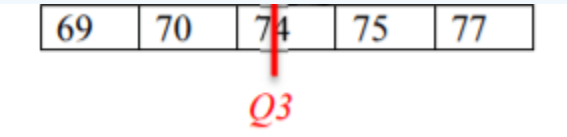


Figure 3.4.8: *Finding Q_3*

Q_1=65 years and Q_3=74 years

The five-number summary is (in years)

Minimum: 64

Q_1: 65

Median: 69

Q_3: 74

Maximum: 77

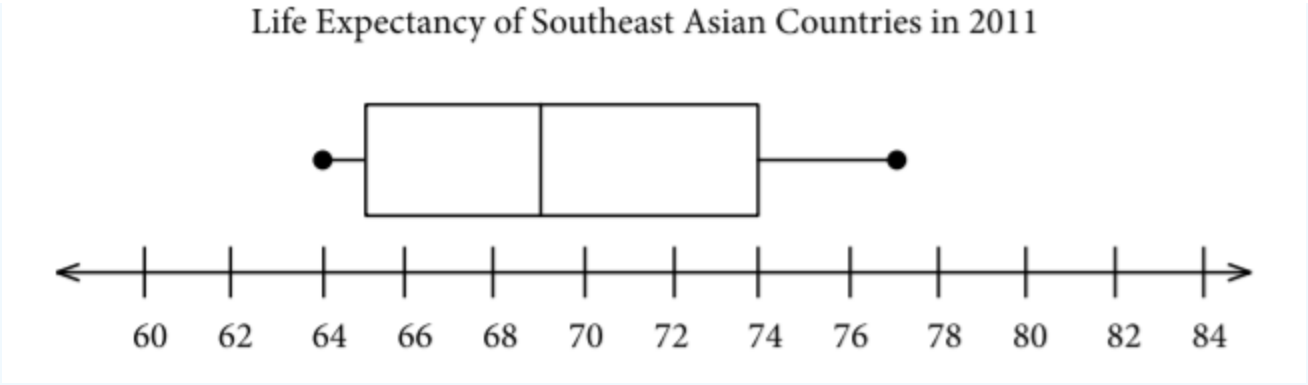


Figure 3.4.9: Box Plot of Life Expectancy

This distribution looks approximately skewed to the right since the whisker is longer on the right. However, it could be considered almost symmetric too since the box looks somewhat symmetric.

2. Find the interquartile range (IQR) and determine if there are any outliers in the raw data.

   To find the interquartile range (*IQR*)

   $$IQR=Q_3-Q_1=74-65=9 \text{ years}$$

   The middle 50% of life expectancies are within 9 years.

   $$\textit{Lower Fence} = Q_1 - 1.5 \cdot IQR = 65 - 1.5 \cdot 9 = 65 - 13.5 = 51.5$$

   $$\textit{Upper Fence} = Q_3 + 1.5 \cdot IQR = 74 + 1.5 \cdot 9 = 74 + 13.5 = 87.5$$

Data For Life Expectancy

64	65	65	67	68	69	69	70	74	75	77

Table 3.4.4: Data for Life Expectancy

From our arranged data, there is no outlier because there is no number in the raw data below the lower fence, 51.5, and higher than the upper fence, 87.5.

You can draw 2 box plots side by side (or one above the other) to compare 2 samples. Since you want to compare the two data sets, make sure the box plots are on the same axes. As an example, suppose you look at the box-and-whiskers plot for life expectancy for European countries and Southeast Asian countries.

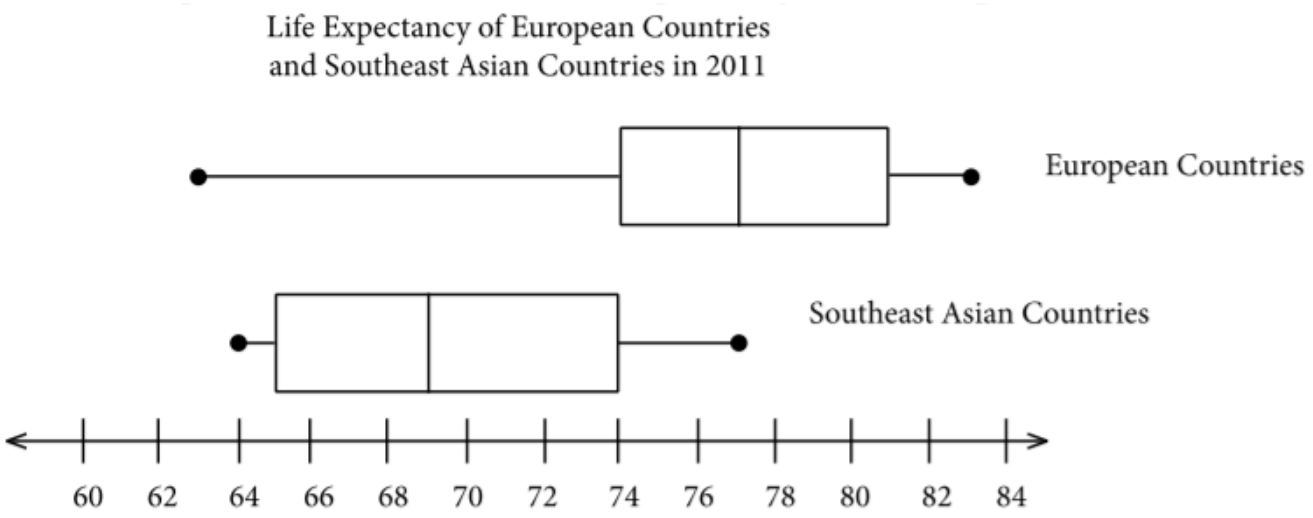


Figure 3.4.10: Box Plots of Life Expectancy of Two Regions

Looking at the box-and-whiskers plot, you will notice that the three quartiles for life expectancy are all higher for the European countries, yet the minimum life expectancy for the European countries is less than that for the Southeast Asian countries. The life expectancy for the European countries appears to be skewed left, while the life expectancies for the Southeast Asian countries appear to be more symmetric. There are of course more qualities that can be compared between the two graphs.

Authors

"3.4: Measures of Position" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"3.3: Ranking" by Kathryn Kozak is licensed CC BY-SA 4.0

This page titled 3.4: Measures of Position is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 3.3: Ranking by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

3.5: Formulas for Chapter 3

Formulas for Chapter 3

Measures of Central Tendency

Mean of a Population

The mean of a population is the average of all the values in an entire population. It is found by adding all the population data values together and dividing by the total number of values in the population. This measure is used to describe the central value of a complete data set.

$$\mu = \frac{\sum x}{N}$$

Where:

- μ is the population mean
- $\sum x$ is the sum of all values in the population
- N is the number of values in the population

Mean of a Sample

The mean of raw data is calculated by adding all the data values and dividing by the total number of values.

$$\bar{x} = \frac{\sum x}{n}$$

Where:

- $\bar{x}$ = the ,mean.
- $\sum x$ = the sum of all data values
- n = the number of values.

Mean for Grouped Data

The mean of grouped data is an estimate of the average for data that has been organized into frequency groups or classes. It is calculated by multiplying each class midpoint by its frequency, adding those products, and dividing by the total frequency.

$$\bar{x} = \frac{\sum f \cdot X_M}{n}$$

Where:

- $\bar{x}$ is the estimated mean
- f is the frequency of each class
- X_M is the midpoint of each class
- $\sum f \cdot x$ is the sum of the products of midpoints and frequencies
- $n = \sum f$ is the total number of data values

Weighted Mean

The mean for weighted data is a type of average where each data value is assigned a weight based on its importance or frequency.

It is calculated by multiplying each value by its weight, adding those products, and then dividing by the total of the weights.

$$\bar{x} = \frac{\sum w \cdot x}{\sum w}$$

Where:

- $\bar{x}$ is the weighted mean
- x is each data value
- w is the weight of each value
- $\sum w \cdot x$ is the sum of the weighted values
- $n = \sum w$ is the sum of the weights

Median for Sample Data

The median, denoted as MD, is the middle value in a data set when the numbers are arranged in order from smallest to largest. To find it, first sort the data. If there is an odd number of values, the median is the value at the midpoint. If there is an even number of values, the median is the average of the two numbers at the midpoint. The median is useful because it represents the center of the data and is not affected by extreme values.

Midrange for Sample Data

The midrange, denoted as MR, is the value halfway between the smallest and largest numbers in a data set. It is found by adding the minimum and maximum values and dividing the result by 2. The midrange gives a quick sense of the central location of the data and is easy to calculate.

$$MR = \frac{\text{Minimum value} + \text{Maximum value}}{2}$$

Mode for Sample Data

The mode is the value that appears most often in a data set. It shows which item occurs the most and is useful for identifying the most common or popular value. A data set can have one mode, more than one, or none at all if no number repeats.

Measures of Variation

Range of Sample Data

The range is a simple measure of variation that shows the difference between the highest and lowest values in a data set. It gives a quick sense of how spread out the data is, but does not account for how values are distributed between the extremes.

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Variance for Sample Data: Traditional Formula

The sample variance for raw data is a measure of how much the values in a sample differ from the sample mean. It represents the average of the squared differences between each value and the mean and is used to understand the variability within a sample.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Where:

- s^2 is the sample variance
- x is each data value
- $\bar{x}$ is the sample mean

- n is the number of data values
- $\sum (x - \bar{x})^2$ is the sum of squared differences from the mean

Variance for Sample Data: Shortcut Formula

The sample variance using the shortcut formula is an alternative method to calculate variance without first computing the mean. It is useful when working with raw data, especially large data sets, as it simplifies calculations by using the sum of the values and the sum of the squared values.

$$s^2 = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$$

Where:

- s^2 is the sample variance
- x is each data value
- n is the number of data values
- $\sum x$ is the sum of all data values
- $\sum x^2$ is the sum of the squares of the data values

Variance for Grouped Data

The grouped variance is an estimate of the variance for data that has been organized into frequency groups or classes. It measures how spread out the data is by using the class midpoints and their frequencies to approximate the average squared deviation from the mean.

$$s^2 = \frac{n\sum f \cdot (X_M)^2 - (\sum f \cdot X_M)^2}{n(n-1)}$$

Where:

- s^2 is the sample variance
- f is the frequency of each class
- X_M is the midpoint of each class
- $\sum fx$ is the sum of the products of frequency and midpoint
- $\sum fx^2$ is the sum of the products of frequency and the square of the midpoint
- n is the total number of data values

Standard Deviation for Sample and Grouped Data

The standard deviation is a measure of how spread out the values in a data set are around the mean. It shows the average distance of each data point from the mean and is useful for understanding the consistency or variability of the data. It is calculated by taking the square root of the variance. This applies whether the variance is from a population, a sample, or grouped data.

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Where:

- $\sigma = \sqrt{\sigma^2}$ for population standard deviation
- $s = \sqrt{s^2}$ for sample or grouped sample standard deviation

Standard Deviation of the Population

The population standard deviation is a measure of how spread out the values in an entire population are from the population mean. It shows the average distance of each data point from the mean and is used to understand variability in a full data set.

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Where:

- σ is the population standard deviation
- x is each data value
- μ is the population mean
- N is the number of values in the population
- $\sum (x - \mu)^2$ is the sum of squared differences from the mean

Measures of Position

Z-Score

A z-score tells you how many standard deviations a data value is from the mean. It helps compare values from different data sets or identify how unusual a value is within a distribution. A positive z-score means the value is above the mean, while a negative z-score means it is below the mean.

$$z = \frac{x - \bar{x}}{s}$$

- z is the z-score
- x is the data value
- $\bar{x}$ is the sample mean
- s is the sample standard deviation

Percentiles

Percentiles are values that divide a data set into 100 equal parts, showing the relative standing of a value within the data. Each percentile indicates the percentage of data that falls below it. For example, the 70th percentile means that 70% of the data values are less than or equal to that point.

Quartiles

Quartiles divide a data set into four equal parts after the values have been arranged in order from least to greatest. The three quartiles mark the boundaries of these parts. The first quartile, or Q_1, is the value that separates the lowest 25% of the data from the rest. The second quartile, or Q_2, is the median, which splits the data in half. The third quartile, or Q_3, marks the point where 75% of the data falls below it. To find the quartiles, you first sort the data set. Then, find the median (Q_2), and use the lower half of the data to find Q_1 and the upper half to find Q_3.

Interquartile Range

The interquartile range (IQR) is a measure of statistical spread that shows the range of the middle 50% of a data set. It is calculated by subtracting the first quartile (Q1) from the third quartile (Q3). The IQR helps identify the spread of the central portion of the data and is useful for detecting outliers.

$$IQR = Q_3 - Q_1$$

Authors

"3.5: Formulas for Chapter 3" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 3.5: Formulas for Chapter 3 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

3.6: Chapter 3 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Interquartile range (IQR): the difference between the third quartile (Q3) and the first quartile (Q1). It measures the spread of the middle 50% of the data.

Mean: the arithmetic average of the numbers. This is the center that most people call the average, though all three – mean, median, and mode – really are averages.

Median: the data value in the middle of a sorted list of data. To find it, you put the data in order and then determine which data value is in the middle of the data set.

Midrange: the sum of the least and the greatest value in the data set, divided by 2.

Mode: the data value that occurs the most frequently in the data. To find it, you count how often each data value occurs, and then determine which data value occurs most often.

Percentile: a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors also use percentiles to track a child's growth.

k^{th} percentile: the data value that has k% of the data at or below that value.

Quartiles: values that divide a data set into four equal parts, each containing 25% of the data. They help describe the spread and center of the data.

First quartile: the value that marks the 25th percentile of a data set. It separates the lowest 25% of the data from the rest and is the **median of the lower half** of the data.

Second quartile: the median of the data set. It marks the 50th percentile, meaning half of the data values are less than or equal to Q2, and the other half are greater.

Third quartile: the value at the 75th percentile of a data set. It separates the lowest 75% of the data from the highest 25% and is the median of the upper half of the data.

Sample variance: a statistical measure that shows how spread out the values in a sample are from the sample mean. It estimates the variability or dispersion within a sample data set.

Standard deviation: a measure of how **spread out** or **dispersed** the values in a data set are from the **mean** (average). It tells you, on average, how much the data values **differ** from the means.

Z-score: measures how far a data point is from the average (mean) of a data set, measured in standard deviations. It helps compare values from different distributions by showing if a number is above or below the average and by how much. Z-scores help standardize values so they can be compared, even if they come from different groups or scales.

Symbols

$\bar{x}$ - sample mean

μ - population mean

x_M - midpoint

MD - median

MR - midrange

Mode - mode

P_k - k-th percentile

Q_1 - first quartile

Q_2 - second quartile (median)

Q_3 - third quartile

R - range

s - sample standard deviation

s^2 - sample variance

σ - population standard deviation

σ^2 - population variance

w - weight

z - z-score

Authors

"3.6: Chapter 3 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 3.6: Chapter 3 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

4: PROBABILITY AND COMBINATORICS

4.1: Empirical Probability

4.2: Theoretical Probability

4.3: Addition Rules for Probabilities

4.4: Multiplication Rules and Conditional Probability

4.5: At Least One Rule and Tree Diagrams

4.6: Counting Rules

4.7: Chapter 4 Formulas

4.8: Chapter 4 - Key Terms and Symbols

This page titled 4: Probability and Combinatorics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

4.1: Empirical Probability

Learning Objectives

- Define empirical probability as based on observed outcomes from experiments rather than theoretical calculations.
- Calculate empirical probability using the ratio of favorable outcomes to total trials.
- Understand the law of large numbers and its role in stabilizing probabilities with increased trials.
- Define key terms such as sample space (all possible outcomes) and event space (outcomes of interest).

One of the earliest stories about probability comes from a gambler who wanted to know when to take risks and when to hold back. He turned to two mathematicians, Pierre de Fermat and Blaise Pascal, who began studying patterns in chance and uncertainty. Their work laid the foundation for what we now call probability, and many others have continued to develop these ideas over time.

Probability is useful in everyday life, even if we don't always realize it. It can help answer simple questions like whether to bring an umbrella or wear a jacket. Still, it also helps with bigger decisions, such as understanding the risk of buying a used car, passing a class, or even being involved in a car accident. It can also explain why events like winning the lottery are so rare.

Many people tend to misunderstand probability. For example, some worry about rare events like terrorist attacks but don't think much about more common risks like car accidents. In reality, the chance of being in a car accident is much higher. Similarly, even though the chances of winning the lottery are very small, people still spend money on tickets. Understanding probability helps us make more informed and rational decisions.

In general, events with a very low probability (less than 5%) are considered unlikely, while events with a high probability (greater than 80%) are likely to occur. This chapter will introduce key ideas that will help you better understand chance and make sense of uncertainty in everyday situations.

Key Definitions

Definition 4.1.1

Experiment: an activity or process that will produce specific outcomes.

Examples) Tossing one die, selecting a card from a standard deck, flipping 1 coin, etc.

Definition 4.1.2

Outcomes: the result of an experiment.

Example) When a die is flipped, it can lead to six possible outcomes: 1, 2, 3, 4, 5, or 6. These are the outcomes of the experiment.

Definition 4.1.3

Event: a set of certain outcomes of an experiment that you want to have happen.

Example) A die is tossed; one event can be getting an odd outcome. The possibilities are 1, 3, or 5.

Definition 4.1.4

Sample Space: a collection of all possible outcomes of the experiment. It is denoted as S.

Example) A coin is flipped; the sample space is S = {Heads, Tails}.

Definition 4.1.5

Event Space: the set of outcomes that make up an event. The symbol is usually a capital letter.

Example) A die is tossed; one event can be getting an odd outcome. The event space is E = {1, 3, 5}.

Visual Example of Definitions

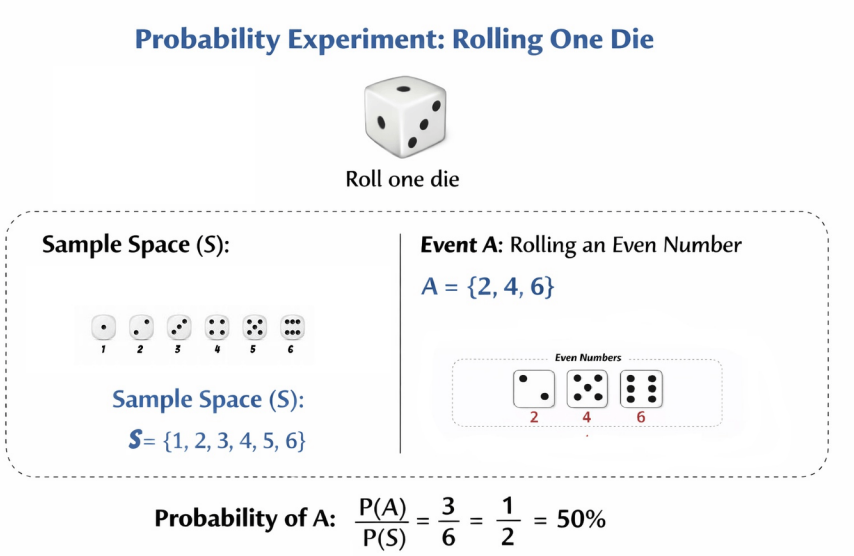


Figure 4.1.1: This figure illustrates a simple probability experiment involving the roll of a single die. The sample space consists of all possible outcomes, S = {1, 2, 3, 4, 5, 6}. An example event is rolling an even number, defined as A = {2, 4, 6}. Since there are 3 favorable outcomes out of 6 possible outcomes, the probability of event A is $\frac{3}{6} = \frac{1}{2}$, or 50%.

Empirical Probability (Relative Frequency) Rule

Start with an experiment. Suppose we roll a die. The sample space is {1, 2, 3, 4, 5, 6}, and the event of interest is getting a 6, so the event space is {6}. Now roll the die 10 times. If a 6 appears 2 times, then the estimated probability of getting a 6 is 2 out of 10, or 1/5. To obtain a more accurate estimate, repeat the experiment many more times. It is helpful to organize the results in a table, where n represents the number of trials. The ratio of the number of 6s to the total number of trials is called the relative frequency.

Table 4.1.1: Trials for Die Experiment

n	Number of 6s	Relative Frequency
10	2	0.2
50	6	0.12
100	18	0.18
500	81	0.162

n	Number of 6s	Relative Frequency
1000	163	0.163

Notice that as n increases, the relative frequency approaches a number. It looks like it is approaching 0.163. You can say that the probability of getting a 6 is approximately 0.163. If you want more accuracy, then increase n even more.

These probabilities are called **experimental (empirical) probabilities** since they are found by experimentation. They come about from the relative frequencies and give an approximation of the true probability. The approximate probability of an event A, $P(A)$, is

Definition 4.1.6

Experimental Probabilities

$$P(A) = \frac{\text{number of times } A \text{ occurs}}{\text{number of times the experiment was repeated}}$$

For the event of getting a 6, the probability would be $\frac{163}{1000} = 0.163$.

Experimental probabilities are done whenever probabilities cannot be calculated using other means. For example, if you want to find the probability that a family has 5 children, you would look at many families and count how many have 5 children. Then you could calculate the probability. Another example is if you want to figure out if a die is fair. You would roll the die many times and count how often each side comes up. Make sure you repeat an experiment many times because otherwise, you won't be able to estimate the true probability. This is due to the law of large numbers.

Definition 4.1.7

Law of large numbers: as n increases, the relative frequency tends towards the actual probability value.

Note

Probability, relative frequency, percentage, and proportion are all different words for the same concept. Also, probabilities can be given as percentages, decimals, or fractions.

Exercises

Exercise 4.1.1

A student tracks how often they get a parking spot close to campus. Over 20 days, they find a close spot 6 times.

a. What is the empirical probability of getting a close parking spot?'
b. Based on this, how likely is it that they will find a close spot tomorrow?

Answer available at the end of this section.

Exercise 4.1.2

A coffee shop records customer drink choices during the morning rush. Out of 50 customers, 18 ordered iced coffee.

a. What is the empirical probability that a randomly selected customer orders iced coffee?

b. Would you consider this a common or uncommon choice?

Answer available at the end of this section.

Exercise 4.1.3

A student flips a coin 30 times and records 19 heads and 11 tails.

a. What is the empirical probability of getting heads?
b. How does this compare to the theoretical probability of heads?

Answer available at the end of this section.

Exercise 4.1.4

A student tracks how often they arrive at class on time. Over 25 class meetings, they arrive on time 20 times.

a. What is the empirical probability of arriving at class on time?
b. Would you consider this a common or uncommon choice?

Answer available at the end of this section.

Exercise 4.1.5

A music app tracks how often a user skips songs. Out of 40 songs played, the user skips 10 songs.

a. What is the empirical probability of a user skipping a song?
b. Would you consider this a common or uncommon choice?

Answer available at the end of this section.

Authors

"4.1: Empirical Probability" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"4.1: Empirical Probability" by Kathryn Kozak is licensed under CC BY-SA 4.0

Answers

Exercise 4.1.1

a. Empirical probability = $\frac{6}{20} = 0.30$.
b. There is a 30% chance of finding a close parking spot, so it is somewhat unlikely.

Exercise 4.1.2

a. Empirical probability = $\frac{18}{50} = 0.36$.
b. There is a 36% chance a customer orders iced coffee, so it is a fairly common choice.

Exercise 4.1.3

a. Empirical probability of heads = $\frac{19}{30} \approx 0.63$.
b. The theoretical probability is 0.50, so this result is higher than expected, but small samples can vary.

Exercise 4.1.4

a. Empirical probability = $\frac{20}{25} = 0.80$.
b. There is an 80% chance the student arrives on time, so this is very likely.

Exercise 4.1.5

a. Empirical probability = $\frac{10}{40} = 0.25$.
b. There is a 25% chance a song will be skipped, so skipping is somewhat uncommon.

This page titled 4.1: Empirical Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 4.1: Empirical Probability by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

4.2: Theoretical Probability

Learning Objectives

- Define theoretical probability as based on known, equally likely possible outcomes.
- Understand that probabilities range from 0 (impossible) to 1 (certain).
- Apply key rules such as the complement rule: the probability of an event not occurring equals 1 minus the probability that it does occur.

Theoretical probability is the branch of probability that calculates the likelihood of an event occurring based on **mathematical reasoning and logical analysis**, rather than on actual experiments or observations. It assumes that all outcomes in a sample space are equally likely.

What is a Sample Space?

In probability theory, a **sample space** (**S**) is the **set of all possible outcomes** of a random experiment. It represents every outcome that could occur when an experiment is performed.

Examples of Sample Spaces:

1. **Flipping a Coin** 🪙

- Possible outcomes: **{Heads, Tails}**
- Sample space: S = {H, T}

2. **Flipping Two Coins** 🪙🪙

- Possible outcomes: **{Heads and Heads, Heads and Tails, Tails and Heads, Tails and Tails}**
- Sample space: S ={HH, HT, TH, TT}

3. **Rolling a Six-Sided Die:**🎲

- Possible outcomes: **1, 2, 3, 4, 5, 6**
- Sample space: S = {1, 2, 3, 4, 5, 6}

4. **Rolling Two Six-Sided Dice** 🎲🎲

- Possible Outcomes—It is best to write them out as ordered pairs. The first value is an outcome for die#1, and the second value is an outcome for die#2. There are 36 outcomes in total. See the table below.

Six-Sided Die

Six-Sided Die	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3.6)	(4,6)	(5,6)	(6,6)

Figure 4.2.1:This table shows the sample space for rolling two six-sided dice. The numbers across the top represent the outcomes of the first die, and the numbers along the left side represent the outcomes of the second die. Each cell contains an ordered pair (x, y), where x is the result of the first die and y is the result of the second die. Since each die has 6 possible outcomes, there are $6 \times 6 = 36$ equally likely outcomes in total. This table can be used to determine probabilities of different events. For example, to find the probability of rolling a sum of 7, count the outcomes that add to 7: $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$. There are 6 such outcomes out of 36, so the probability is $\frac{6}{36} = \frac{1}{6}$. This visual helps identify outcomes and calculate probabilities for sums, doubles, or other events involving two dice.

5. **Drawing a Card from a Standard Deck**

- Sample space: **All 52 cards** in a deck. The table below lists all 52 outcomes. Jokers are excluded, and the deck consists of four suits, each containing 13 cards.

Sample Space for a Deck of Cards

Hearts (♥)	Clubs (♣)	Diamonds (♦)	Spades (♠)
A♥	A♣	A♦	A♠
2♥	2♣	2♦	2♠
3♥	3♣	3♦	3♠
4♥	4♣	4♦	4♠
5♥	5♣	5♦	5♠
6♥	6♣	6♦	6♠
7♥	7♣	7♦	7♠
8♥	8♣	8♦	8♠
9♥	9♣	9♦	9♠
10♥	10♣	10♦	10♠
J♥	J♣	J♦	J♠
Q♥	Q♣	Q♦	Q♠
K♥	K♣	K♦	K♠

Table 4.2.1: The deck contains four suits—hearts, diamonds, clubs, and spades. Hearts and diamonds are red, while clubs and spades are black. Each suit includes thirteen ranks: ace, numbers two through ten, and the face cards jack, queen, and king. The cards are rectangular with white backgrounds, suit symbols (pips), and corner indices showing the rank and suit. Face cards

display illustrated portraits, while numbered cards show repeated suit symbols arranged in patterns. The deck may appear stacked, fanned, or spread to show multiple cards.

Probability Properties

1. The probability of any event in the sample space is between 0 and 1, inclusive, as shown here $0 \leq P(\text{ event }) \leq 1$
2. If the P(event) = 1, then it will happen and is called a certain event.
3. If the P(event) = 0, then it cannot happen and is called an impossible event.
4. The sum of all the probabilities in the sample space is one, as shown here $\sum P(\text{ outcome }) = 1$

Example 4.2.1 calculating theoretical probabilities

Suppose you conduct an experiment where you pull a card from a standard deck. Write the answers as a fraction or a decimal rounded to three place values.

a. What is the probability of getting a Spade?
b. What is the probability of getting a Jack?
c. What is the probability of getting a Spade and an Ace?
d. What is the probability of getting a Spade or an Ace?
e. What is the probability of getting a Jack and an Ace?
f. What is the probability of getting a Jack or an Ace?

Solutions

a. Let A = getting a spade = {2♠, 3♠, 4♠, 5♠, 6♠, 7♠, 8♠, 9♠, 10♠, J♠, Q♠, K♠, A♠} therefore,

$$P(A) = \frac{13}{52} = 0.25$$

b. Let B = getting a Jack = {J♠, J♣, J♥, J♦} therefore,

$$P(B) = \frac{4}{52} = 0.077$$

c. Let E = getting a Spade and an Ace = {A♠} therefore,

$$P(E) = \frac{1}{52} = 0.019$$

d. Let F = getting a Spade or an Ace ={2♠, 3♠, 4♠, 5♠, 6♠, 7♠, 8♠, 9♠, 10♠, J♠, Q♠, K♠, A♠, A♣, A♦, A♥} therefore,

$$P(F) = \frac{16}{52} = 0.308$$

e. Let G = getting a Jack and an Ace = { } since you can't do that with one card. therefore,

$$P(G) = 0$$

f. Let H = getting a Jack or an Ace = {J♠, J♣, J♦, J♥, A♠, A♣, A♦, A♥} therefore,

$$P(H) = \frac{8}{52} = 0.154$$

Example 4.2.2 calculating theoretical probabilities

Assume that two dice are tossed and we desire to compute the probability of an outcome. Typically, the outcomes are computed by adding the two numbers (represented by the pips/dots on the sides of the dice) together. Assume that one pair of dice is tossed one time. Compute the following probabilities. Write the answers as a fraction or a decimal rounded to three place values.

a. The sum is 5.
b. The sum is less than or equal to 5.
c. The sum is 4 or 9.
d. The sum is 13.
e. The probability of getting doubles.

Solutions

a. Let A = getting a sum of 5 = {(1,4), (2,3), (3,2), (4,1)} therefore,

$$P(A) = \frac{4}{36} = 0.111$$

b. Let B = sum less than 5 = {(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (3,1), (3,2), (4,1)} therefore,

$$P(B) = \frac{10}{36} = 0.278$$

c. Let C = sum is 4 or 9. = {(1,3), (2,2), (3,1), (3,6), (4,5), (5,4), (6,3)} therefore,

$$P(C) = \frac{7}{36} = 0.194$$

d. Let D = sum is 13 = { } since there is no sum of 13 for two dice, therefore,

$$P(D) = 0$$

e. Let E = getting doubles = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}

$$P(E) = \frac{6}{36} = 0.167$$

Example 4.2.3

A liberal arts club is organizing a cultural showcase. Students randomly select one event ticket from a box. The box contains the following tickets:

- 5 Music Performance tickets
- 4 Poetry Reading tickets
- 3 Art Exhibition tickets

a. What is the probability that a student selects a poetry reading ticket?
b. What is the probability that a student selects a music performance ticket?

Solution

a. Probability of selecting a Poetry Reading ticket

P(Poetry)= $\frac{4}{12} = \frac{1}{3}$

b. Probability of selecting a Music Performance ticket

P(Music)=$\frac{5}{12}$

Complementary Events

In probability, complementary events are pairs of outcomes that cover all possibilities but cannot happen at the same time. If one event occurs, the other cannot. For example, flipping a coin results in either heads or tails—these are complementary events. The key idea is that the probability of an event and its complement always adds up to 1, which makes complements useful when it's easier to calculate the chance of something not happening.

Definition 4.2.2: complementary events

If two events are **complementary events,** then to find the probability of one, subtract the probability of the other from one. The notation used to complement A is not A or $\bar{A}$.

$$P(A)+P\left(\bar{A}\right)=1, \text{ or } P(\bar{A})=1-P(A) \text{ where } P(\bar{A})=P(\text{not } A)$$

Example 4.2.4 complementary events

a. Suppose you know that the probability of it raining today is 0.45. What is the probability of it not raining?
b. Suppose you know the probability of not getting the flu is 0.24. What is the probability of getting the flu?
c. In an experiment of picking a card from a deck, what is the probability of not getting a card that is a Queen?

Solutions

a. Since not raining is the complement of raining, then

$$P(\text{ not raining })=1-P(\text{ raining })=1-0.45=0.55$$

b. Since getting the flu is the complement of not getting the flu, then

$$P(\text{ getting the flu })=1-P(\text{ not getting the flu })=1-0.24=0.76$$

c. You could solve this problem by listing all the ways to not get a queen, but that set is fairly large. One advantage of the complement is that it reduces the workload. You use the complement in many situations to make the work shorter and easier. In this case it is easier to list all the ways to get a Queen, find the probability of the Queen, and then subtract from one. Queen = {Q♠, Q♣, Q♦, Q♥} so

$$P(\text{ Queen })=\frac{4}{52} \text{ and}$$

$$P(\text{ not Queen })=1-P(\text{ Queen })=1-\frac{4}{52}=\frac{48}{52}$$

Example 4.2.5 using the complement to find probabilities

a. In an experiment of rolling a fair die once, find the probability of rolling a 4 on the die. Write the answer as a fraction or a decimal number rounded to three places.
b. In an experiment of pulling a card from a fair deck, find the probability of not getting a jack. Write the answer as a fraction or a decimal number rounded to three places.

Solutions

a. The sample space for this experiment is {1, 2, 3, 4, 5, 6}. You want the event of getting at least a 4, which is the same as thinking of getting 4 or more The event space is {4, 5, 6}. The probability is

$$P(\text{ at least } 4) = \frac{3}{6} = 0.5$$

b. The sample space for this experiment is the entire deck of cards.

Getting a card that is not a jack would involve listing all of the cards that are not jacks; it would be much easier to list the outcomes that make up the complement. The complement is all the jacks. The event space for the complement would be {J♠, J♣, J♦, J♥}. The probability of the complement would be $\frac{4}{52}$. The probability of not getting a jack is

$$P(\text{ not a jack }) = 1 - \frac{4}{52} = \frac{48}{52} = 0.923$$

Exercises

Exercise 4.2.1

Two fair six-sided dice are rolled at the same time. Each die has six possible outcomes: 1, 2, 3, 4, 5, and 6.

a. What is the probability that the sum of the two dice is 8?
b. What is the probability that the sum of the two dice is 10?
c. What is the probability that the sum of the two dice is greater than 9?

Answer available at the end of this section.

Exercise 4.2.2

A standard deck of playing cards contains 52 cards. The deck has four suits: hearts, diamonds, clubs, and spades. Each suit contains 13 cards (Ace through King).
One card is randomly selected from the deck.

a. What is the probability of selecting a heart?
b. What is the probability of selecting a king?
c. What is the probability of selecting a black card (clubs or spades)?

Answer available at the end of this section.

Exercise 4.2.3

An economics professor prepares a box of topic cards for a class activity. The cards represent different areas of economics:

- Macroeconomics: 6 cards
- Microeconomics: 5 cards
- International Economics: 3 cards
- Behavioral Economics: 2 cards

A student randomly selects one card from the box.

a. What is the probability that the student selects a Macroeconomics card?
b. What is the probability that the student selects a Microeconomics card?
c. What is the probability that the student selects either an International Economics card or a Behavioral Economics card?

Answer available at the end of this section.

Exercise 4.2.4

A survey is conducted on a college campus. One student is selected at random. The results show that 60% of students participate in a campus club, 50% work part-time, and 30% do both.

a. What is the probability that the selected student participates in a club and works part-time?
b. What is the probability that the selected student participates in a club or works part-time?
c. What is the probability that the selected student does neither activity?

Answer available at the end of this section.

Exercise 4.2.5

A standard deck of playing cards contains 52 cards. The deck has four suits: hearts, diamonds, clubs, and spades. There are 13 cards in each suit.
One card is randomly selected from the deck.

a. Find the probability of not selecting a heart.
b. Find the probability of not selecting a king.
c. Find the probability of not selecting a red card.

Answer available at the end of this section.

Authors

"4.2: Theoretical Probability" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"4.2: Theoretical Probability" by Kathryn Kozak is licensed under CC BY-SA 4.0

Answers

Exercise 4.2.1

a. The outcomes that sum to 8 are $(2,6),(3,5),(4,4),(5,3),(6,2)$, so there are 5 outcomes. $P(\text{sum } = 8) = \frac{5}{36}$
b. The outcomes that sum to 10 are $(4,6),(5,5),(6,4)$, so there are 3 outcomes. $P(\text{sum } = 10) = \frac{3}{36} = \frac{1}{12}$
c. Sums greater than 9 are 10, 11, and 12. The outcomes are: 10 → 3 outcomes, 11 → 2 outcomes, 12 → 1 outcome (total = 6). $P(\text{sum } > 9) = \frac{6}{36} = \frac{1}{6}$

Exercise 4.2.2

a. There are 13 hearts out of 52 cards. $P(\text{heart}) = \frac{13}{52} = \frac{1}{4}$
b. There are 4 kings out of 52 cards. $P(\text{king}) = \frac{4}{52} = \frac{1}{13}$
c. There are 26 black cards (clubs and spades) out of 52 cards. $P(\text{black card}) = \frac{26}{52} = \frac{1}{2}$

Exercise 4.2.3

a. Total cards = $6 + 5 + 3 + 2 = 16$. $P(\text{Macroeconomics}) = \frac{6}{16} = \frac{3}{8}$

b. $P(\text{Microeconomics}) = \frac{5}{16}$

c. Favorable cards = $3 + 2 = 5$. $P(\text{International or Behavioral}) = \frac{5}{16}$

Exercise 4.2.4

a. $P(\text{club and work}) = 0.30$

b. $P(\text{club or work}) = 0.60 + 0.50 - 0.30 = 0.80$

c. $P(\text{neither}) = 1 - 0.80 = 0.20$

Exercise 4.2.5

a. $P(\text{not a heart}) = 1 - P(\text{heart}) = 1 - \frac{13}{52} = \frac{39}{52} = \frac{3}{4}$

b. $P(\text{not a king}) = 1 - P(\text{king}) = 1 - \frac{4}{52} = \frac{48}{52} = \frac{12}{13}$

c. $P(\text{not a red card}) = 1 - P(\text{red}) = 1 - \frac{26}{52} = \frac{26}{52} = \frac{1}{2}$

This page titled 4.2: Theoretical Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Theoretical Probability by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

4.3: Addition Rules for Probabilities

Learning Objectives

- Understand the two addition rules for probability to calculate the chance of either one event or another occurring.
- Apply the addition rule for mutually exclusive events by summing their probabilities.
- Apply the addition rule for events that can occur together by adding their probabilities and subtracting the overlap to avoid double-counting.

When computing the probability that at least one of several events occurs, addition rules can be applied to simplify the calculation. For example, if a card is selected from a standard deck, what is the chance that the card is a jack or a queen? This probability can be computed rapidly by computing the probability of getting a jack and a queen separately and then adding the two probabilities. This process can be done because the events are mutually exclusive. In other words, the events have no outcomes in common.

Definition: 4.3.1

Mutually exclusive events (disjoint) are events with no outcomes in common.

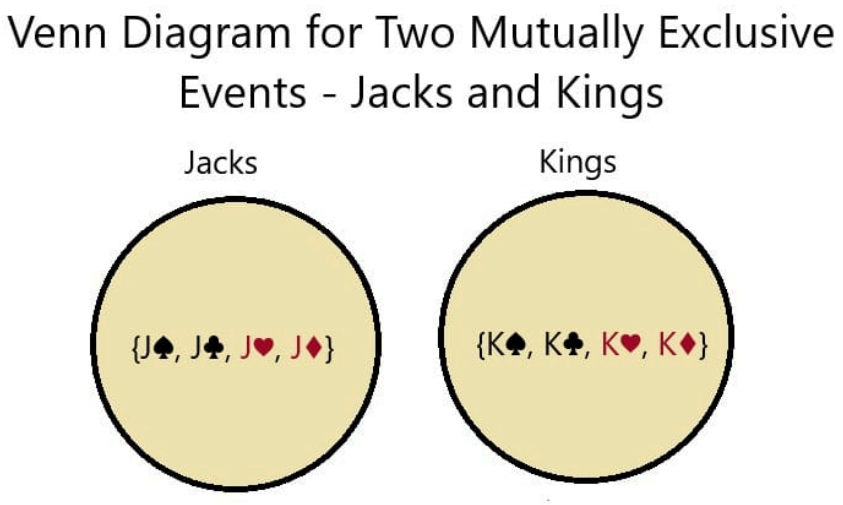


Figure 4.3.1: Two Mutually Exclusive Events

Example 4.3.1

1. Roll a single die. The events are to get a 2 or a 5.
2. Select a single card. The event is to get a jack or a queen.
3. A room has 5 math majors, 12 chemistry majors, 6 biology majors, and 4 physics majors. A single person is selected, and no person is double-majoring. The person is a math major or a biology major.

Solution

None of the events has outcomes in common.

Definition: 4.3.2

Addition Rule I

For two events A and B that are mutually exclusive, the probability that A or B will occur is P(A or B) = P(A) + P(B).

Example 4.3.2

Use the addition rule I to compute the following probabilities. Write the final answers as fractions or decimals rounded to three places.

1. Roll a single die. What is the chance of getting a 2 or a 5?
2. Select a single card. What is the chance the card is a jack or a queen?
3. A room has 5 math majors, 12 chemistry majors, 6 biology majors, and 4 physics majors. A single person is selected, and no person is double-majoring. What is the chance that the person is a math or biology major?

Solutions

1. P(2 or 5) = $\frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} = 0.333$
2. P(ace or king) = $\frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13} = 0.154$
3. P(math major or biology major) = $\frac{5}{27} + \frac{6}{27} = \frac{11}{27} = 0.407$

Definition: 4.3.3

Two events are **not mutually exclusive** if they **can happen at the same time**, meaning they have some outcomes in common.

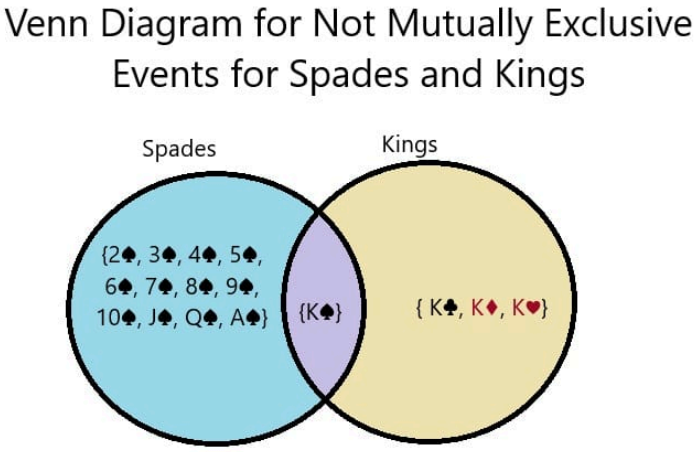


Figure 4.3.2: Two Not Mutually Exclusive Events

When two or more events have outcomes in common, they are labeled non-mutually exclusive events. These types of events have one or more outcomes in common. Care must be taken not to over-count these outcomes. For example, suppose a card is selected from a deck. What is the chance that the card is a jack or a black card? If the addition rule 1 is applied, this would result in the wrong answer. The reason is that two blackjacks are being over-counted. To correctly compute the probability, these extra two blackjacks must be subtracted from the total number of outcomes. This process is outlined in Addition Rule II.

Definition: 4.3.4

Addition Rule II

For two events A and B that are not mutually exclusive, the probability that A or B will occur is P(A or B) = P(A) + P(B) – P(A and B)

Example 4.3.3

Use the addition rule II to compute the following probabilities. Write the final answers as fractions or decimals rounded to three place places.

1. Select a single card. What is the chance the card is a jack or a black card?
2. Select a single card. What is the chance the card is a king or a diamond?
3. A room has 5 math majors, 12 chemistry majors, 6 biology majors, and 4 physics majors. In this group, 2 people are double-majoring in math and physics. A single person. What is the chance that the person is a math or physics major?

Solutions

1. P(Jack or black) = $\frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13} = 0.538$
2. P(king or diamond) = $\frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} = 0.308$
3. P(math major or physics major) = $\frac{5}{27} + \frac{4}{27} - \frac{2}{27} = \frac{7}{27} = 0.259$

Sometimes data is presented in a tabular form. This type of table is called a contingency table, and the table is used to show a relationship between two data types. The data are listed in rows and columns. We can treat the contingency table as two sets where the column data types represent one set, and the row data types represent another data type. To compute probabilities using a contingency table, the first step is to find the sum of each row and column. For example, students are categorized into those who don't work, work part-time, and work full-time, along with the modality of the class they are taking for Math 165: regular, with support, hybrid, or online. The contingency table is displayed below for 120 randomly selected students.

Contingency Table Involving Modality of Classes and Work Status

	Regular	With Support	Hybrid	Online	Total
Does not work	15	14	5	3	37
Works part-time	9	16	10	5	40
Works full-time	7	8	13	15	43
Total	31	38	28	23	120

Table 4.3.1: Contingency Table Involving Modality of Classes and Work Status

Example 4.3.4

Using the contingency table above, compute the probability for one randomly selected student. Write the final answers as fractions or decimals rounded to three places.

1. The student works full-time or takes a class online.
2. The student is part-time and takes a hybrid class.
3. The student does not work or take a class with support.

Solutions

1. P(works full-time or takes an online class) = $\frac{43}{120} + \frac{23}{120} - \frac{15}{120} = \frac{51}{120} = 0.425$
2. P(part-time and takes a hybrid class) = $\frac{10}{120} = 0.083$
3. P(Does not work or takes a class with support) = $\frac{37}{120} + \frac{38}{120} - \frac{14}{120} = \frac{61}{120} = 0.508$

Exercises

Exercise 4.3.1

A single fair six-sided die is rolled.

- Event A: Rolling a 2
- Event B: Rolling a 5

Since a die can only land on one number at a time, these events are mutually exclusive.
What is the probability of rolling a 2 or a 5?

Answer available at the end of this section.

Exercise 4.3.2

A psychology researcher is studying students' preferred study methods. In a group of students:

- 35% prefer studying alone
- 25% prefer studying in a group

Assume that no student prefers both methods at the same time. What is the probability that a randomly selected student prefers studying alone OR in a group?

Answer available at the end of this section.

Exercise 4.3.3

At a college, students in the humanities division were surveyed about their interests:

- 55% of students are interested in literature.
- 40% are interested in history.
- 20% are interested in both literature and history.

What is the probability that a randomly selected student is interested in literature or history?

Answer available at the end of this section.

Exercise 4.3.4

A political science professor surveys 200 students about their interests in public policy and international relations.

The results are shown in the table below:

Data Table for 200 Students

	International Relations	Not International Relations	Total
Public Policy	50	70	120
Not Public Policy	30	50	80
Total	80	120	200

What is the probability that a randomly selected student is interested in public policy or international relations?

Answer available at the end of this section.

Exercise 4.3.5

A standard deck of 52 playing cards is used. The cards are categorized by whether they are a heart and whether they are a face card (jack, queen, king).
The table below summarizes the counts:

Data Table for 52 Cards

	Face Card	Not a Face Card	Total
Heart	3	10	13
Not a Heart	9	30	39
Total	12	40	52

What is the probability that a randomly selected card is a heart or a face card?

Answer available at the end of this section.

Author

"4.3: Addition Rules for Probabilities"(opens in new window) by Alfie Swan(opens in new window) is licensed under CC BY 4.0(opens in new window)

Answers

Exercise 4.3.1

These events are mutually exclusive; thus, we use the addition rule I.

P(2 or 5) = $\frac{1}{6}+\frac{1}{6}=\frac{2}{6}=\frac{1}{3}$

Exercise 4.3.2

These events are mutually exclusive; thus, we use the addition rule I.

P (studying alone or in a group) = P (Studying alone) + P (Studying in a group) = 0.35 + 0.25 = 0.6 = 60%

Exercise 4.3.3

These events are mutually exclusive; thus, we use the addition rule II.

P (literature or history) = P (literature) + P (history) – P (literature and history) = 0.55 + 0.40 – 0.20 = 0.75 = 75%

Exercise 4.3.4

These events are mutually exclusive; thus, we use the addition rule II.

P (public policy or international relations) = P (public policy) + P (international relations) – P (public policy and international relations) = $\frac{120}{200}+\frac{80}{200}-\frac{50}{200}$

P (public policy or international relations) =$\frac{180}{200}=\frac{9}{10}$

Exercise 4.3.5

These events are mutually exclusive; thus, we use the addition rule II.

P (heart or face card) = P (heart) + P (face card) – P (heart and face card) = $\frac{13}{52} + \frac{12}{52} - \frac{3}{52}$

P (public policy or international relations) = $\frac{22}{52} = \frac{11}{26}$

This page titled 4.3: Addition Rules for Probabilities is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

4.4: Multiplication Rules and Conditional Probability

Learning Objectives

- Use multiplication rules to determine the probability of two events occurring together.
- Calculate the probability of independent events by multiplying their probabilities.
- Apply conditional probability for dependent events, adjusting calculations based on prior occurrences.

In probability, **independent events** are events where the occurrence of one does not affect the occurrence of the other. For example, flipping a coin and rolling a die are independent events because the outcome of the coin flip (heads or tails) does not influence the number that appears on the die. Moreover, when selecting a **small sample from a large population**, individual selections can often be treated as **independent events** because the removal of one item does not significantly alter the probabilities of selecting subsequent items. This is particularly true when the sample size is **much smaller** than the total population.

Multiplication Rule#1

Definition: Multiplication Rule#1

When two events A and B are independent, then $P(A \text{ and } B) = P(A) \cdot P(B)$

Example 4.4.1

A coin is tossed once, followed by the rolling of a single die. What is the probability of getting a head on the coin and a 5 on the die?

Sample space for the Problem ={ (Heads, 1), (Heads, 2), (Heads, 3), (Heads, 4), (Heads, 5), (Heads, 6), (Tails, 1), (Tails, 2), (Tails, 3), (Tails, 4), (Tails, 5), (Tails, 6)}

Solution

There are a total of 12 possible outcomes, and 1 of them is heads and 5. Thus, the probability is P(heads and 5) = $\frac{1}{12}$.

Using the formula above, we get the same answer. is P(heads and 5) = $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

Example 4.4.2

Assume that two dice are tossed twice. Compute the following probabilities. Write the answers as a fraction or a decimal rounded to three places.

a. Get a sum of 7 on both tosses.
b. Get a sum of 5 on the first toss and a sum of 8 on the second toss.

Solution

Since tossing two dice are independent events. The multiplication Rule#1 can be used to answer the question.

a. $P(\text{ First is a 7 and Second is a 7}) = \frac{6}{36} \cdot \frac{6}{36} = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0.028$

b. $P(\text{ First is a 5 and Second is an 8}) = \frac{4}{36} \cdot \frac{5}{36} = \frac{1}{9} \cdot \frac{5}{36} = \frac{5}{324} = 0.015$

Example 4.4.3

In the California Community College system, the chance of passing an introductory statistics course is around 48%. If three students are randomly selected from the system, what is the chance that all three will pass the class in one term? Write the answer as a decimal rounded to three places.

Solution

Since a small sample is selected from a large population, the events are independent, and Multiplication Rule#1 can be used to answer the question.

$P(\text{ All three pass the class}) = 0.48 \cdot 0.48 \cdot 0.48 = 0.48^3 = 0.111$

Example 4.4.4

A local elementary school is hiring new staff, and each applicant has a 72% chance of being hired. Four Citrus College students majoring in child development apply for positions at the school. Assuming each hiring decision is independent, what is the probability that all four students are hired?

Solution

$P(\text{ All four are hired at the elementary school}) = 0.72 \cdot 0.72 \cdot 0.72 \cdot 0.72 = 0.72^4 = 0.269$

Multiplication Rule#2

Dependent events are events where the outcome of one event affects the probability of the other occurring. This means that when one event happens, it changes the likelihood of the second event happening. For example, if you draw a card from a deck and do **not** put it back, the probability of drawing a second card of a certain type changes because the total number of cards in the deck has decreased.

Definition: Multiplication Rule#2

When two events A and B are dependent, the $P(A \text{ and } B) = P(A) \cdot P(B|A)$

where $P(B|A)$ is the probability that B occurs given that A has occurred.

Example 4.4.5

A person selects three cards from a standard deck without replacement. Compute the chance that all three cards are jacks. Write the answer as a decimal rounded to three non-zero decimal places.

Solution

Since the cards are selected without replacement, the events are dependent, and Multiplication Rule #2 is applied to solve the answer.

$P(\text{ All 3 Cards are Jacks}) = \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{1}{5525} = 0.000181$

Example 4.4.6

Three scholarships of $1000 are to be awarded to three liberal arts majors. The liberal arts majors who qualify for the scholarship consist of 7 economics majors, 4 sociology majors, 3 political science majors, and 6 anthropology majors. Compute the chance that all three scholarships will be awarded to the sociology majors. Write the answer as a decimal rounded to three non-zero decimal places.

Solution

Since a person can only win the award once, the events are dependent, and Multiplication Rule #2 is applied to solve the answer.

$$P(\text{ All 3 are sociology Majors}) = \frac{4}{20} \cdot \frac{3}{19} \cdot \frac{2}{18} = \frac{1}{285} = 0.00351$$

Conditional Probabilities

Conditional probability is the probability of an event occurring given that another event has already happened. It helps us understand how the likelihood of one event changes based on new information. An example of **dependent events** is drawing two jacks from a standard deck of 52 cards **without replacement**. When you draw the first card, the probability of getting a jack is **4/52** since there are 4 jacks in the deck. If the first card is a jack, only 3 jacks remain, and the total number of cards is now 51. This means the probability of drawing a second jack is **3/51**. Since the outcome of the first draw affects the probability of the second, these events are **dependent**.

Definition: Conditional Probability

The probability that event B will occur given that A has occurred can be computed as follows.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example 4.4.7 conditional probabilities

Suppose you select two cards from a standard deck without replacement. Find the following probabilities. Write the answer as a decimal rounded to three non-zero decimal places.

a. The second card is a queen, given that the first card is a jack.
b. The second card is a club, given that the first card is a heart.
c. The second card is a king, given that the first card is also a king.

Solutions

a. Since one jack has been removed from the deck, there is now a total of 51 cards and 4 queens. Thus the answer is $P(\text{ Second is a queen}|\text{ First is a jack }) = \frac{4}{51} = 0.078$

b. Since one heart has been removed from the deck, there is now a total of 51 cards and 13 clubs. Thus the answer is $P(\text{ Second is a club}|\text{ First is a heart }) = \frac{13}{51} = 0.255$

c. Since one king has been removed from the deck, there is now a total of 51 cards and 3 kings. Thus the answer is $P(\text{ Second is a king}|\text{ First is a king}) = \frac{3}{51} = 0.059$

Example 4.4.8

An urn contains 5 red marbles and 7 blue marbles, making a total of 12 marbles. Two marbles are drawn one after the other without replacement. What is the probability that the second marble drawn is red, given that the first marble drawn was red?

Solution

In this problem, the formula will be used to find the answer.

$$P(\text{First is Red}) = \frac{5}{12}$$

$$P(\text{First and Second is Red}) = \frac{5}{12} \cdot \frac{4}{11} = \frac{20}{132}$$

$$P(\text{Second is Red} \mid \text{First is Red}) = \frac{P(\text{First and Second are Red})}{P(\text{First is Red})} = \frac{\frac{20}{132}}{\frac{5}{12}} = \frac{4}{11} = 0.364$$

Example 4.4.9

The contingency table below presents data from a survey conducted among students at a local college. The survey respondents are categorized by gender and blood type. Write the answer as a fraction or decimal rounded to 3 nonzero places.

Contingency Table for Blood Type and Gender

	A	B	AB	O	Total
Male	61	81	30	96	268
Female	24	70	92	84	270
Total	85	151	122	180	538

Table 4.4.1: Contingency Table for Gender and Blood Type

If a student is selected, compute the following:

a. What is the probability that a student is a male given that the student has blood type B?
b. What is the probability that the student has blood type O, given that the student is a female?
c. What is the probability that the student is a female given that the student has blood type B or AB?

Solution

a. The condition is type B. Focus on the B column. $P(\text{Male}|\text{Blood type B}) = \frac{81}{151} = 0.536$
b. The condition is female. Focus on the female row. $P(\text{Blood type O}|\text{Female}) = \frac{84}{270} = 0.311$
c. The condition is blood type B or AB. Focus on the B or AB rows. $P(\text{Female}|\text{Blood types B or AB}) = \frac{162}{273} = 0.593$

Exercises

Exercise 4.4.1

Two fair six-sided dice are rolled.

- Event A: The first die shows a 4
- Event B: The second die shows a 6

Since the dice are rolled independently, the events are independent.
What is the probability that the first die shows a 4 and the second die shows a 6?

Answer available at the end of this section.

Exercise 4.4.2

An economist is studying consumer behavior.

- The probability that a randomly selected consumer prefers online shopping is 0.60
- The probability that a randomly selected consumer uses digital payment methods is 0.70

Assume these two events are independent.

What is the probability that a randomly selected consumer prefers online shopping and uses digital payment methods?

Answer available at the end of this section.

Exercise 4.4.3

An economist is analyzing income and investment behavior.

- The probability that a randomly selected individual is in the top income bracket is 0.20
- The probability that a randomly selected individual owns stocks is 0.50

Assume these two events are independent.

What is the probability that a randomly selected individual is in the top income bracket and owns stocks?

Answer available at the end of this section.

Exercise 4.4.4

A standard deck of 52 playing cards is used. Two cards are drawn without replacement.

- Event A: The first card drawn is a Heart
- Event B: The second card drawn is also a Heart

What is the probability that both cards drawn are Hearts?

Answer available at the end of this section.

Exercise 4.4.5

A university is awarding scholarships to students from different majors.

- 10 are child development majors
- 7 are economics majors
- 5 are sociology majors

Three students are selected at random without replacement.

a. What is the probability that the first student selected is a child development major, the second is an economics major, and the third is a sociology major?
b. What is the probability that all 3 are economics majors?

Answer available at the end of this section.

Answer available at the end of this section.

Exercise 4.4.6

A standard deck of 52 playing cards is used. One card is randomly selected.

- Event A: The card is a Face Card (Jack, Queen, or King)
- Event B: The card is a Heart

What is the probability that the card is a Face Card given that it is a Heart?

Answer available at the end of this section.

Exercise 4.4.7

A sociologist surveys 200 adults about community involvement and education level.
The results are shown in the table below:

	College Degree	No College Degree	Total
Volunteer	60	40	100
Does Not Volunteer	30	70	100
Total	90	110	200

What is the probability that a person volunteers, given that they have a college degree?

Answer available at the end of this section.

Authors

"4.4: Multiplication Rules and Conditional Probability" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"4.3: Conditional Probability" by Kathryn Kozak is licensed under CC BY-SA 4.0

Answers

Exercise 4.4.1

Since these events are independent, use the multiplication rule I.

P(first die is a 4 and second die is a 6) = $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

Exercise 4.4.2

Since these events are independent, use the multiplication rule I.

P(prefers online shopping and uses digital payment methods) = $0.6 \cdot 0.7 = 0.42$

Exercise 4.4.3

Since these events are independent, use the multiplication rule I.

P(top income bracket and owns stocks) = $0.2 \cdot 0.5 = 0.10$

Exercise 4.4.4

Since these events are dependent, use the multiplication rule II.

P(both cards drawn are Hearts) = $\frac{13}{52} \cdot \frac{12}{51} = \frac{1}{17}$

Exercise 4.4.5

Since these events are dependent, use the multiplication rule II.

a. P (child development major, economics major, and sociology major) = $\frac{10}{22} \cdot \frac{7}{21} \cdot \frac{5}{20} = \frac{5}{132}$

b. P (all 3 are economics majors) = $\frac{7}{22} \cdot \frac{6}{21} \cdot \frac{5}{20} = \frac{1}{44}$

Exercise 4.4.5

Exercise 4.4.6

The condition is that the first card is a heart, and there are 13 hearts. Of those 13 hearts, 3 of them are face cards.

Thus, the answer is P (face card | heart) = $\frac{3}{13}$

Exercise 4.4.7

Focus on the column for a college degree. Of those 90, 60 volunteer.

P (volunteer | college degree) = $\frac{60}{90} = \frac{2}{3}$

This page titled 4.4: Multiplication Rules and Conditional Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 4.3: Conditional Probability by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

4.5: At Least One Rule and Tree Diagrams

Learning Objectives

- Apply the "at least one" rule to calculate the probability that an event occurs one or more times, often by first finding the probability that it never occurs.
- Use tree diagrams as visual tools to map all possible outcomes step-by-step.
- Organize and compute complex probabilities effectively with these methods.
- Solve multi-step probability problems and reduce the risk of missing outcomes.

Assume there is a 25% chance that people will favor a new education policy proposed by a state government. If 5 people are randomly selected, what is the chance that at least one of the 5 will favor the policy? To directly compute this probability by looking at all possible variations of one or more factors favoring the policy would be cumbersome. It turns out that a quick way to calculate this type of probability is to compute the chance that none of them favors the policy and then subtract it from 1. This method is a general rule that can be applied to a wider range of applications. The rule is defined below.

Definition: 4.5.1

At Least One Rule for Probabilities

The probability of **"at least one"** refers to the likelihood that an event occurs at least once in multiple trials. It means that the event happens one or more times, as opposed to not happening at all.

P(at least one time) = 1 – P(none)

Where,

- P(none) is the probability that the event does **not** happen in any trial.
- By subtracting this from 1, we find the probability that the event **occurs at least once**.

In the example described at the beginning of this section, selecting a small sample from a large population will make the events independent. Thus, to compute the probability, the complement rule must be used to calculate the probability that one person does not favor the policy. This is P(not in favor) = 1 – 0.25 = 0.75. The chance that all five are not in favor is computed using the probability of the complement and the multiplication rule. P(All three are not in favor) = 0.75 · 0.75 · 0.75 · 0.75 · 0.75 = 0.237. Finally, to compute the probability that at least one favors the probability, use the formula. P(at least one favors the policy) = 1 – 0.237 = 0.763. Thus, the answer is 0.763.

Example 4.5.1

The chance that a person is a "night owl" is around 12%. Night owls are those people who believe they are more productive in the late hours of the night. If a sample of 3 people from the population is selected, compute the following probabilities.

a. All three are night owls.
b. None of the three is a night owl.
c. At least one of the three is a night owl.

Solutions

a. These events are independent. We can use the multiplication rule to compute the answer. P(all three are night owls) = 0.12 · 0.12 · 0.12 = 0.0017
b. The complement must be computed first. It is P(not a night owl) = 1 – 0.12 = 0.88. The answer is P(all three are not night owls) = 0.88 · 0.88 · 0.88 = 0.681

c. We use at least one rule and the result of part 2 to get the answer. P(At least one is a night owl) = 1 – 0.681 = 0.319.

Example 4.5.2

For a standard deck of cards, assume five cards are selected without replacement. Compute the probability of getting at least one spade.

Solution

Since the cards are selected without replacement, the events are dependent. That means for each selection, a card must be removed, and this must be reflected in the probability. To use at least one rule, compute the probability that five cards are not spades. There are 13 spades and 39 cards that are not spades. The answer will be rounded to a decimal with three place places.

P(none is a spade) = $\frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} \cdot \frac{36}{49} \cdot \frac{35}{48}$ = 0.222

The computation of the final answer is provided below.

P(at least one is a spade) = 1 – 0.222 = 0.778

Some probability problems may be extremely complex and have multiple outcomes. A tree diagram can be constructed for these problems. The tree diagram is used to help map out all the possibilities for the sequence of events in the problem. The events are represented by dots called nodes, and the outcomes of each event are represented by the branches or lines of the diagram. The first step is to construct the tree diagram. The second step is to assign probabilities to each branch in the tree diagram. The third step requires multiplying all probabilities along each pathway of the tree diagram that leads to the desired result. The final step is to add all the probabilities of the different pathways that resulted in the desired outcome.

Example 4.5.3

A coin is flipped two times. Compute the chance of getting exactly two tails.

Solution

1. First, construct the tree diagram.

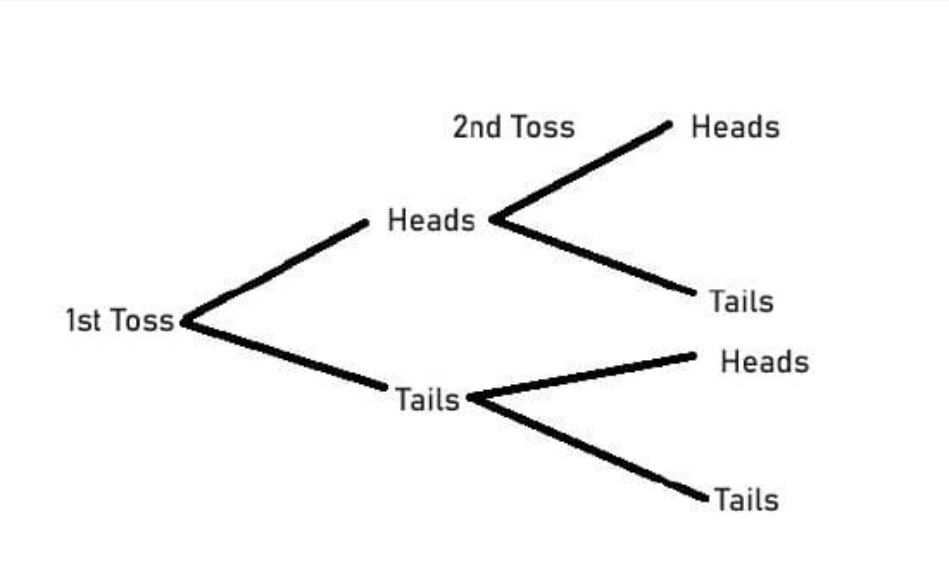


Figure 4.5.1 Tree Diagram

2. Second, assign probabilities to each branch. In this case, each branch will have a probability of 1/2 or 0.5 because there are only two possibilities per toss.

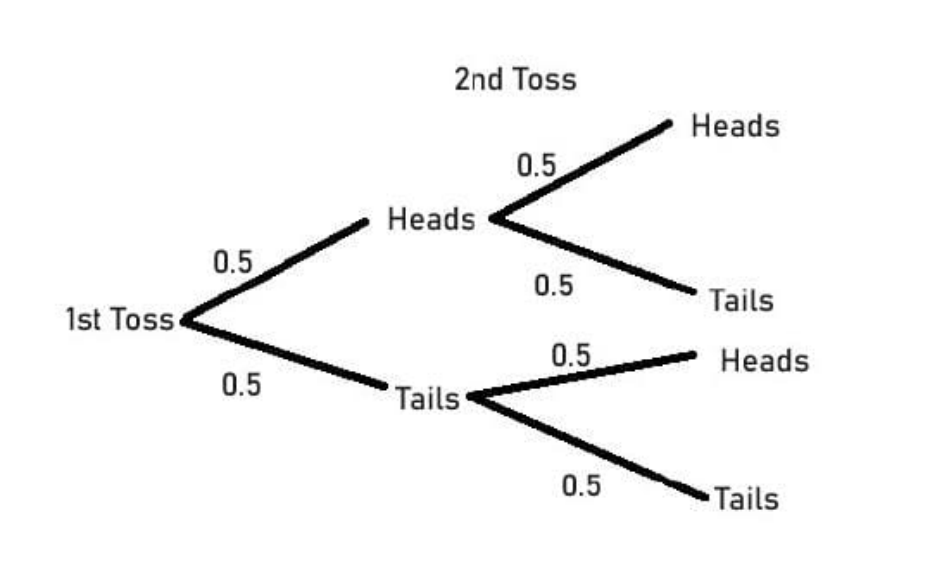


Figure 4.5.2 Tree Diagram with Probabilities

3. Third, follow each branch that leads to the desired outcome of two tails and multiply probabilities along the pathway.

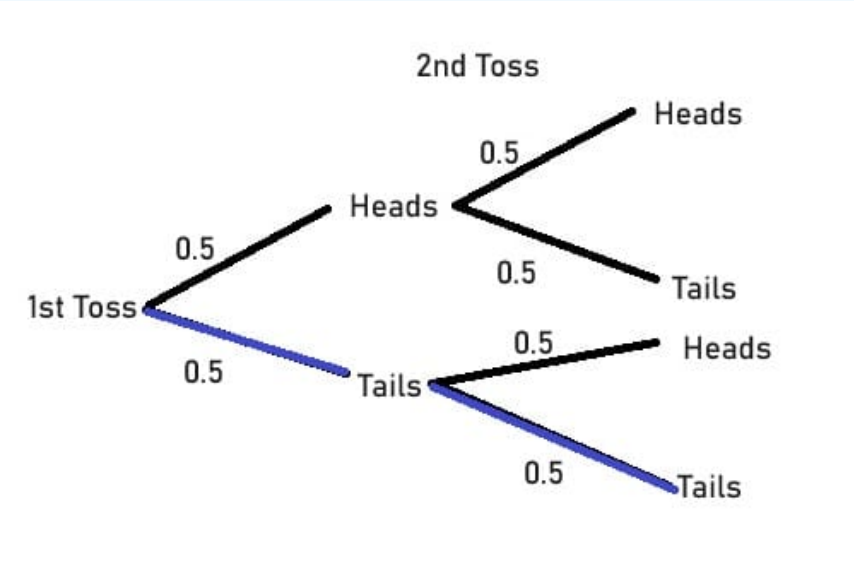


Figure 4.5.3 Tree Diagram with Desired Pathway

P(two tails) = 0.5 · 0.5 = 0.25

Exercises

Exercise 4.5.1

A literature professor randomly selects 2 books from a collection of 10 books:

- 4 are poetry books.
- 6 are novels.

The books are selected without replacement.
What is the probability that at least one of the selected books is a poetry book?

Answer available at the end of this section.

Exercise 4.5.2

A standard deck of 52 playing cards is used. Two cards are drawn without replacement.
What is the probability that at least one of the two cards drawn is a heart?

Answer available at the end of this section.

Exercise 4.5.3

A student takes 3 multiple-choice quiz questions. Each question has 4 answer choices, and the student randomly guesses on each question. Each question is independent. What is the probability that the student gets at least one question correct?

Answer available at the end of this section.

Exercise 4.5.4

A sociologist is studying college students and whether they participate in community service and student clubs. At a local college, 60% of students participate in community service. Of the students who participate in community service, 70% also belong to a student club. Of the students who do not participate in community service, 40% belong to a student club. Create a tree diagram and use it to answer the following questions:

a. What is the probability that a student participates in community service and belongs to a student club?
b. What is the probability that a student does not participate in community service and does not belong to a student club?
c. What is the probability that a student belongs to a student club?

Answer available at the end of this section.

Author

"4.5: At Least One Rule and Tree Diagrams" by Alfie Swan is licensed under CC BY 4.0

Answers

Exercise 4.5.1

Use the "at least one" formula to compute the probability. Compute the probability that none of the books is a poetry book.

P(none is a poetry book) = $\frac{6}{10} \cdot \frac{5}{9} = \frac{30}{90} = \frac{1}{3}$

P(at least one of the selected books is a poetry book) = 1 – P(none is a poetry book) = $1 - \frac{1}{3} = \frac{2}{3}$

Exercise 4.5.2

Use the "at least one" formula to compute the probability. Compute the probability that none of the books is a heart.

P(none is a heart) = $\frac{39}{52} \cdot \frac{38}{51} = \frac{19}{34}$

P(at least one of the cards is a heart) = 1 – P(none is a heart) = $1 - \frac{19}{34} = \frac{15}{34}$

Exercise 4.5.3

First, calculate the probability of guessing correctly. It is p(correct) = $\frac{1}{4}$

Second, calculate the probability of guessing incorrectly. It is p (incorrect) = $1 - \frac{1}{4} = \frac{3}{4}$

Third, calculate the probability that all three are guessed incorrectly. It is $P(\text{all three are incorrect}) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}$

Finally, calculate the probability that at least one problem is guessed correctly.

P(at least one guessed correctly) = 1 – P(none is correct) = $1 - \frac{27}{64} = \frac{37}{64}$

Exercise 4.5.4

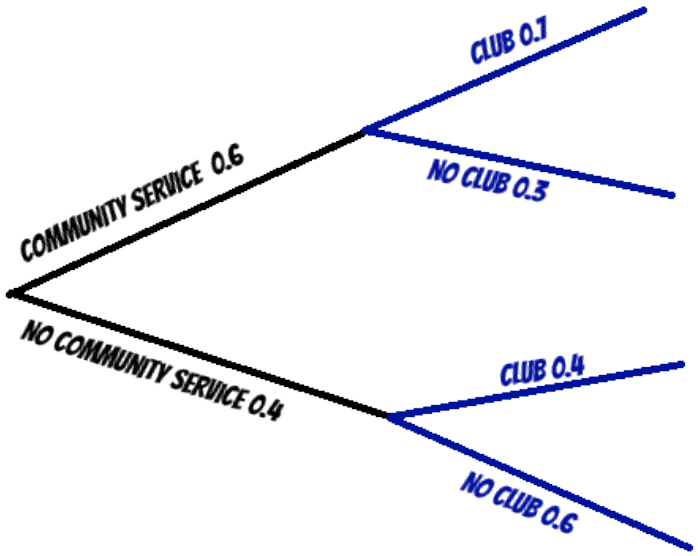


Figure 4.5.4: Tree Diagram of Student Involvement and Club Participation: This tree diagram models student involvement in community service and student clubs. It begins by splitting students into those who participate in community service (60%) and those who do not (40%). Each group then branches into students who belong to a club and those who do not, with corresponding probabilities shown at each stage. The diagram illustrates how to multiply along branches to find joint probabilities and combine outcomes to determine the overall probability that a student belongs to a club.

The answers are provided below using the tree diagram and multiplying adjacent probabilities.

a. P(community service and student club) = $0.6 \cdot 0.4 = 0.42$
b. P(community service and not student club) = $0.6 \cdot 0.3 = 0.18$
c. P(Student Club) = $0.6 \cdot 0.4 + 0.4 \cdot 0.4 = 0.42 + 0.16 = 0.58$

This page titled 4.5: At Least One Rule and Tree Diagrams is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

4.6: Counting Rules

Learning Objectives

- Apply counting rules to determine the number of possible ways events can occur.
- Use the fundamental counting rule by multiplying the number of choices for each step in a sequence.
- Calculate permutations to count arrangements where order matters.
- Calculate combinations to count selections where order does not matter.

Sometimes the sample space or event space is extremely large, and it isn't feasible to write it out. In that case, mathematical tools can be used for counting the size of the sample space and event space. These tools are known as counting techniques.

Definition 4.6.1

Multiplication Rule

Is used to compute the total number of possibilities for a sequence of events.

If event#1 can be done m_1 ways, event#2 can be done m_2 ways, and so forth to event#*n* being done m_n ways. Then the number of ways to do all the events together would be $m_1 \cdot m_2 \cdots m_n$.

Example 4.6.1 multiplication rule in counting

A menu offers a choice of 3 salads, 8 main dishes, and 5 desserts. How many different meals consisting of one salad, one main dish, and one dessert are possible?



Figure 4.6.1: Three-Course Meal: Salad, Entrée, and Dessert: The image shows three plated dishes arranged on a clean, light background. On the left is a fresh salad filled with leafy greens, cherry tomatoes, cucumbers, olives, croutons, and crumbled cheese. In the center foreground is a main dish featuring grilled chicken and a steak, served with roasted potatoes, carrots, and asparagus, with a lemon wedge on the side. On the right is a rich slice of layered chocolate cake topped with chocolate ganache, a strawberry, and raspberries, with sauce drizzled around the plate.

Solution

There are three tasks: picking a salad, a main dish, and a dessert. The salad task can be done 3 ways, the main dish task can be done 8 ways, and the dessert task can be done 5 ways. The ways to pick a salad, main dish, and dessert are

$$3 \cdot 8 \cdot 5 = 120$$

There are 120 different meals that are possible.

Example 4.6.2 Multiplication rule in counting

How many three-letter "words" can be made from the letters a, b, and c with no letters repeating? A "word" is just an ordered group of letters. It doesn't have to be a real word in a dictionary.

Solution

Three tasks must be done in this case. The tasks are to pick the first letter, then the second letter, and then the third letter. The first task can be done 3 ways since there are 3 letters. The second task can be done 2 ways since the first task took one of the letters. The third task can be done 1 way since the first and second tasks took two of the letters. There are

$$3 \cdot 2 \cdot 1 = 6$$

So, there are 6 different possibilities.

In Example 4.6.2, the solution was found by find $3 * 2 * 1 = 6$. Many counting problems involve multiplying a list of decreasing numbers. This is called a **factorial**. There is a special symbol for this and a special button on your calculator.

Definition 4.6.2

Factorial

$$n! = n(n-1)(n-2)\cdots(3)(2)(1)$$

As an example:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

$$8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40320$$

0 factorial is defined to be 0!=1 and 1 factorial is defined to be 1!=1.

Note: Calculator Instructions for Computing a Factorial

To use a TI-84 + calculator to compute a factorial perform the following steps for the example.

5!

1. Press the [**MATH**] button.
2. Use the right arrow key to scroll to the **[PRB]** menu.
3. Select option **[4: !]** by either scrolling down and pressing **ENTER** or simply pressing **4**.
4. You should now see **5!** on the screen.
5. Press **ENTER** to compute the factorial.

The answer for **5!** should be **120**.

Definition: 4.6.3

Permutation Rule #1:Permutation Rule for "n" Distinct Objects

The total number of permutations for n distinct objects that do not repeat is n!

Example 4.6.3

a. How many different batting orders are possible for a group of 9 batters?
b. A group of 6 students enters a line to purchase coffee at a local coffee shop. How many different lines are possible?
c. An athlete has 4 trophies on her shelf. How many ways can she arrange the trophies to display them?

Solutions

a. $9! = 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 362,880$
b. $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$
c. $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

If objects in the group are not distinct and repeat, a different formula must be used to compute the total number of permutations. This formula will divide out all the repetitions.

Definition:4.6.4

Permutation Rule #1: Permutation Rule for "n" Distinct Objects in Which Objects Repeat

The total number of permutations of n objects in which object 1 repeats r_1 times, object 2 repeats r_2 times, object 3 repeats r_3 times, etc. is computed as follows:

$$\frac{n!}{r_1! \cdot r_2! \cdots r_n!}$$

Example 4.6.4

a. For the word "STATISTICS," find the total number of permutations.
b. How many different ways can 5 identical dinosaur toys, 3 identical helicopter toys, and 4 identical action figures be arranged in a grocery store counter display?

Solutions

For this problem, there are 10 letters. Thus, n = 10. S repeats two times, T repeats three times, A occurs once, I repeats two times, and C occurs once. When dividing, only include the number of repetitions for letters with repetitions greater than one.

a. $\frac{10!}{2! \cdot 3! \cdot 2!} = \frac{10!}{2 \cdot 6 \cdot 2} = \frac{3,628,800}{24} = 151,200$
b. $\frac{12!}{5! \cdot 3! \cdot 4!} = \frac{12!}{120 \cdot 6 \cdot 24} = \frac{479,001,600}{17,280} = 27,720$

Sometimes a sample space depends on selecting *r* objects from *n* total objects. The number of ways to do this depends on whether the order in which the *r* objects matters or if it doesn't. For example, if a person needs to call another on the phone, they must have the correct number in the right order. Otherwise, the wrong number will be called. In this case, the order of the numbers matters. However, if numbers were selected randomly for playing the lottery, it doesn't matter which number is picked first. As long as the selection matches the numbers chosen by the lottery, the jackpot will be won. In this case, the order doesn't matter. A **permutation** is an arrangement of items in a specific order. Permutations are used to count items when the ordering of the items matters. When the order of the items doesn't matter, combinations are used. A **combination** is an arrangement of items when order is not important. When computing a counting problem, the first question should be "does order matter?"

Definition 4.6.5

Permutation Rule #3: Permutation Formula for Permutations of "r" Objects Chosen from "n" Objects

Picking *r* objects from *n* total objects when order matters

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

Example 4.6.5 calculating the number of ways

In a club with 15 members, how many ways can a slate of 3 officers consisting of a president, vice-president, and secretary/treasurer be chosen?

Solution

In this case, the order matters. If you pick person 1 for president, person 2 for vice-president, and person 3 for secretary/treasurer you would have different officers than if you picked person 2 for president, person 1 for vice-president, and person 3 for secretary/treasurer. This is a permutation problem with n=15 and r=3.

$$ {}_{15}P_3 = \frac{15!}{(15-3)!} = \frac{15!}{12!} = 2,730 $$

Note: Calculator Instructions for Computing the Permutation Formula

To use a TI-84 + calculator to compute a permutation for 15 choose 3 and perform the following steps.

${}_{15}P_3$

1. **Enter 15** (since we are choosing from 15 items).
2. Press the **[MATH]** button.
3. Use the **right arrow key** to navigate to the **[PRB]** menu.
4. Select **[2: nPr]** by either scrolling down and pressing **[ENTER]** or simply pressing **2**.
5. **Enter 3** (since we are selecting 3 items).
6. Press **[ENTER]** to compute the result.

The answer is 2,730.

Example 4.6.6

A school club has 25 members, and we need to select a President, Vice President, Treasurer, and Secretary from the group. How many ways can this be done?

Solution

Since the positions are different, the order in which they are selected is important. In this $n = 25$ and $r = 4$.

$$ {}_{25}P_4 = \frac{25!}{(25-4)!} = \frac{25!}{21!} = 303,600 $$

Example 4.6.7

In how many different ways can an ID card number be created using a 6-digit sequence?

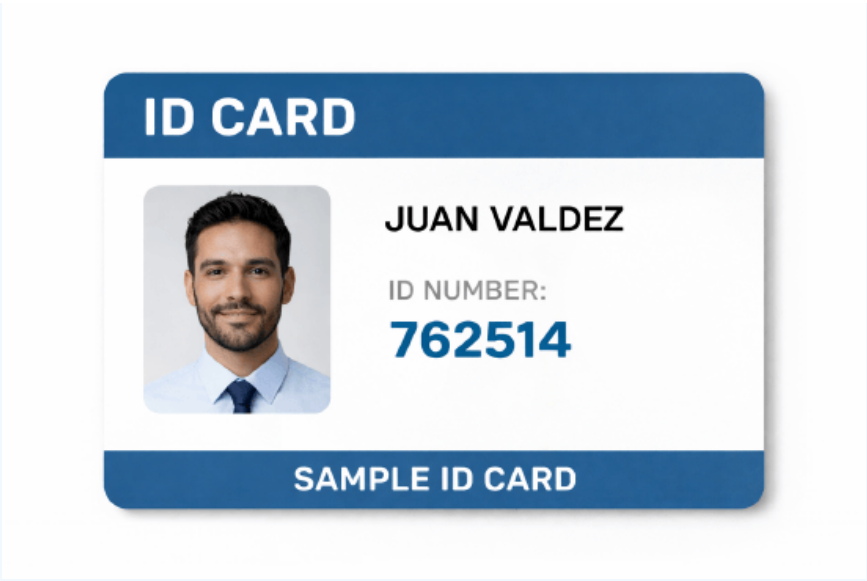


Figure 4.6.2: Sample ID Card for Juan Valdez: The image shows a professionally designed sample ID card with a blue and white color scheme. At the top, a bold blue header displays the text “ID CARD.” On the left side is a headshot of a man with short dark hair and a trimmed beard, wearing a dress shirt and tie. To the right of the photo, the name “JUAN VALDEZ” appears in large, bold text, with the label “ID NUMBER” beneath it followed by the six-digit number 762514 in blue. A blue footer at the bottom reads “SAMPLE ID CARD.”

Solution

Since order matters, each unique sequence of ID numbers represents a distinct identification number. Additionally, since there are 10 possible digits to choose from (0–9), we have $n = 10$, and since 6 digits are selected, we set $r = 6$.

$$ {}_{10}P_6 = \frac{10!}{(10-6)!} = \frac{10!}{4!} = 151,200 $$

Definition: \({4{.}6{.}\eatSpaces{6} \endSpaces}\

Combination Formula

Picking *r* objects from *n* total objects when order doesn’t matter

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

Example 4.6.8 calculating the number of ways

Suppose a college basketball team needs to select 7 players from a roster of 20 players to participate in a special training camp. How many ways can you do this?

Solution

In this scenario, the order of selection does not matter, as you are simply choosing 7 players from the group. Therefore, this is a combination problem, where $n = 20$ (total players) and $r = 7$ (players selected).

$$ {}_{20}C_7 = \frac{20!}{7!(20-7)!} = \frac{20!}{7!13!} = 77,520 $$

Note: Calculator Instructions for Computing the Combination Formula

To use a TI-84 + calculator to compute a combination for 20 choose 7 and perform the following steps.

$_{20}C_7$

1. **Enter 20** (since we are choosing from 20 items).
2. Press the **[MATH]** button.
3. Use the **right arrow key** to navigate to the **[PRB]** menu.
4. Select **[3: nCr]** by either scrolling down and pressing **[ENTER]** or simply pressing **3**.
5. **Enter 7** (since we are selecting 7 items).
6. Press **[ENTER]** to compute the result.

The answer is 77,520.

Example 4.6.9

How many ways can 5 cards be selected from a deck of 52 cards?



Figure 4.6.3: Five Cards Drawn from a Standard Deck: The image displays five playing cards drawn from a standard deck and arranged in a fanned layout on a green felt surface. The cards shown are an ace, king, queen, jack, and ten, each from different suits, forming a sequential hand. The clean arrangement and contrasting colors make it easy to distinguish each card, providing a clear visual representation of a five-card draw commonly used in probability and statistics examples.

Solution

In this scenario, the order of selection does not matter, as you are simply choosing 5 cards from the group of 52 cards. Therefore, this is a combination problem, where $n = 52$ (total cards) and $r = 5$ (cards selected).

$$_{52}C_5 = \frac{52!}{5!(52-5)!} = \frac{52!}{5!47!} = 2,598,960$$

Most calculators have a factorial button, and many have combination and permutation functions.

Exercises

Exercise 4.6.1

A sociology department is forming a research panel consisting of 1 sociologist, 1 psychologist, and 1 economist. There are 5 sociologists, 4 psychologists, and 6 economists available. How many different panels can be formed?

Answer available at the end of this section.

Exercise 4.6.2

A community study surveys individuals based on three characteristics: education level (high school, associate degree, bachelor's degree, graduate degree), employment status (employed, unemployed), and housing status (rent, own). How many different possible combinations of responses are there?

Answer available at the end of this section.

Exercise 4.6.3

A political science researcher is designing a focus group by selecting 1 participant from each of the following groups: 10 urban residents, 12 suburban residents, and 8 rural residents. How many different focus groups can be created?

Answer available at the end of this section.

Exercise 4.6.4

Compute the following factorials.

a. $10!$
b. $\frac{8!}{4!}$

Answer available at the end of this section.

Exercise 4.6.5

A college baseball coach must decide the batting order for 9 players in the starting lineup. Since the order of the players matters, in how many different ways can the 9 players be arranged in the batting lineup?

Answer available at the end of this section.

Exercise 4.6.6

A speech class is selecting 6 students to present their speeches, and the order of presentations matters. In how many different ways can the 6 students be arranged for presentation?

Answer available at the end of this section.

Exercise 4.6.7

A student has 7 different books and wants to arrange them on a bookshelf. Since the order of the books matters, in how many different ways can the 7 books be arranged?

Answer available at the end of this section.

Exercise 4.6.8

A linguistics student is analyzing the word “SUCCESS” for a class project. How many different, distinct arrangements of the letters in the word can be formed?

Answer available at the end of this section.

Exercise 4.6.9

A computer science student is creating a code using the symbols ####, &&&, and ** (that is, 4 hashtags, 3 ampersands, and 2 asterisks). If all symbols are arranged in a sequence, how many different unique code words can be formed?

Answer available at the end of this section.

Exercise 4.6.10

An art history class has 40 famous paintings to feature in a gallery spotlight. The curator will choose 4 paintings and decide the order in which they are displayed (first and second). How many different orders of 4 paintings can be selected?

Answer available at the end of this section.

Exercise 4.6.11

A literature club is selecting 3 different novels from a list of 10 to discuss over two weeks, where the order of discussion matters (which book is read first and which is read second). How many possible ordered selections are there?

Answer available at the end of this section.

Exercise 4.6.12

A communications department is selecting 4 students from 28 to deliver opening and closing speeches at an event. Since the roles are different, order matters, how many different ways can the students be assigned to these four roles?

Answer available at the end of this section.

Exercise 4.6.13

A music appreciation class has 8 songs, and the instructor will select 3 of them to analyze in class. Since the order does not matter, how many different groups of 3 can be selected?

Answer available at the end of this section.

Exercise 4.6.14

A film studies club is choosing 5 movies from a list of 30 to screen during the semester. The order in which they are shown is not important. How many different combinations of 5 movies can be selected?

Answer available at the end of this section.

Exercise 4.6.15

A psychology study is forming a pair of participants by selecting 3 people out of 16 volunteers. Since both participants have equal roles, how many different groups of 3 can be formed?

Answer available at the end of this section.

Exercise 4.6.16

An economics conference is selecting a panel of 4 speakers from two groups: 6 graduate students and 5 faculty members. The panel must include exactly 2 graduate students and 2 faculty members. How many different panels can be formed?

Answer available at the end of this section.

Exercise 4.6.17

A cultural studies class is forming a project team of 6 students from a group that includes 10 history majors, 6 psychology majors, and 9 communication majors. The team must include exactly 2 students from each major. How many different teams can be formed?

Answer available at the end of this section.

Authors

"4.6: Counting Rules" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"4.4: Counting Techniques" by Kathryn Kozak is licensed under CC BY-SA 4.0

Answers

Exercise 4.6.1

Use the counting rule to answer the question.

$5 \cdot 4 \cdot 6 = 120$

Exercise 4.6.2

Use the counting rule to answer the question.

$4 \cdot 2 \cdot 2 = 16$

Exercise 4.6.3

Use the counting rule to answer the question.

$10 \cdot 12 \cdot 8 = 960$

Exercise 4.6.4

a. 10! = 3,628,800

b. $\frac{8!}{4!} = \frac{40,320}{24} = 1,680$

Exercise 4.6.5

Use permutation rule #1.

9! = 362,880

Exercise 4.6.6

Use permutation rule #1.

6! = 720

Exercise 4.6.7

Use permutation rule #1.

7! = 5,040

Exercise 4.6.8

Use permutation rule #2.

$$\frac{7!}{3! \cdot 2! \cdot 1! \cdot 1!} = \frac{5040}{12} = 420$$

Exercise 4.6.9

Use permutation rule #2.

$$\frac{9!}{4! \cdot 3! \cdot 2!} = \frac{3,628,880}{288} = 1,260$$

Exercise 4.6.10

Use permutation rule #3.

$${}_{40}P_4 = 2,193,360$$

Exercise 4.6.11

Use permutation rule #3.

$${}_{10}P_3 = 720$$

Exercise 4.6.12

Use permutation rule #3.

$${}_{28}P_4 = 491,400$$

Exercise 4.6.13

Use the combination rule.

$${}_8C_3 = 56$$

Exercise 4.6.14

Use the combination rule.

$${}_{30}C_5 = 142,506$$

Exercise 4.6.15

Use the combination rule.

$${}_{16}C_3 = 560$$

Exercise 4.6.16

Use the combination rule.

$${}_6C_2 \cdot {}_5C_2 = 15 \cdot 10 = 150$$

Exercise 4.6.17

Use the combination rule.

$${}_{10}C_2 \cdot {}_6C_2 \cdot {}_9C_2 = 45 \cdot 15 \cdot 36 = 24,300$$

This page titled 4.6: Counting Rules is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 4.4: Counting Techniques by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

4.7: Chapter 4 Formulas

Chapter 4 Formulas

General Formula for the Probability of an Event

The probability that an event E will occur is the ratio of the number of outcomes in the event to the total number of outcomes in the sample space. It gives a numerical measure of the likelihood that the event will happen.

$$P(E) = \frac{\text{number of outcomes in the event}}{\text{total number of outcomes}}$$

Addition Rule #I for Probability for Mutually Exclusive Events

The Addition Rule I in probability theory is used to find the probability that one of two events occurs when the events are mutually exclusive. Mutually exclusive events cannot occur at the same time. This means that if events A and B cannot happen at the same time, the probability of either A or B happening is the sum of their probabilities.

$$P(A \text{ or } B) = P(A) + P(B)$$

Addition Rule #II for Probability for Non-Mutually Exclusive Events

The Addition Rule II in probability theory is used to find the probability that one of two events occurs, even if the events are not mutually exclusive. When events can occur at the same time, we must subtract the probability of both events occurring together to avoid double-counting.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

At Least One Probability

The "at least one" rule in probability is used to calculate the probability that an event occurs at least once in multiple trials. Rather than computing all possible ways the event can happen, it is often simpler to use the complement rule: subtract the probability that the event does not occur at all from 1.

$$P(\text{at least one}) = 1 - P(\text{none})$$

Complement Rule for Probability

The complement rule in probability is used to find the probability that an event does not occur. It is based on the fact that the total probability of all possible outcomes is 1. Therefore, the probability that event A does not happen is equal to 1 minus the probability that it does happen.

$$P(\text{Not } A) = 1 - P(A)$$

Conditional Probability

Conditional probability is the probability that one event occurs given that another event has already occurred. It helps us understand how the occurrence of one event affects the likelihood of another. This means the probability of B given A is equal to the probability of both A and B occurring divided by the probability of A.

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Multiplication Rule I for Probability for Independent Events

The Multiplication Rule I in probability theory is used to find the probability that two events occur in sequence, assuming the events are independent. Independent events are those where the outcome of one event does not affect the outcome of the other.

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Multiplication Rule II for Probability for Dependent Events

The Multiplication Rule II in probability theory is used to find the probability that two events occur in sequence when the events are dependent. Dependent events are those where the outcome of the first event affects the outcome of the second event. This means the probability of both A and B occurring is equal to the probability of A times the probability of B, given that A has occurred.

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

Combination Rule

The combination rule is used to determine the number of ways to choose r items from a group of n items when the order of selection does not matter. This is useful in situations where the arrangement of selected items is irrelevant.

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Fundamental Counting Rule

The fundamental counting rule is used to determine the total number of possible outcomes when a sequence of events occurs and each event has a specific number of possible outcomes. If the first event can occur in m_1 ways, the second in m_2 ways, and so on, then the total number of combined outcomes is the product of all these values.

$$\text{Total outcomes} = m_1 \cdot m_2 \cdot m_3 \cdots m_k$$

n-Factorial

The term "n factorial" (written as $n!$) represents the product of all positive integers from 1 to n. It is used frequently in counting and probability problems, especially in permutations and combinations.

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1$$

Permutation of n Distinct Objects

The permutation of n items is an arrangement of those items in a specific order. If all n items are used, the total number of different permutations is computed using the formula below.

$$n!$$

Permutation with Repetition of Identical Items

The permutation rule with objects that repeat is used when some objects are identical (i.e., indistinguishable from each other); the total number of distinct permutations is reduced. The formula accounts for repeated items by dividing by the factorial of each group of repeated objects, where there are n total items, with r_1 indistinguishable of one kind, r_2 of another, and so on is provided below.

$$\frac{n!}{r_1! \cdot r_2! \cdot \cdots \cdot r_k!}$$

Permutation Rule for Selecting r Objects from n Distinct Items

The permutation rule for n choose r gives the number of ways to arrange r items from a set of n distinct items, where the order matters and no repetitions are allowed. The number of such permutations is computed using the formula below.

$$_nP_r = \frac{n!}{(n-r)!}$$

Authors

"4.7: Chapter 4 Formulas" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 4.7: Chapter 4 Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

4.8: Chapter 4 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Combination: A selection of items where the order does not matter.

Combination Rule: The number of ways to choose r items from n distinct items without regard to order.

Complementary events: Complementary events are two outcomes of a probability experiment that are the only two possible outcomes and together make up the entire sample space.

Event: a set of certain outcomes of an experiment that you want to have happen.

Event Space: the set of outcomes that make up an event. The symbol is usually a capital letter.

Experiment: an activity or process that will produce specific outcomes.

Factorial: The product of all positive integers up to a number n.

Fundamental Counting Rule (Multiplication Rule): If one event can occur in m ways and another in n ways, then both events can occur in $m \times n$ ways.

Independent events are events where the occurrence of one does not affect the occurrence of the other.

Law of large numbers: a rule in which**,** as "n" increases, the relative frequency tends towards the actual probability value.

Mutually exclusive events (disjoint) are events with no outcomes in common.

Outcomes: the result of an experiment.

Permutation: An arrangement of items where the order matters.

Permutation Rule (no repetition): The number of ways to arrange r items from n distinct items.

Permutation with Repetition: Used when some items are repeated.

Probability: the measure of how likely an event is to occur.

Sample Space: a collection of all possible outcomes of the experiment. It is denoted as SS.

Key Terms

${}_nC_r$ - number of combinations (n choose r).

E, X, A, B - events (subsets of the sample space).

$\bar{E}$ - complement of event E.

! - factorial symbol (e.g., $n!$).

n - total number of items.

$n!$ - factorial of n.

${}_nP_r$ - number of permutations (arrangements of r items from n).

$P(E)$ - probability that event E occurs.

$P(\bar{E})$ - probability that event E does not occur.

$P(B \mid A)$ - conditional probability of B given A.

r - number of items selected.

\(SS \) - sample space (the set of all possible outcomes).

Authors

"4.8: Chapter 4 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 4.8: Chapter 4 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

5: DISCRETE PROBABILITY DISTRIBUTIONS

5.1: Basics of Probability Distributions
5.2: Binomial Probability Distribution
5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution
5.4: Poisson Distribution
5.5: Chapter 5 Formulas
5.6: Chapter 5 - Key Terms and Symbols

This page titled 5: Discrete Probability Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

5.1: Basics of Probability Distributions

Learning Objectives

- Define a discrete probability distribution as a list of all possible outcomes of a random variable, along with their associated probabilities.
- Use discrete probability distributions to model real-world scenarios such as dice rolls or item counts.
- Calculate key measures from the distribution, including mean, variance, standard deviation, and expected value.

As a reminder, a variable, or what will be called the random variable from now on, is represented by the letter *x, which* means a quantitative (numerical) variable that is measured or observed in an experiment.

Also, remember that there are different types of quantitative variables, called discrete or continuous. What is the difference between discrete and continuous data? **Discrete** data can only take on particular values in a range, **while continuous** data can take on any value in a range. Discrete data usually arises from counting, while continuous data usually arises from measuring.

Probability Distribution

How tall is a plant given a new fertilizer? Continuous. This is something you measure. How many fleas are on prairie dogs in a colony? Discrete. The number of fleas can be counted.

A **random variable** has outcomes that are determined by chance, and it is what is being measured. A **discrete random variable** is a random variable whose outcomes are countable whole-number values. For example, in the problem of "how tall is a plant given a new fertilizer?", the random variable is the height of the plant given a new fertilizer. For "how many fleas are on prairie dogs in a colony," the discrete random variable is the number of fleas on a prairie dog in a colony.

Suppose all the outcomes of a discrete random variable are listed together with their corresponding probabilities. A distribution is formed, but now it is called a probability distribution since it involves probabilities. A **probability distribution** is an assignment of probabilities to the values of the random variable. Furthermore, the sum of all the probabilities in the distribution must equal 1.

Note: The abbreviation of pdf is used for a probability distribution function.

For probability distributions, $0 \leq P(x) \leq 1 \text{ and } \sum P(x) = 1$

Example 5.1.1: Probability Distribution

The 2010 U.S. Census found the chance of a household being a certain size. Is the variable X a discrete random variable? Also, determine whether the distribution represents a probability distribution. If it does not, explain why. The data is in Example 5.1.1 ("Households by age," 2013).

Table 5.1.1: Household Size from US Census of 2010

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

Solution

In this case, the random variable is x = the number of people in a household. Since you are counting the number of people in a household, this is a discrete random variable.

This is a probability distribution since you have the x value and the probabilities that go with it. All of the probabilities are between zero and one, and the sum of all is one.

You can give a probability distribution in table form (as in the Example above) or as a graph. The graph looks like a histogram. A probability distribution is a relative frequency distribution based on a very large sample.

Example 5.1.2 graphing a probability distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013). Draw a histogram of the probability distribution.

Table 5.1.2: Household Size from US Census of 2010

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

Solution

State random variable:

x = number of people in a household

You draw a histogram, with the x values on the horizontal axis representing the classes (for the 7 or more categories, just call it 7) and the probabilities on the vertical axis representing the probabilities.

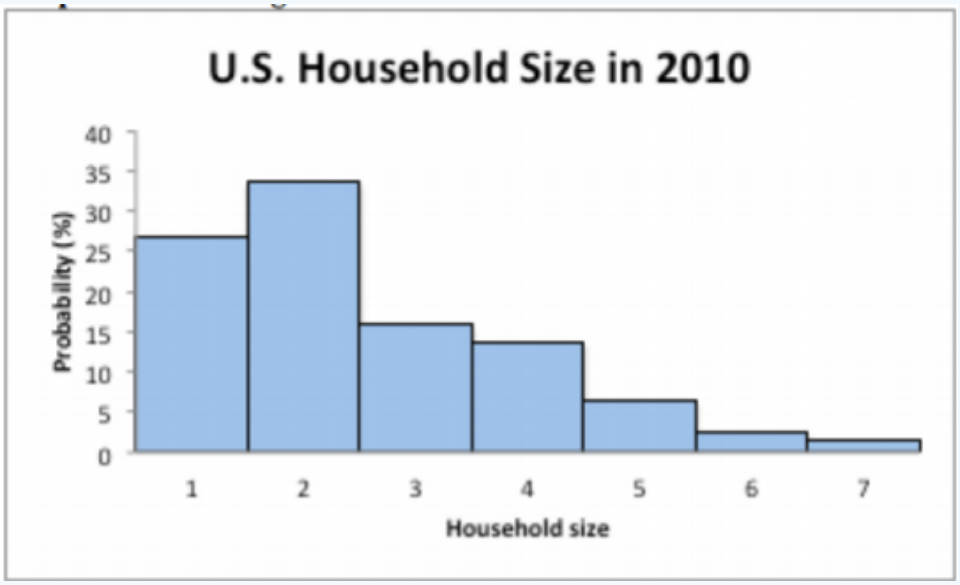


Figure 5.1.1: Histogram of Household Size from US Census of 2010

Notice this graph is skewed right.

Just as with any data set, the mean and standard deviation can be computed. In problems involving a probability distribution function (PDF), the probability distribution is considered the population, even though the PDF, in most cases, comes from repeating an experiment many times. This is because you are using the data from repeated experiments to estimate the true probability. Since a PDF is a population, the mean and standard deviation that are calculated are the population parameters and not the sample statistics. The notation used is the same as the notation for population mean and population standard deviation that was used in Chapter 3.

Expected Value, Variance, and Standard Deviation

Note

The mean can be thought of as the **expected value**. It is the value you expect to get if the trials are repeated an infinite number of times. The mean or expected value does not need to be a whole number, even if the possible values of *x* are whole numbers.

For a discrete probability distribution function,

The mean or expected value is $\mu = \sum xP(x)$

The variance is $\sigma^2 = [\Sigma x \cdot p(x)] - \mu^2$

The standard deviation is $\sigma = \sqrt{[\Sigma x \cdot p(x)] - \mu^2}$

where *x* = the value of the random variable and *P(x)* = the probability corresponding to a particular *x* value.

Example 5.1.3: Calculating mean, variance, and standard deviation for a discrete probability distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013).

Table 5.1.3: Household Size from US Census of 2010

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

a. Find the mean
b. Find the variance
c. Find the standard deviation
d. Use a TI-83/84 to calculate the mean and standard deviation

Solution

State random variable:

x= number of people in a household

a. To find the mean, it is easier just to use a table as shown below. Consider the category 7 or more to be just 7. The formula for the mean says to multiply the *x* value by the *P(x)* value, so add a row to the table for this calculation. Also, convert all *P(x)* to decimal form.

Probability Distribution for Calculating the Mean

X	$P(X)$	$X \cdot P(X)$
1	0.267	0.267
2	0.336	0.672
3	0.158	0.474
4	0.137	0.548
5	0.063	0.315
6	0.024	0.144
7	0.015	0.098
No sum needed	$\Sigma = 1$	$\Sigma X \cdot P(X) = 2.525$

Table 5.1.4: Calculating the Mean for the Discrete Probability Distribution

This is the mean or the expected value is the sum of the last column, which is μ = 2.525 people. This means that you expect a household in the U.S. to have 2.525 people in it. Now, of course, you can't have half a person, but what this tells you is that you expect a household to have either 2 or 3 people, with a little more 3-person households than 2-person households.

b. To find the variance, again, it is easier to use a table version than to try to use just the formula in a line. Looking at the formula, you will notice that the first operation that you should do is to subtract the mean from each *x* value. Then, you square each of these values. Then, you multiply each of these answers by the probability of each *x* value. Finally, you add up all of these values.

Probability Distribution for Calculating the Variance and Standard Deviation

X	$P(X)$	$X^2 \cdot P(X)$
1	0.267	0.267
2	0.336	1.344
3	0.158	1.422
4	0.137	2.192
5	0.063	1.575
6	0.024	0.864
7	0.015	0.735
No sum needed	$\Sigma = 1$	$\Sigma X^2 \cdot P(X) = 8.399$

Table 5.1.5: Calculating the Variance for the Discrete Probability Distribution

To find the variance, add the values in the last column and subtract the square of the mean from the sum. It is \(\sigma^{2}=8.399 - 2.525^{2} = 8.399- 6.376 = 2.023. (Note: Try not to round your numbers too much so you aren't creating rounding errors in your answer. The numbers in the table above were rounded off because of space limitations, but the answer was calculated using many decimal places.)

c. To find the standard deviation, just take the square root of the variance, $\sigma = \sqrt{2.023375} \approx 1.422454$ people. This means that you can expect a U.S. household to have 2.525 people in it, with a standard deviation of 1.42 people.

d. Go into the **[STAT]** menu, then the **[Edit]** menu. Type the *x* values into **[L1]** and the *P(x)* values into **[L2]**. Then go into the **[STAT]** menu, then the **[CALC]** menu. Choose **[1:1-Var Stats]**. This will put 1-Var Stats on the home screen. Now type in **[L1]**, **[L2]** (there is a comma between **[L1]** and **[L2]**) and then press **[ENTER]**. If you have the newer operating system on the TI-84, then your input will be slightly different. You will see the output in the figure below.

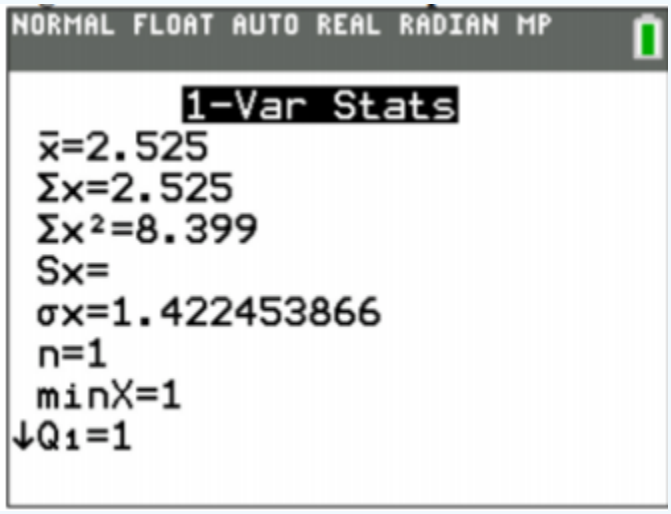


Figure 5.1.2: TI-83/84 Output

The mean is 2.525 people, and the standard deviation is 1.422 people.

To compute the variance from the calculator output. Type the standard deviation, including all digits, into the calculator. Then select the square function and enter. Round the answer to three decimals. Thus, the variance is 2.023.

Exercises

Exercise 5.1.1

A political analyst studies how many major policy reforms a legislative body passes in a single session. Based on historical data, the probability distribution is provided below. Perform the following steps.

a. Draw a probability distribution graph (bar chart) with:

- X-axis: Number of policies supported
- Y-axis: Probability

b. Describe the shape of the distribution (e.g., symmetric, skewed left, skewed right).
c. Identify the most likely outcome and explain what it means in context.
d. Based on the graph, are voters more likely to support fewer policies (0–1) or more policies (3–4)? Justify your answer using probabilities.
e. What does this distribution suggest about overall voter behavior in this election?

Probability Distribution for Number of Reforms

Number of Reforms (X)	0	1	2	3	4
P(X)	0.10	0.25	0.30	0.20	0.15

Answer available at the end of this section.

Exercise 5.1.2

A college is studying the number of philosophy majors in randomly selected student discussion groups, where X represents the number of philosophy majors in a group. The probability that there are 0 philosophy majors is 0.15, the probability that there is 1 philosophy major is 0.30, the probability that there are 2 philosophy majors is 0.25, the probability that there are 3 philosophy majors is 0.20, and the probability that there are 4 philosophy majors is 0.10. Using this information, answer the following questions.

a. Construct a discrete probability distribution table.
b. Interpret the meaning of P(2).
c. Draw a bar graph of the probability distribution
d. Determine which outcome is most likely based on your graph.

Answer available at the end of this section.

Exercise 5.1.3

A researcher is studying how many new public policies a city council adopts in a year. The probability distribution is given below, and then answer the questions below the table.

Probability Distribution for Number of Reforms

Number of Reforms (X)	0	1	2	3	4
P(X)	0.10	0.25	0.30	0.20	0.15

a. Find the mean of the distribution.
b. Find the variance of the distribution.
c. Find the standard deviation.
d. Interpret the meaning of the mean and standard deviation in context.

Answer available at the end of this section.

Exercise 5.1.4

A music artist schedules a varying number of live shows each month. Based on past data, the probability distribution for the number of shows is given below, and answer the questions below it.

Probability Distribution for Number of Shows

Number of Shows (X)	1	2	3	4	5
P(X)	0.15	0.25	0.30	0.20	0.10

a. Find the mean.
b. Find the variance.
c. Find the standard deviation.
d. Interpret the results in the context of the artist's performance schedule.

Answer available at the end of this section.

Authors

"5.1: Basics of Probability Distributions"(opens in new window) by Toros Berberyan(opens in new window), Tracy Nguyen(opens in

new window), and Alfie Swan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

Attributions

"5.1: Basics of Probability Distributions"(opens in new window) by Kathryn Kozak(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

Answers

Exercise 5.1.1

a. See the graph below

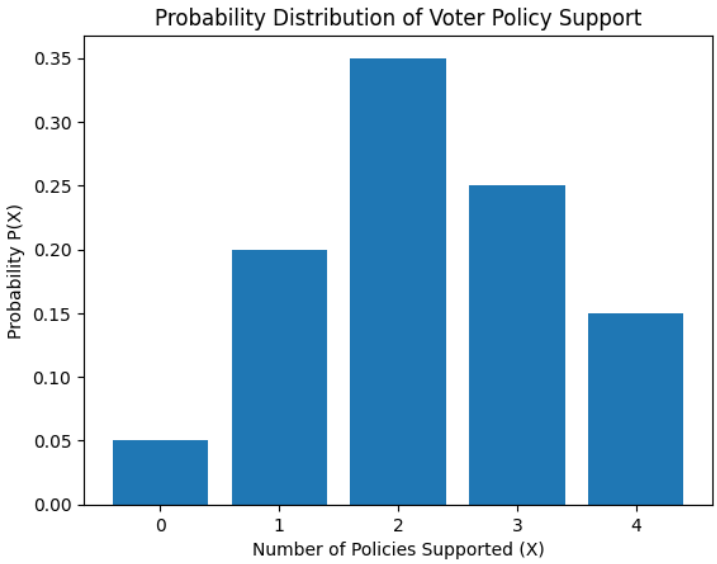


Figure 5.1.3: The image shows a bar graph representing the probability distribution of the number of reforms. The horizontal axis is labeled “Number of Reforms (X)” and displays values from 0 to 4, while the vertical axis is labeled “Probability P(X).” Each bar corresponds to a specific number of reforms and its associated probability: 0 reforms (0.10), 1 reform (0.25), 2 reforms (0.30), 3 reforms (0.20), and 4 reforms (0.15). The bars are evenly spaced and labeled with their probability values, making it easy to compare the likelihood of each outcome.

b. The distribution is slightly skewed to the right.
c. The most likely outcome is: X = 2 with probability 0.35. Voters are most likely to support 2 policies in this election.
d. Fewer policies (0–1):P(0)+P(1)=0.05+0.20=0.25 and more policies (3–4): P(3)+P(4)=0.25+0.15=0.40. Voters are more likely to support more policies (3–4) since 0.4 > 0.25.
e. This distribution suggests: Voters tend to support a moderate number of policies (around 2–3), there is very low support for 0 policies, and it is rare. Overall, there is a slight tendency toward supporting multiple policies rather than very few.

Exercise 5.1.2

a. The table is provided below.

Probability Distribution of Number of Philosophy Majors

Number of Philosophy Majors (X)	0	1	2	3	4
P(X)	0.15	0.30	0.25	0.20	0.10

b. There is a 25% chance that a randomly selected discussion group will have exactly 2 philosophy majors.
c. The graph is provided below.

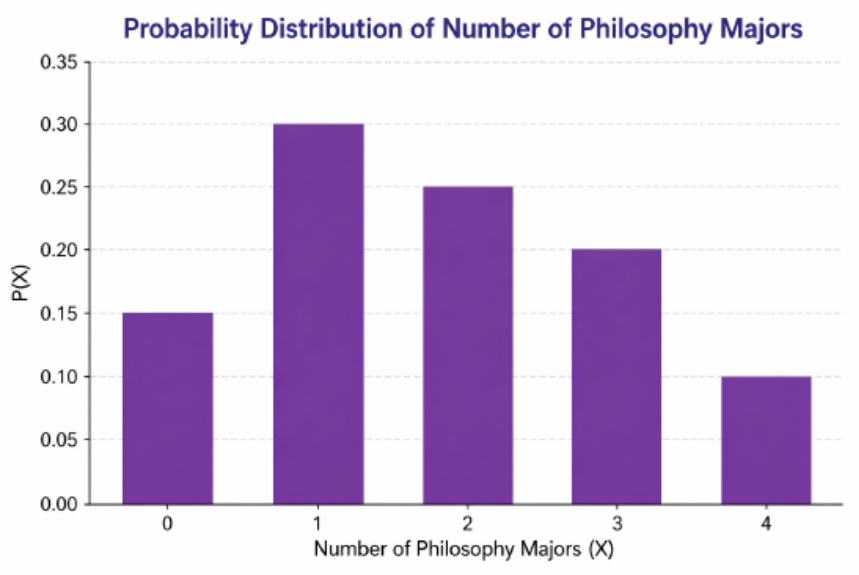


Figure 5.1.4: The image shows a bar graph representing the probability distribution of the number of philosophy majors in a group. The horizontal axis is labeled "Number of Philosophy Majors (X)" and includes values from 0 to 4, while the vertical axis is labeled "P(X)." Each bar is colored purple and represents the probability associated with each value of X. The tallest bar occurs at 2 philosophy majors, indicating it is the most likely outcome, while shorter bars appear at 0 and 4, showing lower probabilities. The graph provides a clear visual comparison of how likely each number of philosophy majors is within a group.

d. The most likely outcome is 1 philosophy major per group.

Exercise 5.1.3

a. $\mu = 2.15$
b. $\sigma^2 = 1.43$
c. $\sigma = 1.19$
d. The mean suggests that, on average, about 2 policies are adopted per year. The standard deviation indicates a moderate spread, meaning the number of policies typically varies by about 1 policy from the mean.

Exercise 5.1.4

a. $\mu = 2.85$
b. $\sigma^2 = 1.43$
c. $\sigma = 1.19$
d. The artist performs about 2.85 shows per month on average (roughly 3 shows). The standard deviation of 1.19 indicates that the number of shows typically varies by about 1 show from the average. This suggests a fairly consistent performance schedule, with occasional busier or lighter months.

This page titled 5.1: Basics of Probability Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 5.1: Basics of Probability Distributions by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

5.2: Binomial Probability Distribution

Learning Objectives

- Define the binomial probability distribution to model the number of successes in a fixed number of repeated, independent trials with two possible outcomes (success or failure).
- Calculate the probability of achieving a specific number of successes.
- Use the total number of trials, the probability of success, and the desired number of successes in the computation.

Section 5.1 introduced the concept of a probability distribution. The focus of the section was on discrete **probability distributions** (PDF). The PDF is found by conducting an experiment to collect data and then computing the experimental probabilities. Normally, the theoretical probabilities cannot be computed. However, certain types of experiments allow for the calculation of the theoretical probability. One of those types is called a **Binomial Experiment**.

Properties of a binomial experiment (or Bernoulli trial)

1. Fixed number of trials, *n*, which means that the experiment is repeated a specific number of times.
2. The *n* trials are independent, which means that what happens on one trial does not influence the outcomes of other trials.
3. There are only two outcomes, which are called success and failure.
4. The probability of success doesn't change from trial to trial, where p = probability of success and *q* = probability of failure, *q* = 1-*p*.

Once a binomial experiment has been verified, then binomial probabilities can be calculated. This is important because binomial probabilities come up often in real life. Examples of binomial experiments are:

- Toss a fair coin ten times, and find the probability of getting two heads.
- Question twenty people in class, and look for the probability of more than half being women.
- Shoot five arrows at a target, and find the probability of hitting it five times.

To develop the process for calculating the probabilities in a binomial experiment, consider Example 5.2.1.

Binomial Distribution Formula

The random variable in a binomial experiment is *x* = number of successes. However, success is not always a good thing. Sometimes, success is bad, like finding a defect. Success means you observed the outcome you wanted to see happen.

Definition 5.2.1

The binomial Formula for the probability of *x* successes in *n* trials is

$$P(x) = {}_n C_x p^x q^{n \cdot x} \text{ where } {}_nC_x = \frac{n!}{x!(n-x)!}$$

The ${}_nC_x$ is the number of combinations of *n* things taking *x* at a time. It is read "*n* choose x".

When solving problems, make sure you define your random variable and state what *n, p, q,* and *x* are. Without doing this, the problems are a great deal harder.

Example 5.2.1: Calculating Binomial Probabilities

When looking at a person's eye color, it turns out that 1% of people in the world have green eyes ("What percentage of," 2013). Consider a group of 20 people. Round the results to three decimal places.

a. State the random variable.
b. Verify that this is a binomial experiment.
c. Find the probability that none have green eyes.
d. Find the probability that exactly nine have green eyes.
e. Find the probability that at most one has green eyes.
f. Find the probability that at least two have green eyes.

Solution

a. *x* = number of people with green eyes
b. Verification is presented below.
   - There are 20 people, and each person is a trial, so there is a fixed number of trials. In this case, $n = 20$.
   - If you assume that each person in the group is chosen at random, the eye color of one person doesn't affect the eye color of the next person, thus, the trials are independent.
   - Either a person has green eyes or they do not have green eyes, so there are only two outcomes. In this case, success is when a person has green eyes.
   - The probability of a person having green eyes is 0.01. This is the same for every trial since each person has the same chance of having green eyes. *p* = 0.01 and *q* = 1 - 0.01 = 0.99
c. $P(x = 0) =_{20} C_0(0.01)^0(0.99)^{20-0} \approx 0.818$
d. $P(x = 9) =_{20} C_9(0.01)^9(0.99)^{20-9} \approx 1.50 \times 10^{-13} \approx 0.000$
e. At most one means that one is the highest value you will have. Find the probability that *x* is less than or equal to three.

$$\begin{aligned} P(x \leq 1) &= P(x = 0) + P(x = 1) \\ &=_{20} C_0(0.01)^0(0.99)^{20} +_{20} C_1(0.01)^1(0.99)^{19} \\ &= 0.818 + 0.165 = 0.983 \end{aligned}$$

f. To find the probability that at least two people have green eyes, it is simpler to calculate it as $1 - P(\text{At Most } 1)$.

$$\begin{aligned} P(x \geq 2) &= 1 - P(x \leq 1) \\ &= 1 - 0.983 \\ &= 0.017 \end{aligned}$$

The binomial formula is cumbersome to use, and probabilities can be easily computed using technology. On the TI-83/84 calculator, the commands on the TI-83/84 calculators when the number of trials is equal to *n* and the probability of a success is equal to *p* are $\text{binompdf}(n, p, r)$ when you want to find *P*(x=r) and $\text{binomcdf}(n, p, r)$ when you want to find $P(x \leq r)$. If you want to find $P(x \geq r)$, then you use the property that $P(x \geq r) = 1 - P(x \leq r - 1)$, since $x \geq r$ and \(x[DISTR] menu.

Example 5.2.2 Using the Binomial Probability Table

A fair coin is flipped 10 times. Count how many times heads appear. Since each flip is independent with two outcomes, this is a binomial situation with a probability of 0.5 for heads. Use a binomial probability table to find the following probabilities.

a. Get exactly 6 heads.
b. Get at least 6 heads.
c. Get at most 3 heads.

Solution

Identify n and p; these values are n = 10 (number of trials) and p = 0.5 (probability of success).

Go to chapter 13.1: The Binomial Distribution Table for n = 2 to n = 10. In the table, search for n = 10 and p = 0.5. The column where these intersect will contain the probabilities.

a. Go to n = 10, X = 6, then find the probability in the column where p = 0.5. From the table, the probability of exactly 6 successes is P(6) = 0.205.
b. At least 6 translates to 6 or more. Go to n = 10, find all the probabilities from X = 6 until 10 under p =0.5. Add these probabilities together. Also, since we are finding the probability of at least 6, then we use the notation P(X $\geq$ 6). Thus, the answer is P(X $\geq$ 6) = P(6) + P(7) + P(8) + P(9) + P(10) = 0.205 + 0.117 + 0.044 + 0.010 + 0.001 = 0.377.
c. At most 3 translates to 3 or fewer. Go to n = 10, find all the probabilities from X = 0 until 3 under p =0.5. Add these probabilities together. Also, since we are finding the probability of at most 3, then we use the notation P(X $\leq$ 3). Thus, the answer is P(X $\leq$ 3) = P(0) + P(1) + P(2) + P(3) = 0.001 + 0.010 + 0.044 + 0.017 = 0.172.

Example 5.2.3 Using the Binomial Command on the TI-83/84

When looking at a person's eye color, it turns out that 1% of people in the world have green eyes ("What percentage of," 2013). Consider a group of 20 people.

a. State the random variable.
b. Find the probability that none have green eyes.
c. Find the probability that exactly nine have green eyes.
d. Find the probability that at most three have green eyes.
e. Find the probability that at most two have green eyes.

Solution

a. *x* = number of people with green eyes

b. You are looking for *P* (x=0). Since this problem is x=0, you use the binompdf command on the TI-83/84. On the TI83/84, you go to the **[DISTR]** menu, select the binompdf, and then type into the parentheses your *n*, *p*, and *x* values into your calculator, making sure you use the comma to separate the values. The command will look like $\text{binompdf}(20, .01, 0)$ and when you press **[ENTER]** you will be answered. (If you have the new software on the TI-84, the screen looks a bit different.)

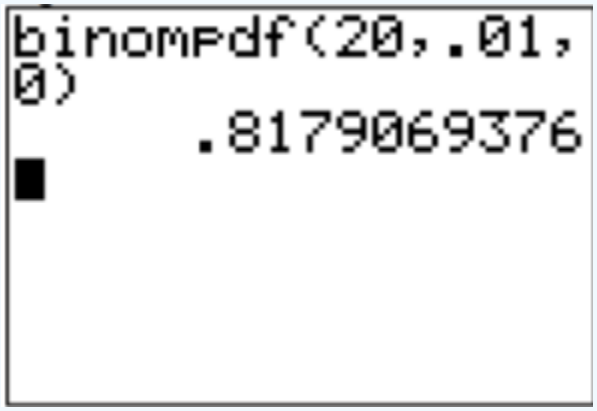


Figure 5.2.1: Calculator Results for binompdf

P (x=0) = 0.8179. Thus, there is an 81.8% chance that in a group of 20 people, none of them will have green eyes.

c. In this case, you want to find the *P* (x=9). Again, you will use the binompdf command or the dbinom command. Following the procedure above, you will have binompdf(20, .01, 9) on the TI-83/84. Your answer is $P(x = 9) = 1.50 \times 10^{-13}$. (Remember when the calculator gives you $1.50E - 13$, this is how they display scientific notation.) The probability that out of twenty people, nine have green eyes is a very small chance.

d. At most three means that three is the highest value you will have. Find the probability of *x* being less than or equal to three, which is $P(x \leq 3)$. This uses the binomcdf command on the TI-83/84. You use the command on the TI-83/84 of binomcdf(20, .01, 3).

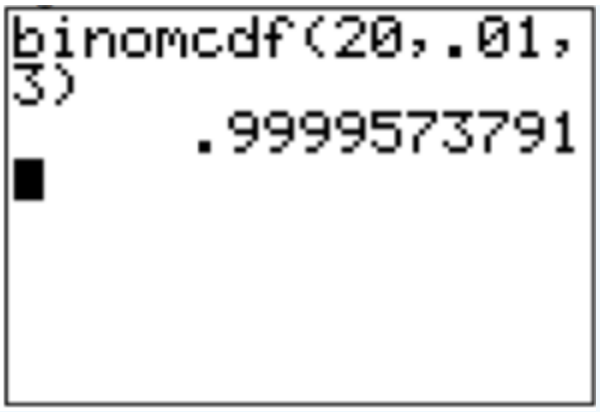


Figure 5.2.2: Calculator Results for binomcdf

Your answer is 0.99996. Thus, there is a good chance that in a group of 20 people, at most three will have green eyes. (Note: don't round this to one, since one means that the event will happen when there is a slight chance that it won't happen. It is best to write the answer out to enough decimal points so it doesn't round off to one.

e. You are looking for $P(x \leq 2)$. Again, use binomcdf or pbinom. Following the procedure above you will have $\text{binomcdf}(20, .01, 2)$ on the TI-83/84 and pbinom(2,20,0.01), with $P(x \leq 2) = 0.998996$. Again, there is a good chance that at most two people in the room will have green eyes.

Example 5.2.4 Calculating Binomial Probabilities

According to the Centers for Disease Control (CDC), about 1 in 88 children in the U.S. has been diagnosed with autism ("CDC-data and statistics," 2013). Suppose you consider a group of 10 children.

a. State the random variable.
b. Verify that this is a binomial experiment.
c. Find the probability that none have autism.
d. Find the probability that exactly seven have autism.
e. Find the probability that at least five have autism.
f. Find the probability that at most two have autism.
g. Suppose five children out of ten have autism. Is this unusual? What does that tell you?

Solution

a. *x* = number of children with autism

b.

1. There are 10 children, and each child is a trial, so there is a fixed number of trials. In this case, *n* = 10.
2. If you assume that each child in the group is chosen at random, then whether a child has autism does not affect the chance that the next child has autism. Thus, the trials are independent.
3. Either a child has autism or they do not have autism, so there are two outcomes. In this case, the success is that a child has autism.
4. The probability of a child having autism is 1/88. This is the same for every trial since each child has the same chance of having autism. $p = \frac{1}{88}$ and $q = 1 - \frac{1}{88} = \frac{87}{88}$.

c. Using the formula:

$$P(x = 0) = {}_{10}C_0 \left(\frac{1}{88}\right)^0 \left(\frac{87}{88}\right)^{10-0} \approx 0.892$$

Using the TI-83/84 Calculator:

$$P(x = 0) = \text{ binompdf } (10, 1 \div 88, 0) \approx 0.892$$

d. Using the formula:

$$P(x = 7) = {}_{10}C_7 \left(\frac{1}{88}\right)^7 \left(\frac{87}{88}\right)^{10-7} \approx 0.000$$

Using the TI-83/84 Calculator:

$$P(x = 7) = \text{ binompdf } (10, 1 \div 88, 7) \approx 2.84 \times 10^{-12}$$

e. Using the formula:

$$\begin{aligned} P(x \geq 5) &= P(x = 5) + P(x = 6) + P(x = 7) \\ &\quad + P(x = 8) + P(x = 9) + P(x = 10) \\ &= {}_{10}C_5 \left(\frac{1}{88}\right)^5 \left(\frac{78}{88}\right)^{10-5} + {}_{10}C_6 \left(\frac{1}{88}\right)^6 \left(\frac{78}{88}\right)^{10-6} \\ &\quad + {}_{10}C_7 \left(\frac{1}{88}\right)^7 \left(\frac{78}{88}\right)^{10-7} + {}_{10}C_8 \left(\frac{1}{88}\right)^8 \left(\frac{78}{88}\right)^{10-8} \\ &\quad + {}_{10}C_9 \left(\frac{1}{88}\right)^9 \left(\frac{78}{88}\right)^{10-9} + {}_{10}C_{10} \left(\frac{1}{88}\right)^{10} \left(\frac{78}{88}\right)^{10-10} \\ &= 0.000 + 0.000 + 0.000 + 0.000 + 0.000 + 0.000 \\ &= 0.000 \end{aligned}$$

Using the TI-83/84 Calculator:

To use the calculator, you need to use the complement.

$$\begin{aligned} P(x \geq 5) &= 1 - P(x < 5) \\ &= 1 - P(x \leq 4) \\ &= 1 - \text{ binomcdf } (10, 1 \div 88, 4) \\ &\approx 1 - 0.9999999 = 0.000 \end{aligned}$$

Notice, the answer is given as 0.000, since the answer is less than 0.000. Don't write 0, since 0 means that the event is impossible to happen. The event of five or more is improbable, but not impossible.

f. Using the formula:

$$\begin{aligned} P(x \leq 2) &= P(x = 0) + P(x = 1) + P(x = 2) \\ &= {}_{10}C_0 \left(\frac{1}{88}\right)^0 \left(\frac{78}{88}\right)^{10-0} + {}_{10}C_1 \left(\frac{1}{88}\right)^1 \left(\frac{78}{88}\right)^{10-1} \\ &\quad + {}_{10}C_2 \left(\frac{1}{88}\right)^2 \left(\frac{78}{88}\right)^{10-2} \\ &= 0.892 + 0.103 + 0.005 > 0.999 \end{aligned}$$

Using the TI-83/84 Calculator:

$$P(x \leq 2) = \text{ binomcdf } (10, 1 \div 88, 2) \approx 0.9998$$

g. Since the probability of five or more children in a group of ten having autism is much less than 5%, it is unusual to happen. If this does happen, then one may think that the proportion of children diagnosed with autism is more than 1/88.

Example 5.2.5 Calculating Binomial Probabilities

A university study finds that 60% of college students regularly drink coffee to stay focused while studying. Suppose we randomly select 10 college students.

a. What is the probability that exactly 7 of the students drink coffee regularly?

b. What is the probability that fewer than 3 students drink coffee regularly?
c. What is the probability that more than 8 students drink coffee regularly?

Solution

a. For this problem $n = 10, p = 0.6, and q = 1 - 0.6 = 0.4$.

$$\begin{aligned} P(X = 7) &= {}_{10}C_7 \cdot (0.6)^7(0.4)^3 \\ &= 0.215 \end{aligned}$$

b. For the problem, all the probabilities P(0), P(1), and P(2) must be computed and added together.

$$\begin{aligned} P(X < 3) &= P(0) + P(1) + P(2) \\ &= {}_{10}C_0 \cdot (0.6)^0(0.4)^{10} + {}_{10}C_1 \cdot (0.6)^1(0.4)^9 + {}_{10}C_2 \cdot (0.6)^2(0.4)^8 \\ &= 0.000 + 0.002 + 0.011 \\ &= 0.013 \end{aligned}$$

c. For the problem, all the probabilities P(9) and P(10) must be computed and added together.

$$\begin{aligned} P(X > 8) &= P(X = 9) + P(X = 10) \\ &= {}_{10}C_9 \cdot (0.6)^9(0.4)^1 + {}_{10}C_{10} \cdot (0.6)^{10}(0.4)^0 \\ &= 0.040 + 0.006 \\ &= 0.046 \end{aligned}$$

Exercises

Exercise 5.2.1

A research team studying social behavior is surveying 10 randomly selected professionals, where each person has a 0.3 probability of working in sociology-related careers (e.g., social work, policy analysis). Answer the following questions.

a. What is the probability that exactly 4 of the professionals work in sociology-related fields?
b. What is the probability that at least 6 of the professionals work in sociology-related fields?
c. What is the probability that at most 3 of the professionals work in sociology-related fields?

Answer available at the end of this section.

Exercise 5.2.2

A university career center is analyzing outcomes for 12 recent graduates, where each graduate has a 0.4 probability of entering a psychology-related profession (e.g., counseling, therapy, behavioral research). Answer the following questions.

a. What is the probability that exactly 5 graduates enter psychology-related careers?
b. What is the probability that more than 7 graduates enter psychology-related careers?
c. What is the probability that fewer than 4 graduates enter psychology-related careers?

Answer available at the end of this section.

Exercise 5.2.3

A publishing company is reviewing 8 job applicants, where each applicant has a 0.25 probability of being an English major qualified for writing and editing roles. Answer the following questions.

a. What is the probability that no more than 2 applicants are English majors?

b. What is the probability that at least 6 applicants are English majors?
c. What is the probability that exactly 4 applicants are English majors?

Answer available at the end of this section.

Exercise 5.2.4

A K-pop entertainment company is analyzing fan preferences at a concert. Each fan in a randomly selected group of 12 has a 0.4 probability of naming BTS as their favorite group.

Answer the following:

a. What is the probability that exactly 5 fans say BTS is their favorite group?
b. What is the probability that at least 8 fans say BTS is their favorite group?
c. What is the probability that no more than 3 fans say BTS is their favorite group?
d. Based on your answers, would you say BTS dominates fan preference in this group? Explain briefly.

Answer available at the end of this section.

Exercise 5.2.1

A fan signing event is being organized for BLACKPINK. Based on past events, each invited fan has a 60% chance of attending. Suppose 9 fans are randomly selected and invited.

Answer the following:

a. What is the probability that exactly 6 fans attend?
b. What is the probability that at least 7 fans attend?
c. What is the probability that at most 3 fans attend?

Answer available at the end of this section.

Authors

"5.2: Binomial Probability Distribution"(opens in new window) by Toros Berberyan(opens in new window), Tracy Nguyen(opens in new window), and Alfie Swan(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

Attributions

"5.2: Binomial Probability Distribution"(opens in new window) by Kathryn Kozak(opens in new window) is licensed under CC BY-SA 4.0(opens in new window)

Answers

Exercise 5.2.1

a. Exactly 4: 0.2001
b. At least 6: 0.0473
c. At most 3: 0.6496

Exercise 5.2.2

a. Exactly 5: 0.2270
b. More than 7: 0.0573
c. Fewer than 4: 0.2253

Exercise 5.2.3

a. No more than 2: 0.6785

b. At least 6: 0.0042
c. Exactly 4: 0.0865

Exercise 5.2.4

a. Exactly 5: 0.2270
b. At least 8: 0.0573
c. No more than 3: 0.2253
d. No, not necessarily. The most likely outcome is around 5 fans choosing BTS, which matches the expected number in a group of 12 when the probability is 0.4. Since the probability of at least 8 fans choosing BTS is only 0.0573, BTS would not be considered overwhelmingly dominant in this group, although they are still quite popular.

Exercise 5.2.1

a. Exactly 6: 0.251
b. At least 7: 0.232
c. At most 3: 0.119

This page titled 5.2: Binomial Probability Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 5.2: Binomial Probability Distribution by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution

Learning Objectives

- Calculate the mean of a binomial distribution by multiplying the number of trials by the probability of success.
- Determine the variance to measure the spread of outcomes, considering both success and failure probabilities.
- Compute the standard deviation as the square root of the variance to represent the typical distance from the mean.

In probability theory and statistics, the binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent trials, each with the same probability of success. When analyzing this distribution, three important measures help summarize its characteristics: the mean, the variance, and the standard deviation.

The mean, often referred to as the expected value, represents the average number of successes that can be expected over many repetitions of the binomial experiment.

Formulas for Mean, Variance, and Standard Deviation

Definition: Formulas

For a Binomial distribution, the following holds.

μ is the expected number of successes and is computed as $\mu = np$.

σ^2 is the variance of the number of successes and is computed as $\sigma^2 = npq$.

σ is the standard deviation for the number of successes and is computed as $\sigma = \sqrt{npq}$.

Where p is the probability of success and q = 1 - p.

Example 5.3.1 Finding the Probability Distribution, Mean, Variance, and Standard Deviation of a Binomial Distribution

When looking at a person's eye color, it turns out that 1% of people in the world have green eyes ("What percentage of," 2013). Consider a group of 20 people.

a. Find the mean.
b. Find the variance.
c. Find the standard deviation.

Solution

a. Since this is a binomial, then you can use the formula $\mu = np$. So $\mu = 20(0.01) = 0.2$ people. You expect on average that out of 20 people, less than 1 would have green eyes.
b. Since this is a binomial, then you can use the formula $\sigma^2 = npq$.

$$q = 1 - 0.01 = 0.99$$

$$\sigma^2 = 20(0.01)(0.99) = 0.198$$

c. Once you have the variance, you just take the square root of the variance to find the standard deviation.

$$\sigma = \sqrt{0.198} \approx 0.445$$

Example 5.3.2

A calculator manufacturer produces scientific calculators with a defect probability of 0.02. Assume that a quality control department for the manufacturer gathers a batch of 800 such calculators.

a. Find the mean.
b. Find the variance.
c. Find the standard deviation.

Solution

a. For the is problem *n = 800, p = 0.02, and q = 1 - 0.02 = 0.98*.

The formula $\mu = np$ is used.

$\mu = 800(0.02) = 16$

The expected number of defective calculators in the batch of 800 is 16.

b. The variance is computed using the formula $\sigma^2 = npq$.

$\sigma^2 = 800(0.02)(0.98) = 15.68$

c. The standard deviation is computed by taking the square root of the variance.

$\sigma = \sqrt{15.68} \approx 3.960$

Exercises

Exercise 5.3.1

A group of 15 college students launches small side businesses while in school (e.g., reselling, tutoring, online shops). Based on past data, each student has a 0.4 probability that their business will still be active after one year. Let X be the number of students whose businesses succeed.

a. Find the mean number of successful student businesses.
b. Find the variance.
c. Find the standard deviation.

Answer available at the end of this section.

Exercise 5.3.2

A group of 18 college students is surveyed about whether they choose to eat at a local restaurant instead of the campus cafeteria on a given day. Each student has a 0.5 probability of choosing the restaurant. Let X be the number of students who choose the restaurant.

a. Find the mean number of successful student businesses.
b. Find the variance.
c. Find the standard deviation.

Answer available at the end of this section.

Exercise 5.3.3

A college sustainability club has 16 student volunteers working in a campus garden. Each student has a 0.6 probability of showing up on a given weekend. Let X be the number of students who show up.

a. Find the mean number of students who attend.
b. Find the variance.
c. Find the standard deviation

Answer available at the end of this section.

Exercise 5.3.4

In a community survey, each selected resident has a 40% chance of responding. A researcher randomly selects 12 residents. Find:

a. The mean
b. The variance
c. The standard deviation

Answer available at the end of this section.

Exercise 5.3.1

A political science club surveys 25 college students about a campus policy issue. Each student has a 0.6 probability of supporting the policy. Let X be the number of students who support the policy.

a. Find the mean number of students who support the policy.
b. Find the variance.
c. Find the standard deviation.

Answer available at the end of this section.

Authors

"5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"5.3: Mean and Standard Deviation of Binomial Distribution" by Kathryn Kozak is licensed under CC BY-SA 4.0

Answers

Exercise 5.3.1

a. Mean: 6.0
b. Variance: 3.60
c. Standard deviation: 1.90

Exercise 5.3.2

a. Mean: 9.0
b. Variance: 4.50
c. Standard deviation: 2.12

Exercise 5.3.3

a. Mean: 9.6
b. Variance: 3.84
c. Standard deviation: 1.96

Exercise 5.3.4

a. Mean: 4.8
b. Variance: 2.88
c. Standard deviation: 1.70

Exercise 5.3.1

a. Mean: 15.0
b. Variance: 6.00
c. Standard deviation: 2.45

This page titled 5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 5.3: Mean and Standard Deviation of Binomial Distribution by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

5.4: Poisson Distribution

Learning Objectives

- Define the Poisson distribution to model the number of times an event occurs in a fixed interval of time or space under conditions of independence and a constant rate.
- Apply the Poisson distribution to rare events such as phone calls per hour or accidents per day.
- Calculate probabilities using the average number of occurrences within the specified interval.

Introduction

The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

two main characteristics of a Poisson experiment

1. The Poisson probability distribution gives the probability of several events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on average, there are five words spelled incorrectly in 100 pages. The interval is 100 pages.
2. The Poisson distribution may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1,000). You will verify the relationship in the homework exercises. n is the number of trials, and p is the probability of a "success."

The Formula for Poisson Distribution

The formula to compute the Poisson Distribution is presented below.

$$P(X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$$

- X = 0, 1, 2, 3, ...
- λ = mean number of occurrences in the interval.
- e = Euler's constant ≈ 2.718280

Examples

Example 5.4.1

Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah will receive more than one call in the next 15 minutes?

Solution

Let X = the number of calls Leah receives in 15 minutes. (The *interval of interest* is 15 minutes or $\frac{1}{4}$ hour.)

$$x = 0, 1, 2, 3, ... \tag{5.4.1}$$

If Leah receives, on average, six telephone calls in two hours, and there are eight 15-minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)(6) = 0.75$ calls in 15 minutes, on average. So, $\lambda = 0.75$ for this problem.

$X \sim P(0.75)$

Find $P(x > 1)$. $P(x > 1) = 0.1734$ (calculator or computer)

- Press 1 – and then press 2nd DISTR.
- Arrow down to poissoncdf. Press ENTER.
- Enter (.75,1).
- The result is $P(x > 1) = 0.1734$.

$P(X) = \frac{e^{-\lambda}\lambda^{X}}{X!}$

Recall that the sum of all the probabilities in a discrete distribution is 1. To use the formula, we must find P(0) and P(1) and subtract the value from 1. The result is P(x > 1).

$p(0) = \frac{e^{-0.75} \cdot 0.75^{0}}{0!}$ = 0.4724

$p(1) = \frac{e^{-0.75} \cdot 0.75^{1}}{1!}$ = 0.3542

P(x > 1) = 1 - P(0) - P(1) = 1 - 0.4724 - 0.3542

P(x > 1) = 0.1734

The probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734:

$P(x > 1) = 1 - \text{poissoncdf}(0.75, 1)$.

The graph of $X \sim P(0.75)$ is:

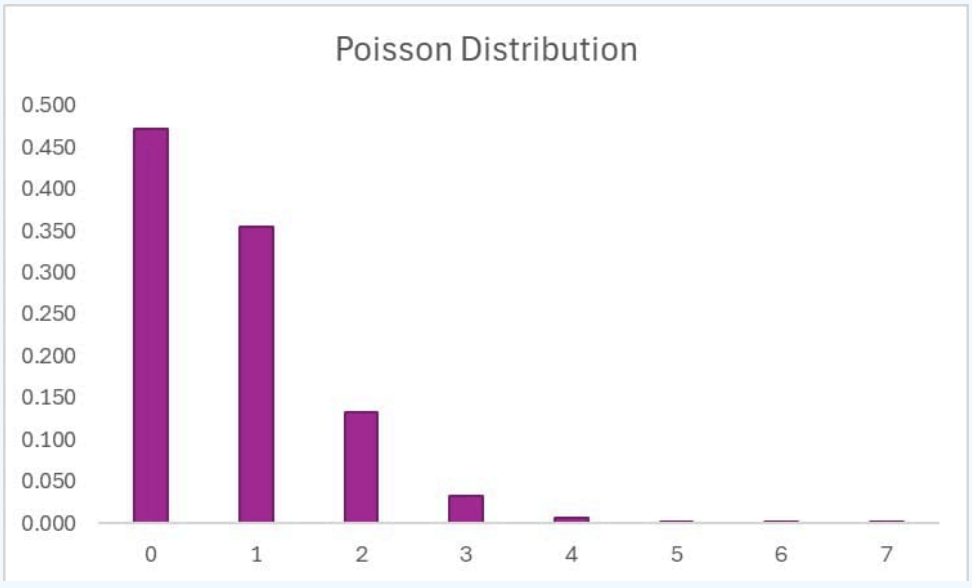


Figure 5.4.1:Graph of Poisson Distribution

The *y*-axis contains the probability of x where $X =$ the number of calls in 15 minutes.

Example 5.4.2

A customer service center receives about ten emails every half-hour. What is the probability that the customer service center will receive more than four emails in the next six minutes?

Solution

$P(x > 4) = 0.0526$

Recall that the sum of all the probabilities in a discrete distribution is 1. To use the formula, we must find P(5), P(6), P(7), and other values until the probabilities become close to zero and then subtract the values from 1. The result is P(x > 4).

$\lambda = \frac{10}{30} \cdot 6 = 2$

$$P(5) = \frac{e^{-2} \cdot 2^5}{5!} = 0.0361$$

$$P(6) = \frac{e^{-2} \cdot 2^6}{6!} = 0.0120$$

$$P(7) = \frac{e^{-2} \cdot 2^7}{7!} = 0.0034$$

$$P(8) = \frac{e^{-2} \cdot 2^8}{8!} = 0.0009$$

$$P(9) = \frac{e^{-2} \cdot 2^9}{9!} = 0.0002$$

All probabilities beyond P(9) are small values that can be treated as 0.

P(x > 4) = P(5) + P(6) + P(7) + P(8) + P(9) = 0.0361 + 0.0120 + 0.0034 + 0.0009 + 0.0002 = 0.0526

P(x > 4) =0.0526.

Example 5.4.3

According to Baydin, an email management company, an email user gets, on average, 144 emails per day. Let $X =$ be the number of emails an email user receives per hour. The discrete random variable X takes on the values $x = 0, 1, 2 \ldots$. The random variable X has a Poisson distribution: $X \sim P(147)$. Assume the interval of time is for 1 hour. The mean is $\frac{144}{24}$ = 6 emails per hour.

a. What is the probability that an email user receives exactly 2 emails per hour?
b. What is the probability that an email user receives at most 2 emails per hour?

Solutions

Using a TI-84+ use the following functions.

a. $P(x = 160) = \text{poissonpdf}(6, 2) \approx 0.0446$
b. $P(x \leq 160) = \text{poissoncdf}(6, 2) \approx 0.0620$

Another approach is to use the formula to calculate the answers.

a. For this part λ = 6 and x = 2.

$P(2) = \frac{e^{-6} \cdot 6^2}{2!} = 0.0446$

b. For this part λ = 6 and the probabilities P(0), P(1), and P(2) must be computed and added together.

$P(0) = \frac{e^{-6} \cdot 6^0}{0!} = 0.0025$

$P(1) = \frac{e^{-6} \cdot 6^1}{1!} = 0.0149$

$P(2) = \frac{e^{-6} \cdot 6^2}{2!} = 0.0446$

$P(x \leq 2)$= P(0) + P(1) + P(2) = 0.0025 + 0.0149 + 0.0446 = 0.0620

$P(x \leq 2)$ =0.0620

Example 5.4.4

An office building contains 100 office spaces of 350 square feet. The carpets in the office space have around 2 stains per 350 square feet. For a sample of 3 office spaces of 350 square feet, compute the following probabilities.

a. What is the probability that the 3 office spaces have 8 stains?
b. What is the probability that the office spaces have at most 3 stains?

Solution

Using a TI-84+ use the following functions.

The average number of stains in the 3 office spaces is $\lambda = \frac{2}{350} \cdot 1050 = 6$

a. $P(x = 8) = \text{poissonpdf}(6, 8) \approx 0.1033$
b. $P(x \leq 3) = \text{poissoncdf}(6, 3) \approx 0.1512$

Another approach is to use the formula to calculate the answers.

a. Using the formula, compute p(8).

$P(8) = \frac{e^{-6} \cdot 6^8}{8!} = 0.1033$

b. Using the formula compute P(0), P(1), P(2), and P(3), and then add them together.

$P(0) = \frac{e^{-6} \cdot 6^0}{0!} = 0.0025$

$P(1) = \frac{e^{-6} \cdot 6^1}{1!} = 0.0149$

$P(2) = \frac{e^{-6} \cdot 6^2}{2!} = 0.0446$

$P(3) = \frac{e^{-6} \cdot 6^3}{3!} = 0.0892$

$P(x \leq 3)$= P(0) + P(1) + P(2) + P(3) = 0.0025 + 0.0149 + 0.0446 + 0.0892 = 0.1512

$P(x \leq 3)$ =0.1512.

Example 5.4.5

A publisher notices that there are 2 small typos in 500 pages of a typical book.

a. How many small typos does a 1,000-page book have on average?
b. What is the probability that a 1000-page book will have 5 small typos?
c. What is the probability that a 1000-page book will have more than 5 small typos?

Solutions

Using a TI-84+ use the following functions.

a. Let $X =$ be the number of texts that a user sends or receives in one hour. The average number of small typos per 1000 pages of a book is $\frac{2}{500} \cdot 1000 = 4$.
b. $X \sim P(4)$, so $P(x = 5) = \text{poissonpdf}(4, 5) \approx 0.1563$
c. $P(x > 5) = 1\text{–}P(x \leq 5) = 1\text{–poissoncdf}(4, 4) \approx 1\text{–}0.6288 = 0.3712$

Another approach is to use the formula to calculate the answers.

a. See part a above.

b. For the formula, compute p(5).

$$P(5) = \frac{e^{-4} \cdot 4^5}{5!} = 0.1563$$

c. Compute P(0), P(1), P(2), P(3), and P(4). Add them together and subtract the sum from 1.

$$P(0) = \frac{e^{-4} \cdot 4^0}{0!} = 0.0183$$

$$P(1) = \frac{e^{-4} \cdot 4^1}{1!} = 0.0733$$

$$P(2) = \frac{e^{-4} \cdot 4^2}{4!} = 0.1465$$

$$P(3) = \frac{e^{-4} \cdot 4^3}{3!} = 0.1954$$

$$P(4) = \frac{e^{-4} \cdot 4^4}{4!} = 0.1954$$

$P(x > 5)$ = 1 - P(0) - P(1) - P(2) - P(3) - P(4) = 1 - 0.0183 - 0.0733 - 0.1465 - 0.1954 - 0.1954 = 1 - 0.6288 = 0.3712

$P(x > 5)$ =0.3712

Example 5.4.6

Atlanta's Hartsfield-Jackson International Airport is the busiest in the world. On average there are 2,500 arrivals and departures each day.

a. How many airplanes arrive and depart the airport per hour?
b. What is the probability that there are exactly 100 arrivals and departures in one hour?
c. What is the probability that there are at most 100 arrivals and departures in one hour?

Solution

a. Let $X =$ the number of airplanes arriving and departing from Hartsfield-Jackson in one hour. The average number of arrivals and departures per hour is $\frac{2,500}{24} \approx 104.1667$.
b. $X \sim P(104.1667)$, so $P(x = 100) = \text{poissonpdf}(104.1667, 100) \approx 0.0366$.
c. $P(x \leq 100) = \text{poissoncdf}(104.1667, 100) \approx 0.3651$.

The Poisson distribution can be used to approximate probabilities for a binomial distribution. This next example demonstrates the relationship between the Poisson and the binomial distributions. Let n represent the number of binomial trials and let p represent the probability of success for each trial. If n is large enough and p is small enough then the Poisson approximates the binomial very well. In general, n is considered "large enough" if it is greater than or equal to 20. The probability p from the binomial distribution should be less than or equal to 0.05. When the Poisson is used to approximate the binomial, we use the binomial mean $\mu = np$. The variance of X is $\sigma^2 = \sqrt{\mu}$ and the standard deviation is $\sigma = \sqrt{\mu}$. The Poisson approximation to a binomial distribution was commonly used in the days before technology made both values very easy to calculate.

Example 5.4.7

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution

Let X = the number of days with low seismic activity.

Using the binomial distribution:

- $P(x = 10) = \text{binompdf}(200, .0102, 10) \approx \ 0.000039$

Using the Poisson distribution:

- Calculate $\mu = np = 200(0.0102) \approx 2.04$
- $P(x = 10) = \text{poissonpdf}(2.04, 10) \approx 0.000045$

We expect the approximation to be good because n is large (greater than 20) and p is small (less than 0.05). The results are close—both probabilities reported are almost 0.

Example 5.4.8

Add example text here.

Solution

Add example text here.

Example 5.4.9

On May 13, 2013, starting at 4:30 PM, the probability of moderate seismic activity for the next 48 hours in the Kuril Islands off the coast of Japan was reported at about 1.43%. Use this information for the next 100 days to find the probability that there will be low seismic activity in five of the next 100 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution

Let $X =$ the number of days with moderate seismic activity.

Using the binomial distribution: $P(x = 5) = \text{binompdf}(100, 0.0143, 5) \approx 0.0115$

Using the Poisson distribution:

- Calculate $\mu = np = 100(0.0143) = 1.43$
- $P(x = 5) = \text{poissonpdf}(1.43, 5) = 0.0119$

We expect the approximation to be good because n is large (greater than 20) and p is small (less than 0.05). The results are close—the difference between the values is 0.0004.

Authors

"5.4: Poisson Distribution" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

Attributions

"4.7: Poisson Distribution", OpenStax is licensed under CC BY 4.0 / A derivative from the original work

This page titled 5.4: Poisson Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

5.5: Chapter 5 Formulas

Chapter 5 Formulas

Mean for a Discrete Probability Distribution

The mean of a discrete probability distribution represents the long-run average or central value of the outcomes if an experiment were repeated many times. It is found by multiplying each possible value by its corresponding probability and summing the results.

$$\mu = \sum x \cdot P(x)$$

Where:

- μ is the mean
- x represents each possible value
- $P(x)$ is the probability of each value x

Variance for a Discrete Probability Distribution

The variance of a discrete probability distribution measures how much the values of the distribution differ from the mean. It represents the average of the squared differences between each value and the mean, weighted by their probabilities.

$$\sigma^2 = \sum x^2 \cdot P(x) - \mu^2$$

Where:

- σ^2 is the variance
- x is each possible value
- $P(x)$ is the probability of each value x
- $\sum x^2 \cdot P(x)$ is the expected value of the square of x
- μ is the mean (expected value)

Standard Deviation for a Discrete Probability Distribution

The standard deviation of a discrete probability distribution is a measure of the spread of the distribution. It tells how much the values typically deviate from the mean. It is calculated by taking the square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{Variance}}$$

Where:

- σ is the standard deviation
- σ^2 is the variance of the distribution

Expected Value

The expected value, denoted as $E(X)$, is the theoretical average of a random variable in a probability distribution. It represents the long-run average outcome if an experiment were repeated many times. For a discrete random variable, it is calculated by multiplying each possible value by its probability and summing the results.

$$E(X) = \sum x \cdot P(x)$$

Where:

- $E(X)$ is the expected value
- x is each possible value of the random variable
- $P(x)$ is the probability of x

Probability Formula for the Binomial Distribution

The binomial distribution models the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: success or failure. Each trial has the same probability of success. The binomial distribution is used when you want to know the likelihood of getting exactly a certain number of successes in repeated trials.

$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}$ or $P(x) = {}_nC_x \cdot p^x \cdot (1-p)^{n-x}$

Where:

- $P(x)$ is the probability of exactly x successes
- n is the number of trials
- x is the number of successes
- p is the probability of success on a single trial
- $1-p$ is the probability of failure
- $\frac{n!}{x!(n-x)!}$ is the number of combinations (n choose x)
- ${}_nC_x$ also represents the number of combinations of n items taken x at a time

Mean for the Binomial Distribution

The mean of a binomial distribution represents the expected number of successes in a fixed number of independent trials, each with the same probability of success. It indicates the long-run average outcome if the experiment were repeated many times.

$\mu = n \cdot p$

Where:

- μ is the mean (expected value)
- n is the number of trials
- p is the probability of success on a single trial

Variance for the Binomial Distribution

The variance of a binomial distribution measures how much the number of successes is expected to vary from the mean across repeated trials. It reflects the spread or dispersion of possible outcomes in a binomial experiment.

$\sigma^2 = n \cdot p \cdot q$

Where:

- σ^2 is the variance
- n is the number of trials
- p is the probability of success
- q is the probability of failure, where $q = 1-p$

Standard Deviaiton for the Binomial Distribution

The standard deviation of a binomial distribution measures the average distance of the number of successes from the mean. It indicates how much the outcomes typically vary in a binomial experiment. It is the square root of the variance and gives a sense of the spread of the distribution.

$\sigma = \sqrt{n \cdot p \cdot q}$

Where:

- σ is the standard deviation
- n is the number of trials
- p is the probability of success
- q is the probability of failure, where $q = 1 - p$

Authors

"5.5: Chapter 5 Formulas" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 5.5: Chapter 5 Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

5.6: Chapter 5 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Binomial Probability: The probability of obtaining a specific number of successes in a fixed number of independent trials, where each trial has only two possible outcomes (success or failure).

Binomial Probability Distribution: A discrete probability distribution that represents the number of successes in a set number of identical and independent trials, with the same probability of success for each trial.

Continuous Variable: A variable that can take on an infinite number of possible values within a given range. These values are measurable, not countable (e.g., height, time, temperature).

Discrete Probability Distribution: A listing or function that shows the probabilities of all possible values of a discrete random variable, with each probability between 0 and 1 and the total adding up to 1.

Discrete Variable: A variable that can only take on specific, separate values, usually whole numbers that can be counted (e.g., number of books, number of people).

Expected Value: The theoretical average of a random variable, representing the long-term outcome if the experiment is repeated many times.

Mean: The average value of a discrete probability distribution, calculated by considering the weighted sum of all possible values.

Probability Distribution: A function or list that shows the likelihood of each possible outcome of a random variable.

Random Variable: A variable whose value is determined by the outcome of a random event or experiment. It can be either discrete or continuous.

Standard Deviation: A measure that describes how spread out the values of a probability distribution are around the mean.

Variance: The average of the squared differences between each value and the mean, indicating the level of dispersion in a distribution.

Key Symbols

X: Random variable

$P(X)$: Probability of the random variable X taking on a specific value

$E(X)$: Expected value (mean) of the random variable X

μ: Population mean (same as $E(X)$ for theoretical distributions)

σ^2: Variance of a probability distribution

σ: Standard deviation of a probability distribution

n: Number of trials in a binomial experiment

p: Probability of success in a single trial

q: Probability of failure in a single trial (where $q = 1 - p$)

X: Number of successes (often used in binomial probability)

$\sum$: Summation symbol, indicating the sum over all values

Authors

"5.6: Chapter 5 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 5.6: Chapter 5 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

6: CONTINUOUS PROBABILITY DISTRIBUTION

6.1: Introduction to the Normal Distribution
6.2: Finding Probabilities for the Normal Distribution
6.3: The Central Limit Theorem
6.4: Formulas for Chapter 6
6.5: Chapter 6 - Key Terms and Symbols

This page titled 6: Continuous Probability Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

6.1: Introduction to the Normal Distribution

Learning Objectives

- Define the normal distribution as a continuous, bell-shaped, symmetric curve modeling many natural phenomena with values clustering around the mean.
- Describe the normal distribution using its mean and standard deviation.
- Use z-scores to standardize values and determine probabilities or areas under the curve with tables or calculators.

Many real-life problems produce a histogram that is a symmetric, unimodal, and bell-shaped continuous probability distribution. For example: height, blood pressure, and cholesterol level. However, not every bell-shaped curve is a normal curve. In a normal curve, there is a specific relationship between its “height” and its “width.”

Normal curves can be tall and skinny, or they can be short and fat. They are all symmetric, unimodal, and centered at μ, the population mean. *Figure* $6.1.1$ shows two different normal curves drawn on the same scale. Both have $\mu = 0$, but the one on the left has a standard deviation of 1, and the one on the right has a standard deviation of .25. Notice that the larger standard deviation makes the graph wider (more spread out) and shorter.

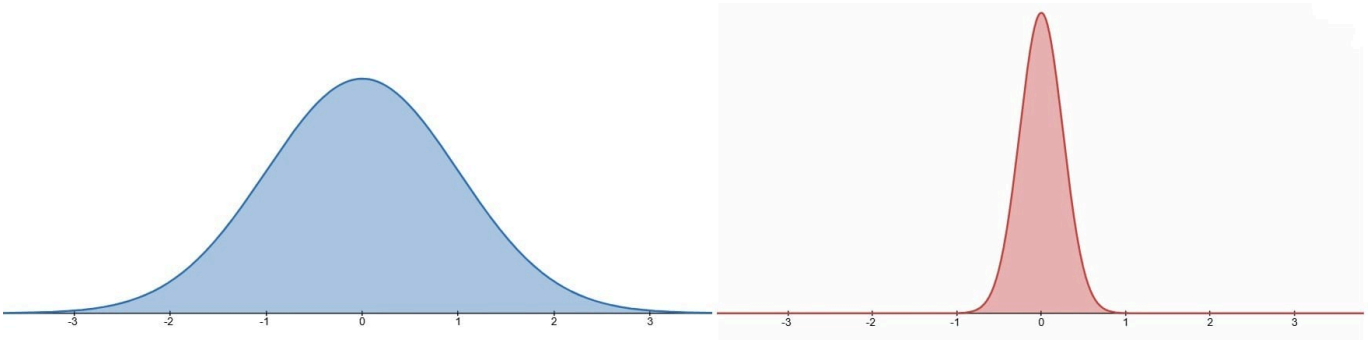


Figure $6.1.1$: Two Normal Distributions with Different Variances

Every normal curve has common features. These are detailed below.

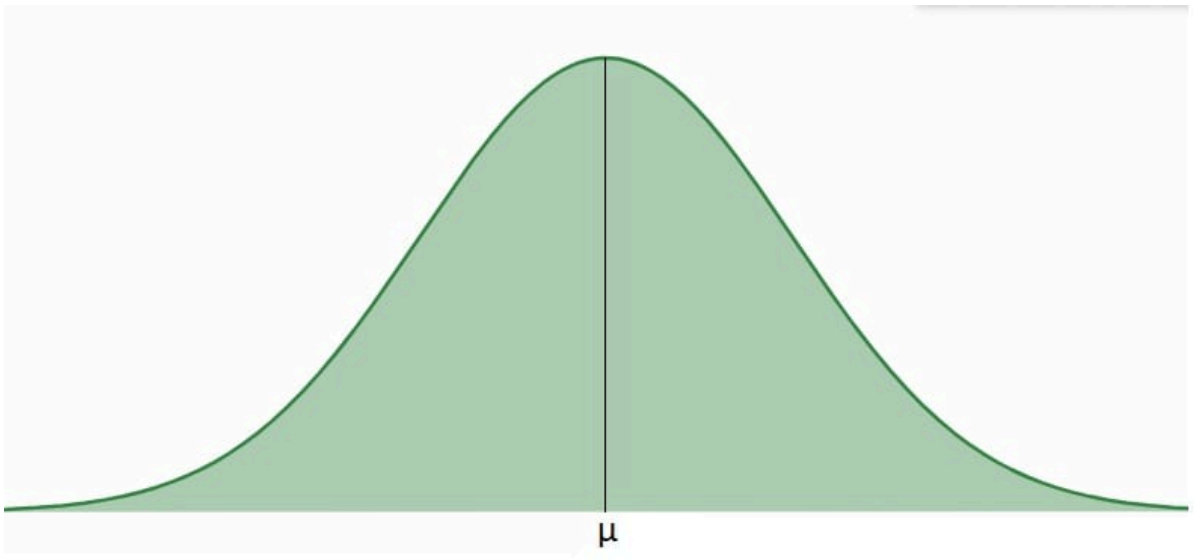


Figure $6.1.2$: Key Elements of a Normal Distribution

- The center, or the highest point, is at the population mean, μ.
- The standard deviation, σ, measures how spread out the distribution is.
- The area under the whole curve is exactly 1. Therefore, the area under the half below or above the mean is 0.5.
- The mean, median, and mode are all equal as the distribution is symmetric about the mean.

Before looking at the process for finding the probabilities under the normal curve, it is somewhat useful to look at the **Empirical Rule** that gives approximate values for these areas. The Empirical Rule is just an approximation, and it will only be used in this section to give you an idea of what the size of the probabilities is for different shadings. A more precise method for finding probabilities for the normal curve will be demonstrated in the next section. Please do not use the empirical rule except for rough estimates.

Definition: Empirical Rule

The **Empirical Rule** for any normal distribution:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- Approximately 99.7% of the data is within three standard deviations of the mean.

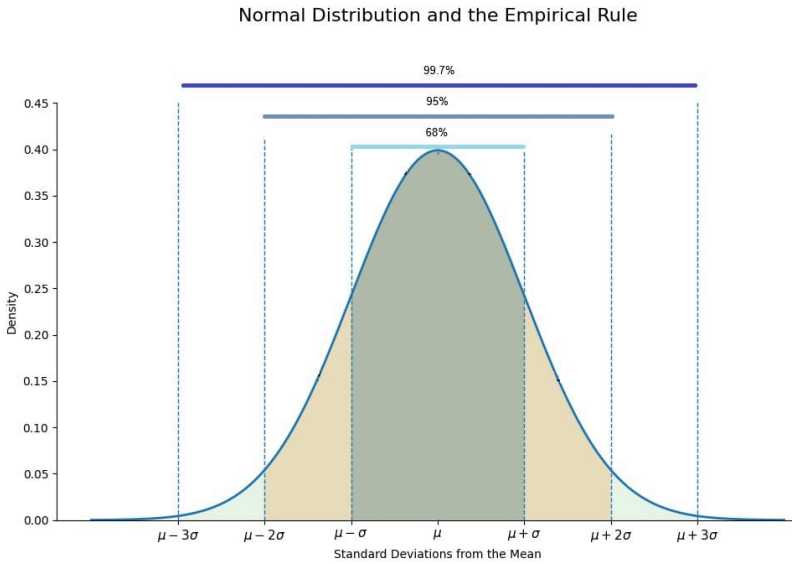


Figure 6.1.3: Empirical Rule for Normal Distribution

Be careful, there is still some area left over at each end. Remember, the maximum probability can be 100%, so if you calculate 100%-99.7%=0.3%, you will see that for both ends together, there is 0.3% of the curve. Because of symmetry, you can divide this equally between both ends and find that there is 0.15% in each tail beyond the $\mu \pm 3\sigma$.

Read and Interpret the Standard Normal Distribution Table

The standard normal distribution is a special type of normal distribution that has a mean of 0 and a standard deviation of 1. It is used as a reference to understand how data values compare to the average in terms of standard deviations. This distribution helps us calculate probabilities using three basic rules.

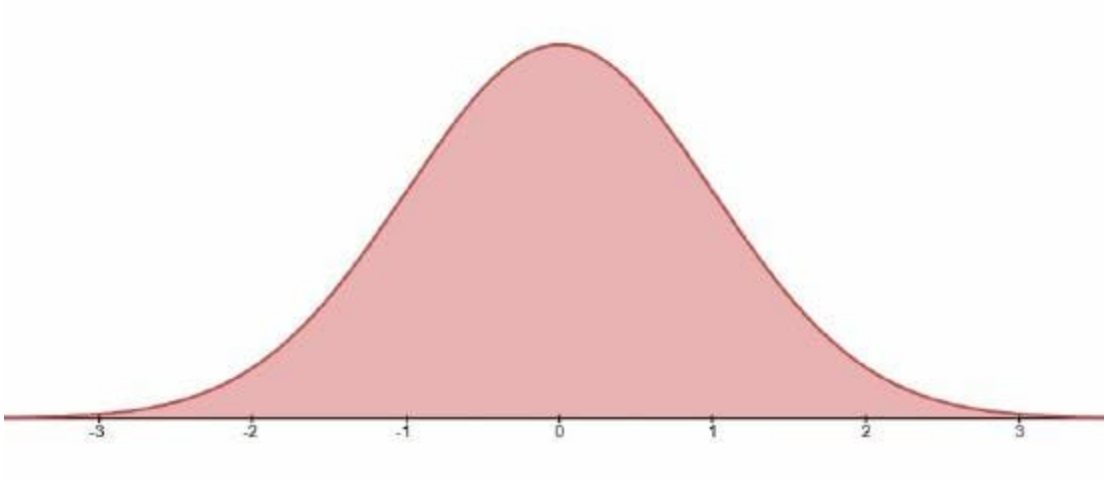


Figure 6.1.4: Graph of the Standard Normal Distribution with Mean = 0 and Standard Deviation = 1.

Definition: Rule #1 Finding the Area to the Left of any Z-value

To find the area to the left of any z-value, look up the z-value in the Standard Normal Distribution Table. The value listed is the area.

Example 6.1.1 Finding the Area to the Left of Any Z-Value

Find the area under the Standard Normal Distribution Curve to the left of the given z-values.

a. To the left of z = 2.43.
b. To the left of z = –1.57

Solution

a. Locate the row for the first part of the z-value (2.4). Move across the columns to the one labeled 0.03. Find the value where the row and column meet. This gives the cumulative area to the left of z = 2.43, which is 0.9925.

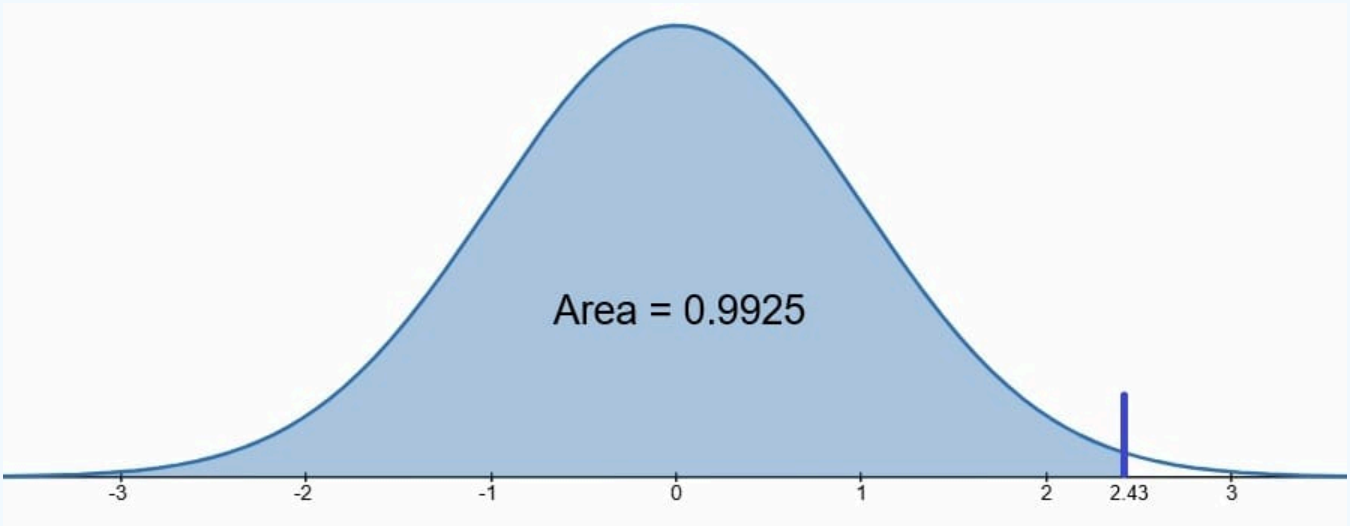


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

b. Locate the row for the first part of the z-value (–1.5). Move across the columns to the one labeled 0.07. Find the value where the row and column meet. This gives the cumulative area to the left of z = –1.57, which is 0.0582.

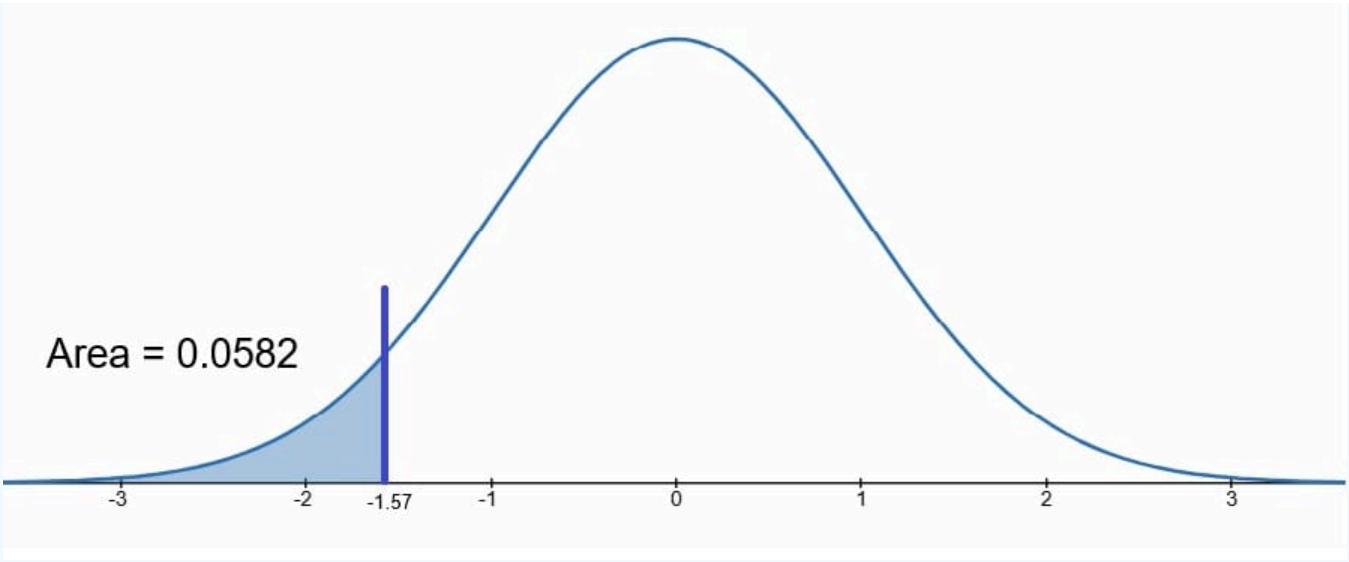


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

Definition: Rule #2 Finding the Area to the Right of any Z-value

To find the area to the right of any z-value, look up the z-value in the Standard Normal Distribution Table, then subtract the listed area from one.

Example \({6{.}1{.}\eatSpaces{2} \endSpaces}\ Finding the Area to the Right of Any Z-Value

Find the area under the Standard Normal Distribution Curve to the right of the given z-values.

a. To the right of z = 0.95.
b. To the right of z = –3.36

Solution

a. Locate the row for the first part of the z-value (0.9). Move across the columns to the one labeled 0.05. Find the value where the row and column meet. This gives the cumulative area to the left of z = 0.95, which is 0.8289. Now, subtract this area from 1 to get 1 – 0.8289 = 0.1711.

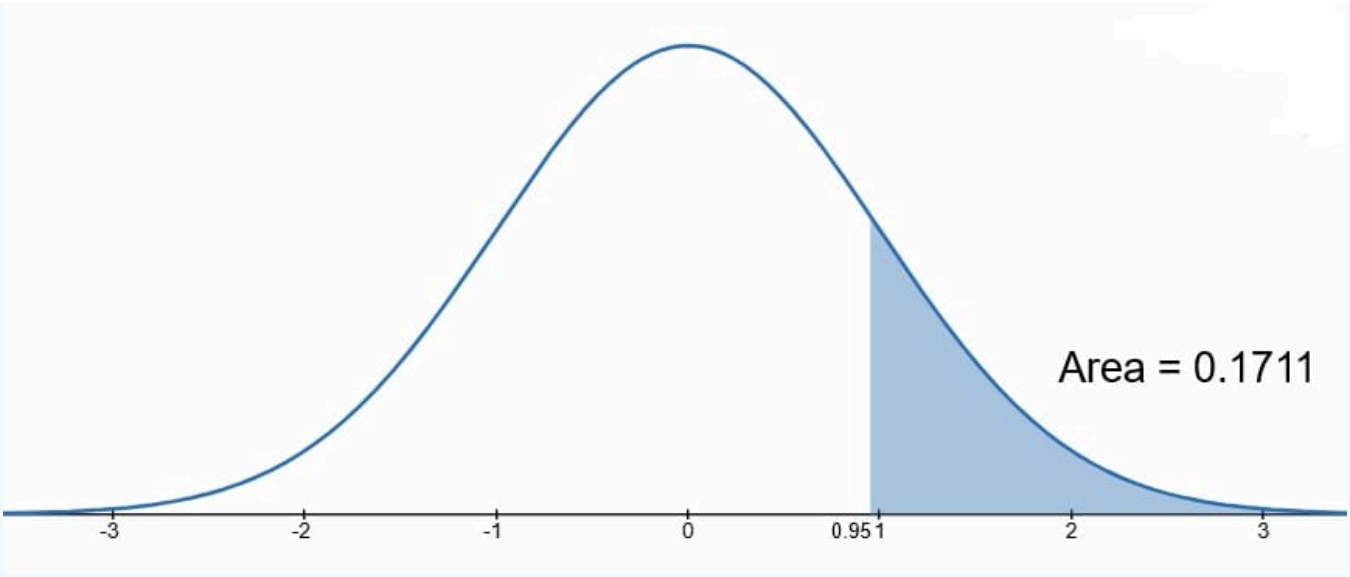


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

b. Locate the row for the first part of the z-value (–3.3). Move across the columns to the one labeled 0.06. Find the value where the row and column meet. This gives the cumulative area to the left of z = –3.36, which is 0.0004. Now, subtract this area from 1 to get 1 – 0.0004 = 0.9996.

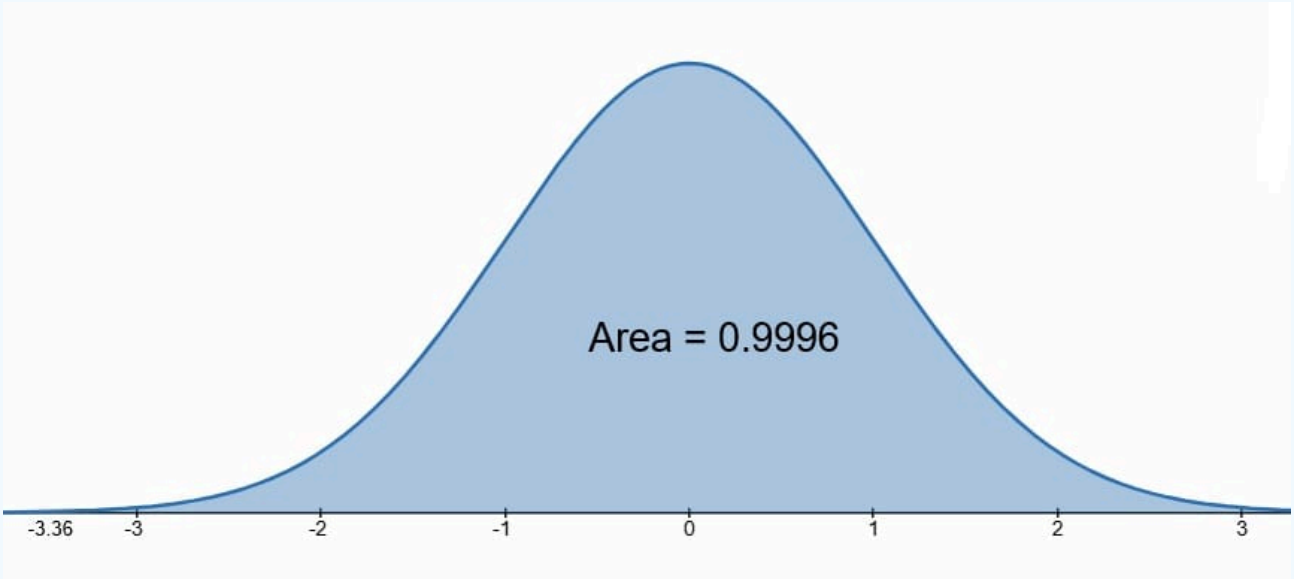


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

Definition: Rule #3 Finding the Area Between Any Two Z-values

To find the area between any two z-values, look up both z-values in the Standard Normal Distribution Table, then subtract the smaller area from the larger area.

Example 6.1.3 Finding the Area Between Any Two Z-Values

Find the area under the Standard Normal Distribution Curve between the two given z-values.

a. Between z = 1.42 and z = 1.80

b. Between z = –0.47 and z = 2.82
c. Between z = –4.02 and –1.67.

Solution

a. Start by finding the area to the left of z = 1.80. According to the table, the area to the left of z = 1.80 is 0.9641. Next, find the area to the left of z = 1.42, which is 0.9222. To get the area between the two z-values, subtract the smaller area from the larger one: 0.9641 minus 0.9222 equals 0.0419.

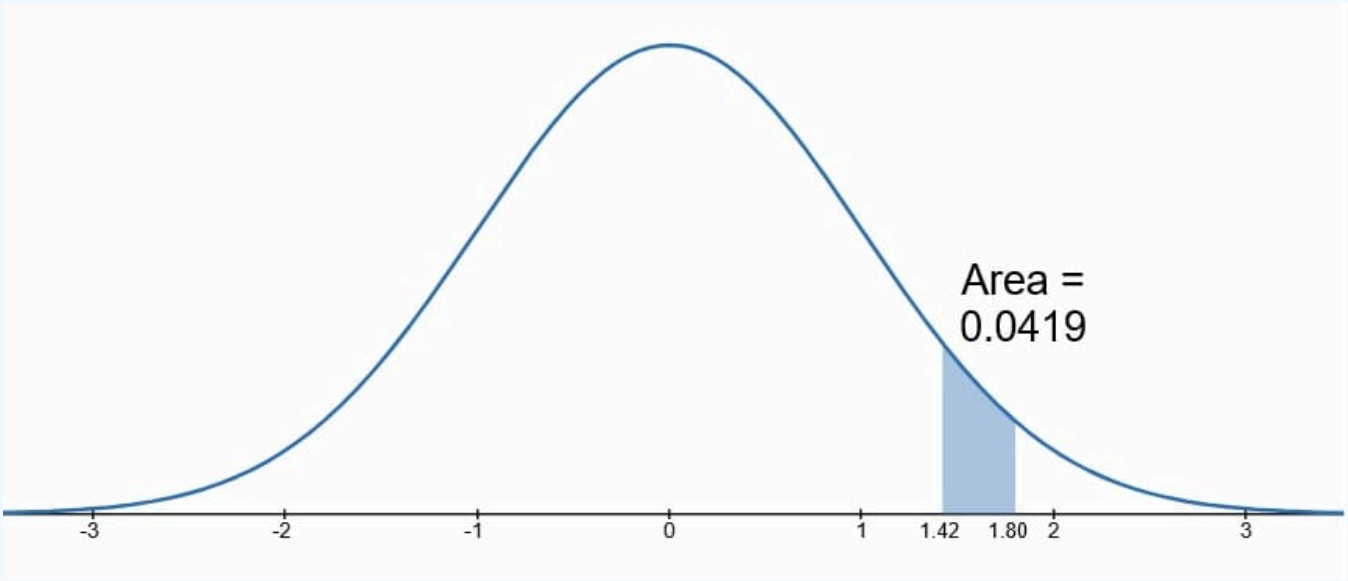


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

b. Start by finding the area to the left of z = 2.82. According to the table, the area to the left of z = 2.82 is 0.9976. Next, find the area to the left of z = –0.47, which is 0.3192. To get the area between the two z-values, subtract the smaller area from the larger one: 0.9976 minus 0.3192 equals 0.6784.

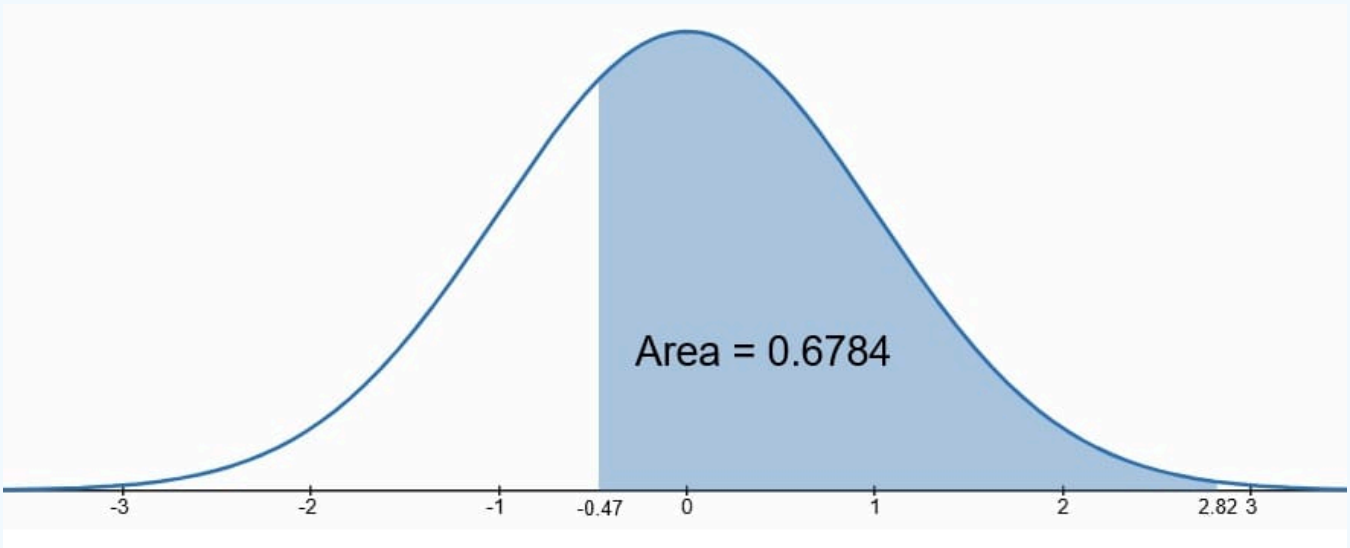


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

c. Start by finding the area to the left of z = –1.67. According to the table, the area to the left of z = –1.67 is 0.0475. Next, find the area to the left of z = –4.02, which is 0.0001. Since this value is less than z = –3.49, we use the area of 0.0001 as noted below the table. To get the area between the two z-values, subtract the smaller area from the larger one: 0.0475 minus 0.0001 equals 0.0474.

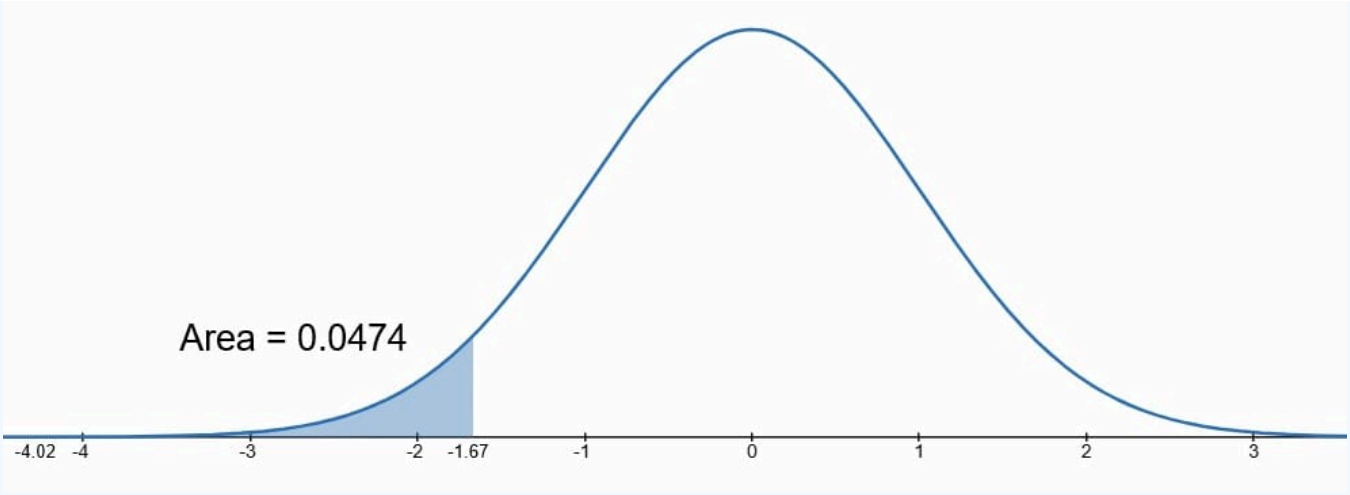


Figure 6.1.1: Copy and Paste Caption here. (Copyright; author via source)

Z-Score Computation

Z-score is a great way to compare different individuals and determine who was ranked relatively higher among his or her group.

The z-score formula: $z = \frac{\text{Individual} - \text{Mean}}{\text{Standard Deviation}}$

Example 6.1.4 Calculating the Z-scores

Kelly scored 50 points on Exam #1 for her statistics class. The mean of the class was 60, and the standard deviation was 4. John scored a 75 on Exam #1 for his geography class. The mean of the class was 90 with a standard deviation of 8. Who did relatively better compared to their class? Assume the distribution for both classes is normally distributed (bell-shaped).

Solution

- In this case, we can't compare their exam scores directly because they are two different exams. In addition, the total points for each exam could be different.
- Since we can't compare their exam scores directly, we need to convert each exam score to a z-score. Then, we will be able to compare their z-scores to know who scored relatively higher among their group.
- We need to use the Z formula to find the relative position of Kelly and John compared to the students taking the same exam.

Z-scores

Calculating Kelly's z-score	Calculating John's z-score
Individual = 50 (Kelly's Score) Mean = 60 Standard Deviation = 4	Individual = 75 (John's Score) Mean = 90 Standard Deviation = 8
$z = \frac{\text{Individual} - \text{Mean}}{\text{Standard Deviation}}$	$z = \frac{\text{Individual} - \text{Mean}}{\text{Standard Deviation}}$
$z = \frac{50 - 60}{4}$	$z = \frac{75 - 90}{8}$
$z = \frac{-5}{4}$	$z = \frac{-15}{8}$
z = -1.25	z = -1.875 $\approx$ -1.88

Table 6.1.5: Computing Z-scores

Since Kelly's z-score was higher than John's (-1.25) is farther to the right on the number line than (-1.88), we say that Kelly scored relatively higher among her classmates than John among his classmates.

Important Notes

- Kelly and John's z-scores are both **negative**. This means their exam scores were below the mean score.
- If the z-score is **positive**, it means the individual's value is above the mean**.**
- If the z-score is **zero**, it means the individual's value is exactly the mean.

Finding Probability Less Than a Z-score

Using the standard normal distribution and z table, the z-score can also tell us roughly the percentage of individuals in the group that was below a value (or someone's value). This can be done if the data is roughly normally distributed.

Example 6.1.5 Finding Probability Less Than z-score

Kelly recently took Exam #1 for her statistics class and scored 55 points. The class average for the exam was 60, with a standard deviation of 4. Kelly is curious about how her score compares to her classmates. Using the standard normal distribution, we determine that her z-score is -1.25.

Assuming the exam scores are normally distributed, approximately what percentage of students scored lower than Kelly? Additionally, what does this tell us about how she performed relative to the rest of the class?

Solution

In this case, we are trying to find $P(z < -1.25)$

We can find our answer by using a standard normal distribution and the related table.

We calculated Kelly's z-score earlier, which was -1.25. Since the value of Kelly's z-score is negative, it will be to the left of the center. We shade the area to the left side, z = -1.25.

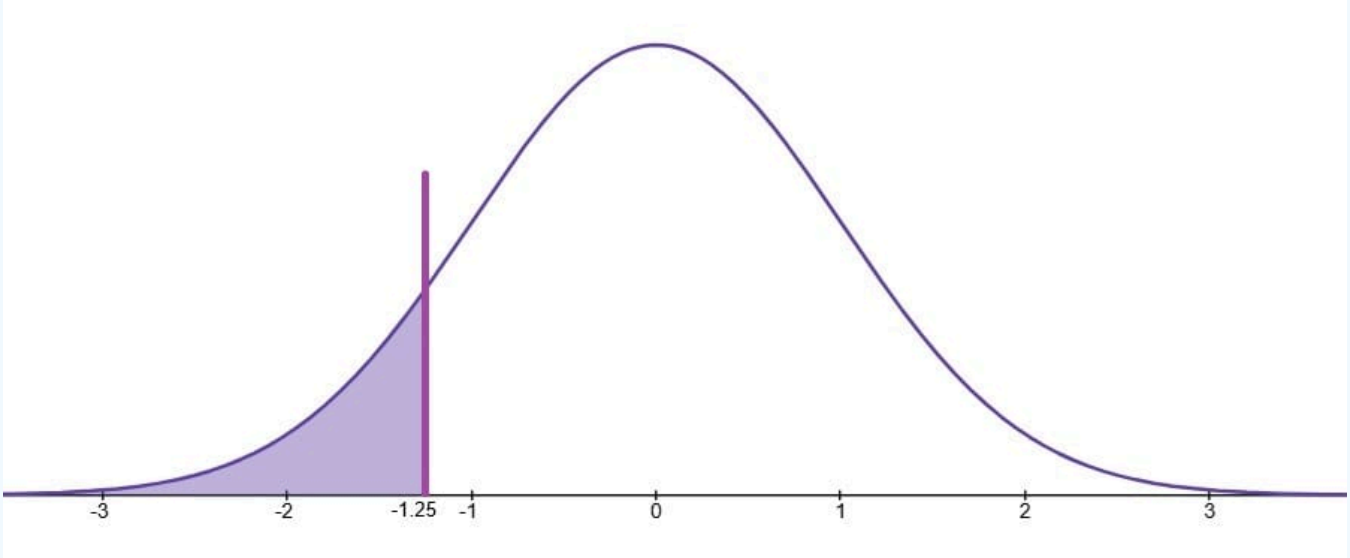


Figure 6.1.6: Standard Normal Distribution, Area Shaded to the Left of z=-1.25

Kelly's z-score = -1.25

To find the area or probability to the left of z = -1.25, we use the standard normal distribution table provided below. We go down -1.2 and move over to the right 0.05 to get the area of 0.1056. This area (probability) represents the area to the left of z=-1.25.

The Standard Normal Distribution (z-table)

z	0.00	0.01	0.02	.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Figure 6.1.7: Standard Normal Distribution Table. Describes How to Find Probability of $z < -1.25$

$$P(z < -1.25) \approx 0.1056 \approx 10.56$$

Thus, the probability of students who scored less than Kelly on the statistics exam is 0.1056 (10.56%).

Note: Computing the Area to the Left of a Z-value Using the TI-84+ Calculator

Steps to Enable Stat Wizard:

1. Press **[MODE]** to open the mode settings.
2. Use the **down arrow key** to scroll down until you see **[STAT WIZARDS]**.
3. Highlight **[STAT WIZARDS]** and use the **right arrow key** to toggle between **ON** and **OFF**.
4. Select **[ON]** and press **[ENTER]** to confirm.
5. Press **[2nd]** then **[QUIT]** to exit the mode settings.

Steps to Find the Area/Probability Using a TI-84+ Calculator:

1. Press the **[2nd]** button.
2. Press **[VARS]** to open the **[DISTR]** menu.
3. Select **[2: normalcdf(]** by either scrolling down and pressing **[ENTER]** or simply pressing **2**.
4. Enter the lower bound, which for a standard normal distribution extends to negative infinity. Since the calculator cannot input infinity, use **-E99** (which represents a very large negative number). To enter **-E99**, press **(-)** (the negative sign is not the same as the subtraction sign), then **[2nd]**, then **[EE]**, and finally type **99**.
5. Enter the upper bound as **-1.25** (the given z-score).
6. Enter the mean as **0** and the standard deviation as **1** for the standard normal distribution.
7. Press **[ENTER]** to compute the probability.

The calculator will display **0.1056.**

Finding Probability Greater Than a z-score

Using the standard normal distribution and z table, the z-score can also tell us roughly the percentage of individuals in the group that was above a value (or someone's value). This can be done as long as the data is roughly normally distributed.

Example 6.1.6 Finding Probability Greater Than z-score

Kelly scored 55 points on Exam #1 for her statistics class. The mean of the class was 60, and the standard deviation was 4. Determine approximately the percentage (or proportion) of students who scored above Kelly's result (z = - 1.25). Assume the exam scores for the class are normally distributed.

Solution

In this case, we are trying to find $P(z > -1.25)$.

The inequality (>) is pointing to the right, so we shade on the right side of our z-score. This shaded area will represent all the z-values greater than z = -1.25.

We can find our answer by using a standard normal distribution curve and a z-table. However, the standard normal distribution table gives the area to the left of -1.25.

Shade on the right side of -1.25 because we want to find the probability of the students who scored greater than Kelly's exam score.

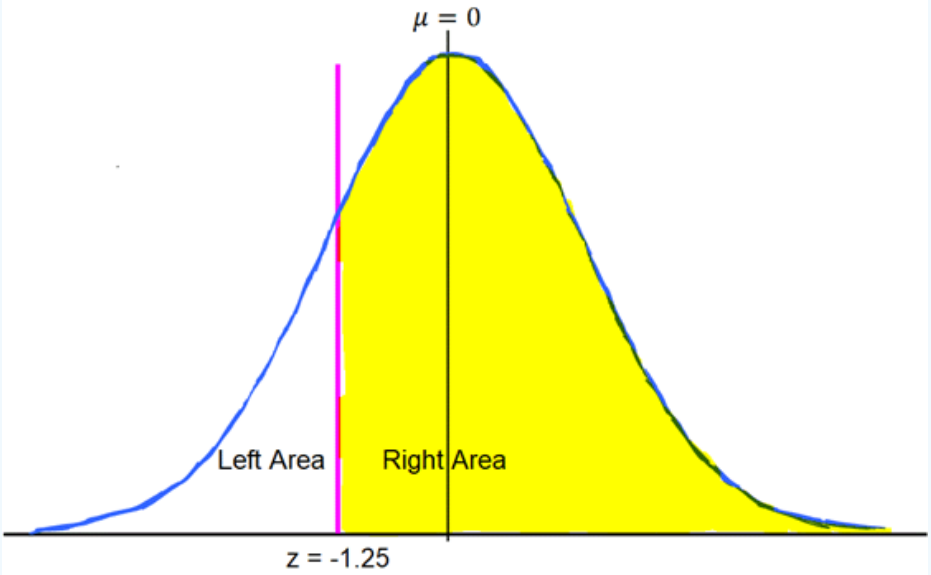


Figure 6.1.8: Standard Normal Distribution, Area Shaded to the Right of z=-1.25

Recall that the area (probability) under the normal distribution is 1 (100%). If we add the left area to the right area, we should get 1.

Left Area + Right Area = 1

In other words, if we subtract the whole area (in this case, it is 1) from the left area, we will be left with only the right area.

Right Area = 1 - Left Area

Kelly's z-score = -1.25

To find the left area, we go down -1.2 and move over to the right 0.05. The left area is 0.1056.

Right Area = 1 - 0.1056 = 0.8944

$$P(z > -1.25) = 0.8944 = 89.44\%$$

Thus, the probability of students who scored more than Kelly on the statistics exam is 0.8944 (89.44%).

The Standard Normal Distribution (z-table)

z	0.00	0.01	0.02	.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Figure 6.1.9: Standard Normal Distribution Table. Describes How to Find Probability of $z \leq -1.25$

Note: Computing the Area to the Right of a Z-Value Using the TI-84+ Calculator

1. Press the **[2nd]** button.
2. Press **[VARS]** to open the **[DISTR]** menu.
3. Select **[2: normalcdf(]** by either scrolling down and pressing **[ENTER]** or simply pressing **2**.
4. Enter the lower bound **-1.25** (the given z-score).
5. Enter the upper bound, which for a standard normal distribution extends to positive infinity. Since the calculator cannot input infinity, use **E99** (which represents a very large negative number). To enter **E99**, press **[2nd]**, then **[EE]**, and finally type **99**.
6. Enter the mean as **0** and the standard deviation as **1** for the standard normal distribution.
7. Press **[ENTER]** to compute the probability.

The calculator will display 0.8944.

Finding Probability Between Two z-scores

Using the standard normal distribution and z-table, the z-score can also tell us roughly the percentage of individuals in the group that were between two values. This can be done as long as the data is roughly normally distributed.

Example 6.1.7

For Exam #1 in a statistics class at Citrus College, students' scores follow a normal distribution. The professor wants to analyze how many students performed within a specific range compared to the class average. Using the standard normal distribution, determine the percentage of students whose z-scores fall between z = -1.53 and z = 2.17.

Solution

In this case, we are trying to find \(P\left(-1.53

We know we need to shade the middle because it asks to find the area (probability) between two z values (z = -1.53 and z = 2.17).

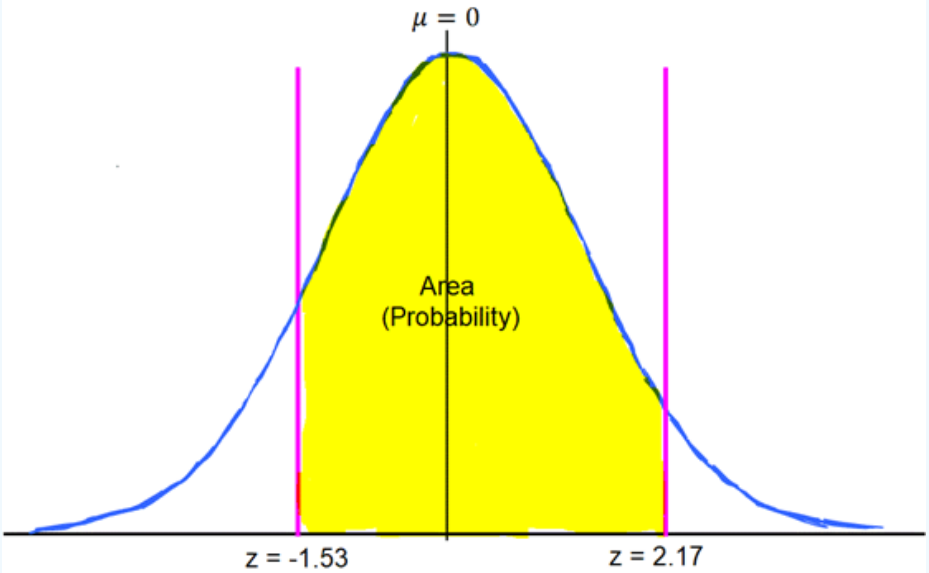


Figure 6.1.10: Standard normal distribution, area shaded between z=-1.53 and z=2.17

We can find our probability to the left of each z-score by using the standard normal distribution table. If we subtract the lower probability value from the higher probability value, we will find the area (probability) between z = -1.53 and z = 2.17.

$$P\left(-1.53 < z < 2.17\right) = p\left(z < 2.17\right) - p\left(z < -1.53\right)< p=""> </z<2.17\right)=p\left(z<2.17\right)-p\left(z<-1.53\right)\)<>$$

To find $P\left(z < 2.17\right)$ and $P\left(z < -1.53\right)$ we use the z-table.

z = 2.17

To find the area or probability to the left of z = 2.17, we use the standard normal distribution table provided below. We go down 2.1 and move over to the right 0.07 to get the area of 0.9850. This area (probability) represents the area to the left of z=2.17.

The Standard Normal Distribution (z-table)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

Figure 6.1.11: Standard Normal Distribution Table. Describes How to Find Probability of $z < 2.17$

$$P\left(z < 2.17\right) = 0.9850$$

z = -1.53

To find the area or probability to the left of z = -1.53, we use the standard normal distribution table provided below. We go down -1.5 and move over to the right 0.03 to get the area of 0.0630. This area (probability) represents the area to the left of z=-1.53.

The Standard Normal Distribution (z-table)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	**0.0630**	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Figure 6.1.12: Standard Normal Distribution Table. Describes How to Find Probability of $z < -1.53$

$$P\left(z < -1.53\right) = 0.0630$$

Now that we found $P\left(z < 2.17\right) = 0.9850$ and $P\left(z < -1.53\right) = 0.0630$, we can use the following formula:

$$P\left(-1.53 < z < 2.17\right) = p\left(z < 2.17\right) - p\left(z < -1.53\right)< p=""> </z<2.17\right)=p\left(z<2.17\right)-p\left(z<-1.53\right)\)<>$$

$$= 0.9850 - 0.0630$$

$$= 0.9220$$

Thus, the probability of students with z-scores between -1.53 and 2.17 is 0.9220 (92.20%).

Note: Computing the Area Between Two Z-Values Using the T-84+ Calculator

1. Press the **[2nd]** button.
2. Press **[VARS]** to open the **[DISTR]** menu.
3. Select **[2: normalcdf(]** by either scrolling down and pressing **[ENTER]** or simply pressing **2**.
4. Enter the lower bound **-1.53** (the given z-score).
5. Enter the upper bound **2.17** (the given z-score).
6. Enter the mean as **0** and the standard deviation as **1** for the standard normal distribution.
7. Press **[ENTER]** to compute the probability.

The calculator will display 0.9220.

Summary: Finding Probability When the z-score is Known

1. To find the area (probability) ***left*** of a z-score (fence), use the z-table.

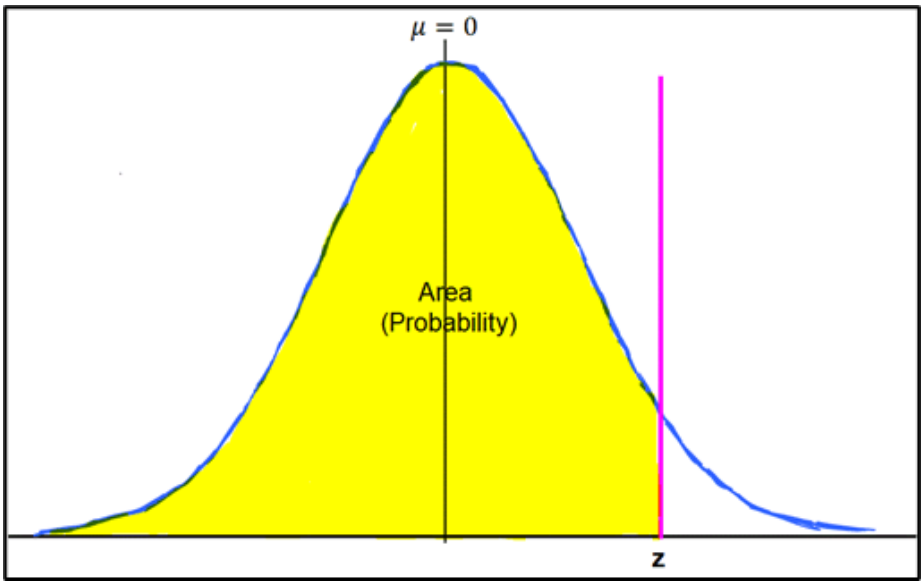


Figure 6.1.13: Standard Normal Distribution Curve Where the Area to the Left of z-score is Shaded.

2. To find the area (probability) on the **right** of the z-score (fence), we perform the following expression **1 - (value obtained from the z table)**

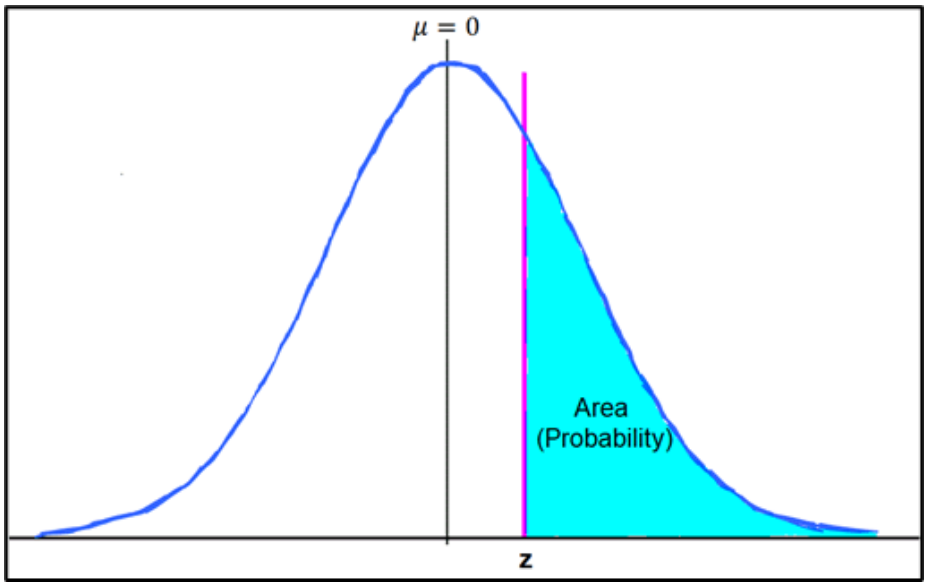


Figure 6.1.14: Standard Normal Distribution Curve Where the Area to the Right of z-score is Shaded.

3. To find the area (probability) ***between*** two z-scores (fences), we subtract the corresponding areas (take the higher area and subtract the lower area).

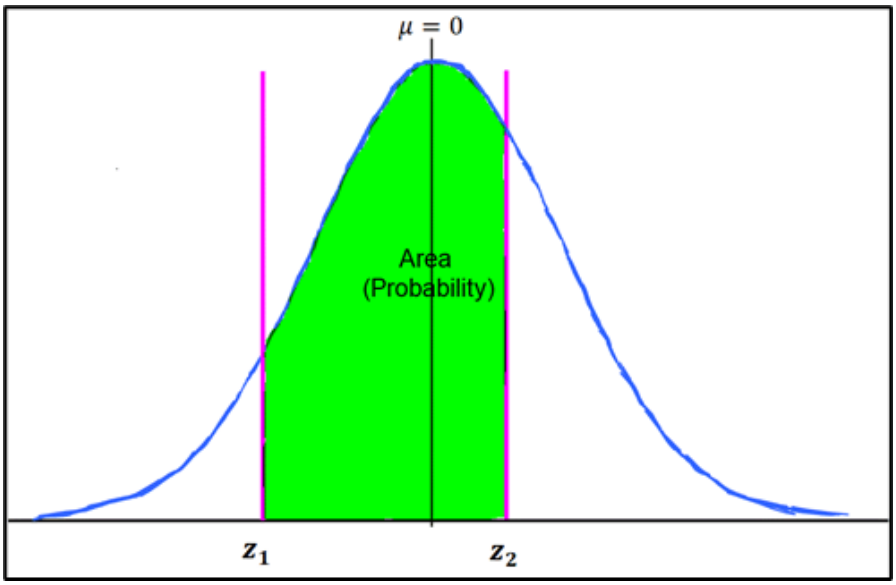


Figure 6.1.15: Standard Normal Distribution Curve Where the Area is Between Two z-values

Finding z-score When the Probability is Known

Finding z-score of Probability to the Left

We can also find the z-score (fence) using the z-table if we know the area (probability) to the left of the z-score.

Example 6.1.8 Finding z-score of Probability to the Left

Find the z-score based on the information provided below.

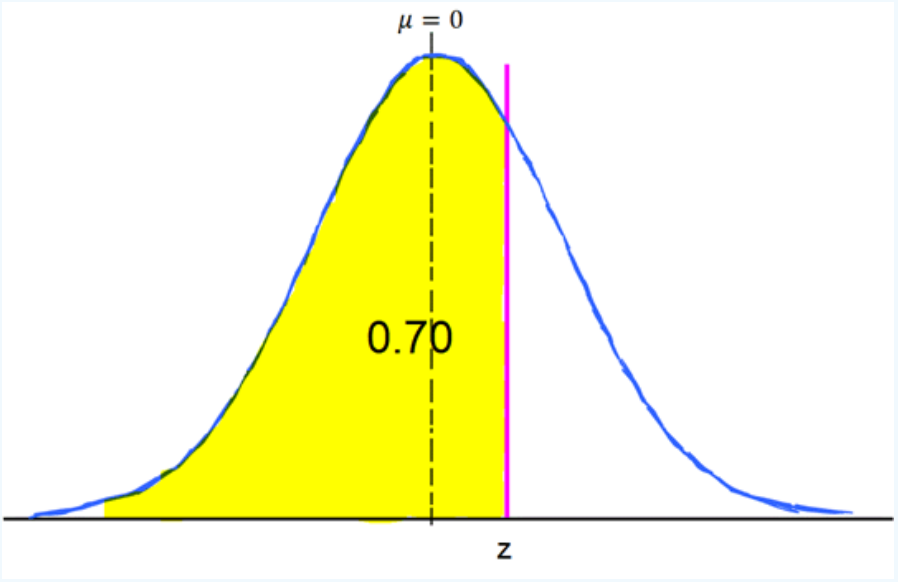


Figure 6.1.16: Standard Normal Distribution Curve where the Area to the Left of the z-score is 0.70

Solution

Based on the diagram above, we have our area (probability) on the left of our z-score (fence). In this case, the area (probability) is 0.70. We can use the z-table to find our z-score.

To find our z-score, we start from the area (probability) region and identify the row and column of the closest number to 0.70. In this case, it will be **0.6985**. The row for this value will be 0.5, and the column will be 0.02. If we add the row and column, we get our z-score value of 0.52.

The Standard Normal Distribution (z-Table)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

Figure 6.1.17: Explains How to Find the Z-score from a Z-Table where the Probability (Area) to the Left is Given

$$z = 0.52$$

For this example, our z-score (fence) will be 0.52.

Finding z-score of Probability to the Right

We can only find the z-score using the z-table if we know the area (probability) on the left of the z-score. If the right area is given, we first need to find the Left Area before we can use the z-table.

Example 6.1.9 Finding z-score of Probability to the Right

Find the z-score based on the information provided below.

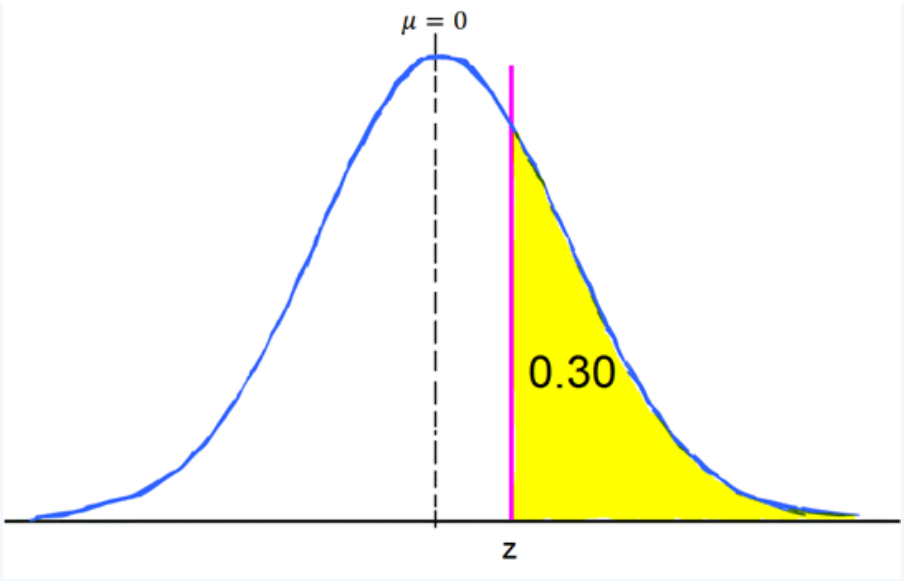


Figure 6.1.18: Standard Normal Distribution Curve Where the Area to the Right of z-score is 0.30

Solution

Recall that the area under the standard normal distribution is 1 (100%). If we add the **Left Area** with the **Right Area** that should give us the whole area (in this case it will be 1).

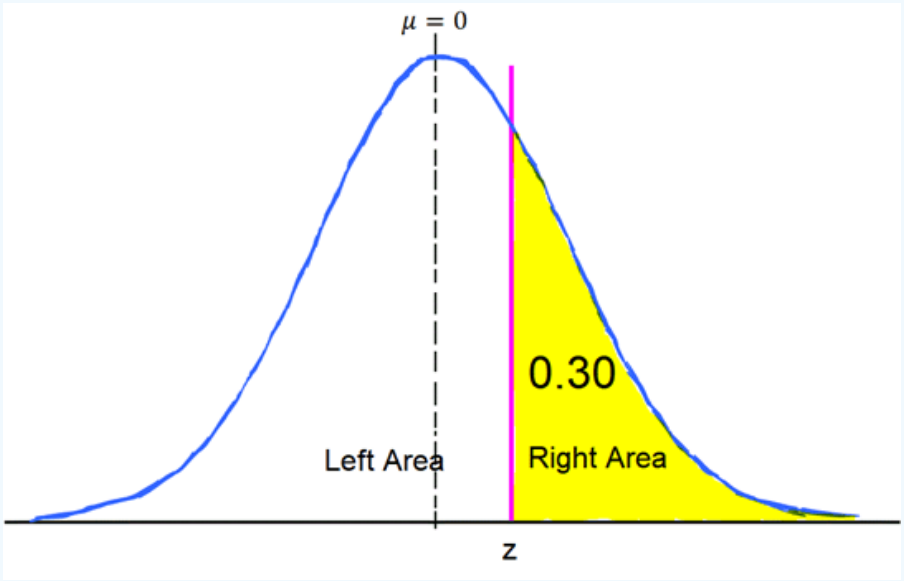


Figure 6.1.19: Standard Normal Distribution Curve Where the Left and Right Area are Separated by the z-score.

$$\text{Left Area} + \text{Right Area} = 1$$

In other words, if we subtract the **Whole Area** (in this case, it is 1) from the **Right Area**, what will remain will be the **Left Area.**

$$\text{Left Area} = 1 - \text{Right Area}$$

$$\text{Left Area} = 1 - 0.30 = 0.70$$

To find our z-score, we start from the area (probability) region and identify the row and column of the closest number to 0.70. In this case, it will be **0.6985**. The row for this value will be 0.5, and the column will be 0.02. If we add the row and column we get our z-score value of 0.52.

The Standard Normal Distribution (z-Table)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.5	~~0.6915~~	~~0.6950~~	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

Figure 6.1.20: Explains How to Find the z-score From a z-Table Where the Probability (Area) to the Left is Given

$$z = 0.52$$

For this example, our z-score (fence) will be 0.52.

Finding the z-scores of the Area (Probability) Between Them

Sometimes we will have a question where that asks us to find two z-scores when the area (probability) of the middle is given. The same rule applies to our previous two examples. We can use the z-table to find our z-scores if we know the area (probability) to the left of each z-score.

Example 6.1.9 Finding the z-scores of the Area (Probability) Between Them

Find the z-scores based on the information provided below.

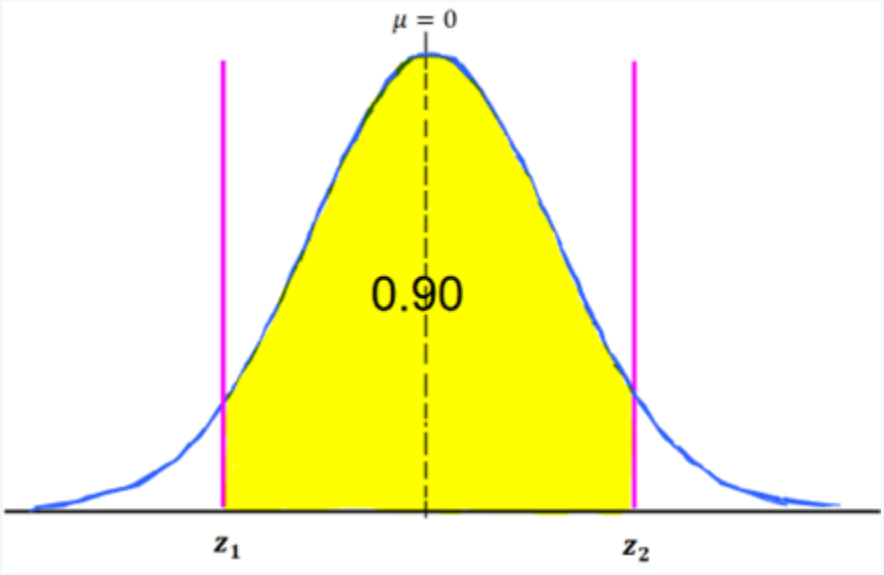


Figure 6.1.21: Standard Normal Distribution Curve Where the Middle Area is 0.90.

Solution

We first need to find the area of all the regions before we can find z_1 and z_2.

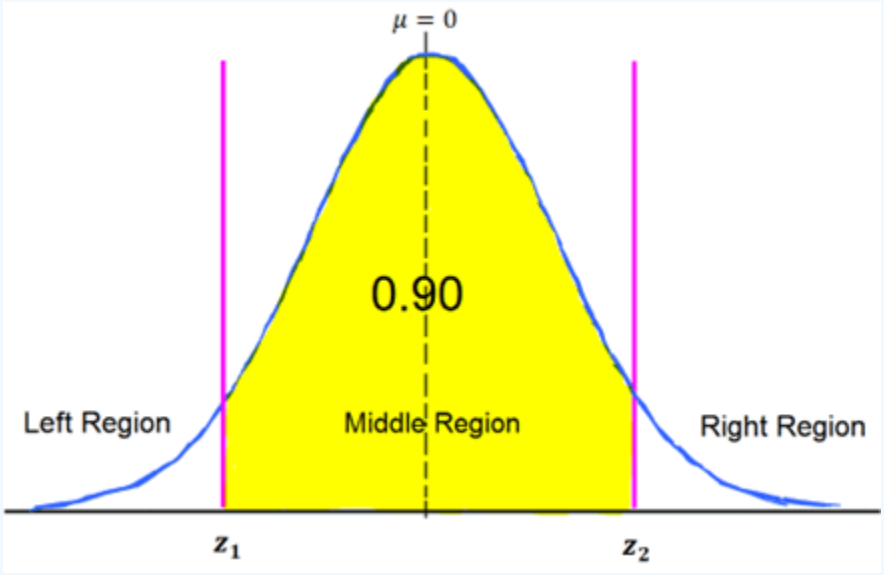


Figure 6.1.22: Standard Normal Distribution Curve Where the Left, Middle, and Right Region are Separated by Our z-score Fences.

Recall area under the standard normal distribution is 1. If we subtract the **Whole Area** (in this case, 1) from the **Middle Region**, then we are left with the **Outside Region (**which we will eventually call).

Outside Region = Whole Area - Middle Region

Outside Region = 1 - 0.90 = 0.10

Note: Outside Region = Left Region + Right Region = 0.10

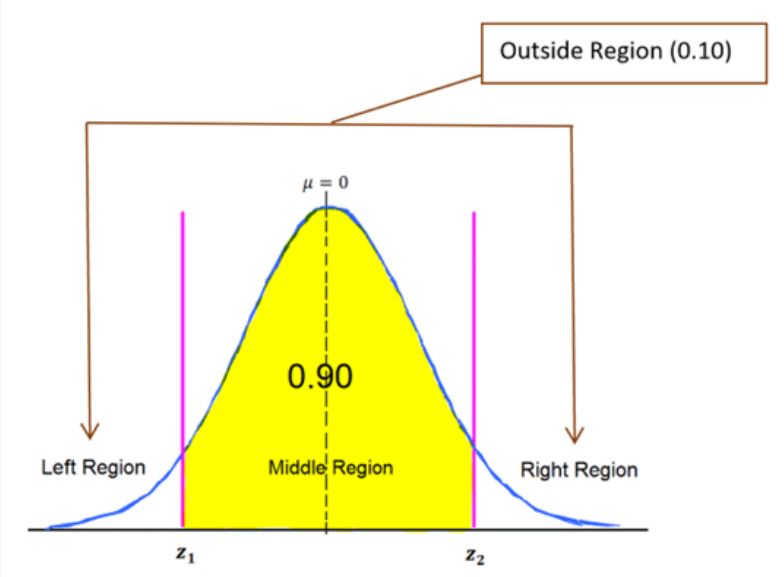


Figure 6.1.23: Normal Distribution Curve that Shows the Middle Region is 0.90 and the Outside Region (Left and Right Region Together) is 0.10.

Authors

"6.1: Introduction to the Normal Distribution" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"6.2: Graphs of the Normal Distribution" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 6.1: Introduction to the Normal Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 6.2: Graphs of the Normal Distribution by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

6.2: Finding Probabilities for the Normal Distribution

Learning Objectives

- Convert values to z-scores to compute probabilities using the normal distribution.
- Use tables or calculators to find the area under the normal curve representing the likelihood of a value occurring.
- Apply the reverse process to determine cut-off scores corresponding to given probabilities.

Finding the Probability of a Normal Distribution

The Empirical Rule is just an approximation and only works for certain values. What if you want to find the probability for *x* values that are not integer multiples of the standard deviation? The probability is the area under the curve. To find areas under the curve, you need calculus. Before technology, you needed to convert every *x* value to a standardized number, called the *z*-score or *z*-value, or simply just *z*. The *z*-score is a measure of how many standard deviations an *x* value is from the mean. To convert from a normally distributed *x* value to a *z*-score, you use the following formula.

Definition 6.2.1: z-score

$$z = \frac{x - \mu}{\sigma} \tag{6.2.1}$$

where μ= mean of the population of the x value and σ= standard deviation for the population of the x value

The z-score is normally distributed, with a mean of 0 and a standard deviation of 1. It is known as the standard normal curve. Once you have the z-score, you can look up the z-score in the standard normal distribution table.

Definition 6.2.2: standard normal distribution

The **standard normal distribution**, z, has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$.

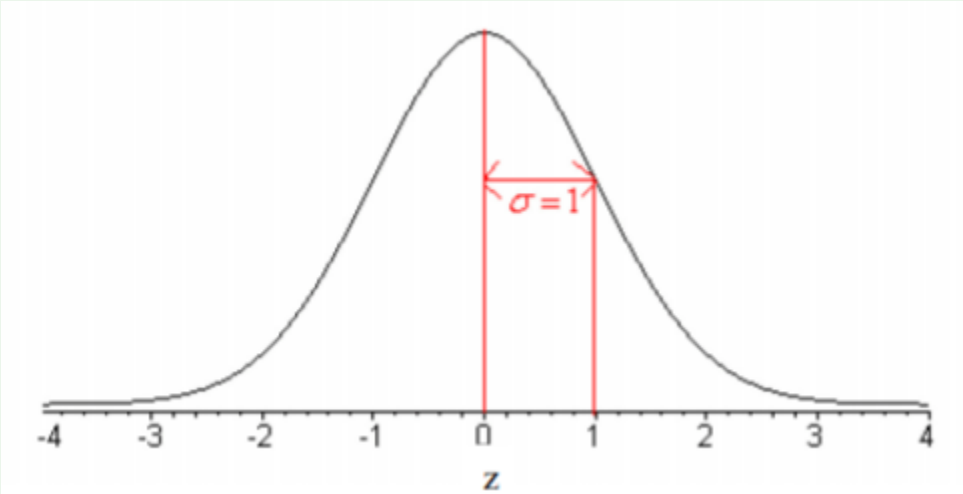


Figure 6.2.1: Standard Normal Curve

Luckily, these days technology can find probabilities for you without converting to the z-score and looking the probabilities up in a table. There are many programs available that will calculate the probability for a normal curve, including Excel and the TI-83/84. There are also online sites available. The following examples show how to do the calculation on the TI-83/84. The command on the TI-83/84 is in the DISTR menu and is normalcdf(. You then type in the lower limit, upper limit, mean, and standard deviation in that order, and including the commas.

Example 6.2.1 general normal distribution

The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9 days (Bhat & Kushtagi, 2006).

Click on blue links to access the normal distribution tables for negative z-values and positive z-values. After the value is found, press "back" to return to the page.

a. State the random variable.
b. Find the probability that a pregnancy lasts more than 280 days.
c. Find the probability that a pregnancy lasts less than 250 days.
d. Find the probability that a pregnancy lasts between 265 and 280 days.
e. Find the cutoff length of pregnancy such that the shortest 10% of all pregnancies fall below this value.
f. Suppose you meet a woman who says that she was pregnant for less than 250 days. Would this be unusual, and what might you think?

Solution

a. x = length of a human pregnancy

b. First, translate the statement into a mathematical statement.

$$P(x>280)$$

Now, draw a picture. Remember that the center of this normal curve is 272.

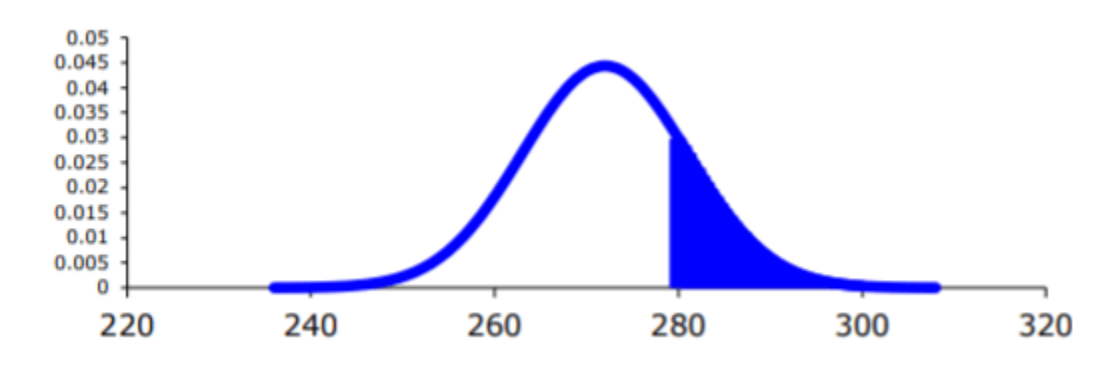


Figure 6.2.2: Normal Distribution With Shaded Area to the Right of 280.

To find the probability on the TI-83/84, looking at the picture, you realize the lower limit is 280. The upper limit is infinity. The calculator doesn't have infinity on it, so you need to put in a really big number. Some people like to put in 1000, but if you are working with numbers that are bigger than 1000, then you would have to remember to change the upper limit. The safest number to use is 1×10^{99}, which you put in the calculator as 1E99 (where E is the EE button on the calculator). The command looks like this:

$$\text{normalcdf}(280, 1E99, 272, 9)$$

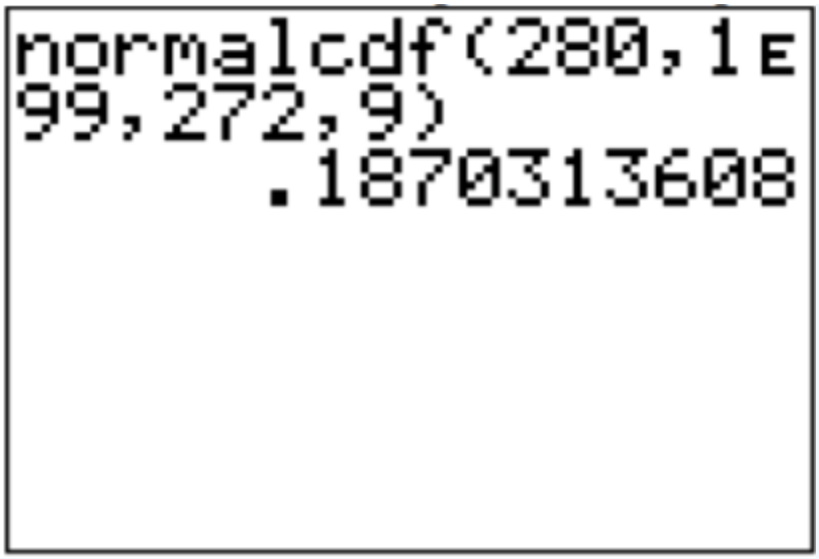


Figure 6.2.3: TI-83/84 Output of Area to the Right of 280.

Thus, $P(x > 280) \approx 0.187$

Thus 18.7% of all pregnancies last more than 280 days. This is not unusual since the probability is greater than 5%.

To find the probability using a z-table, we need to first compute the z-score and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of a pregnancy lasting more than 280 days, so x our random variable x will be 280 days.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{280 - 272}{9}$$

$$z = \frac{8}{9}$$

$z = 0.89$ (rounded to two decimal places)

As a result, *P* (x >280) = P(z > 0.89). Since z is more than 0.89, our shaded area will be on the right of z = 0.89.

We can use the z-table to find our probability to the left. If we convert the normal distribution to the standard normal distribution, we get the following:

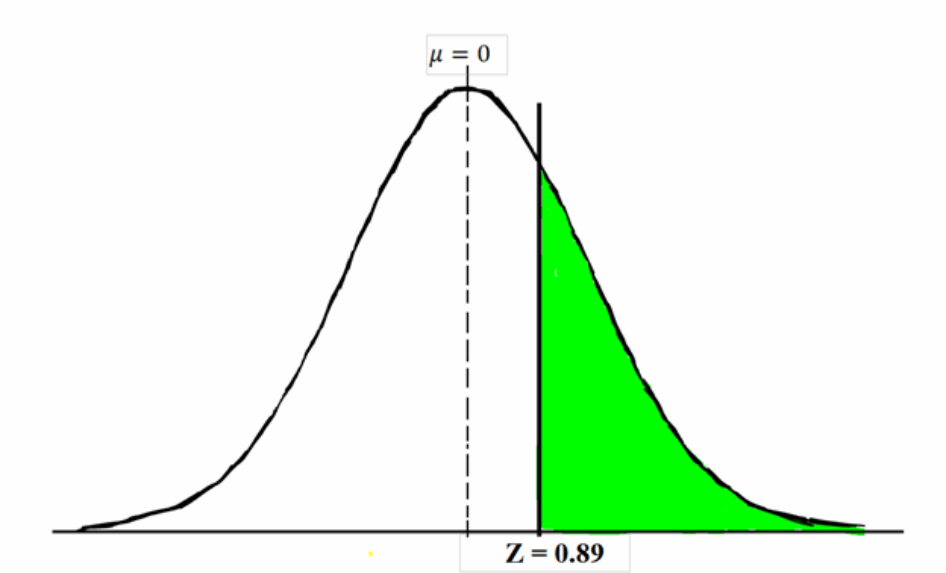


Figure 6.2.4: Normal Distribution, Z = 0.89, Shaded on the Right

From the z-table, it only gives us the probability to the left of z-score. Since we know the total area of the normal distribution is 1. To get the area to the right, we need to take 1 minus the area to the left. The work is provided below.

$$P(z > 0.89) = 1 - P(z < 0.89)$$

To find the area or probability to the left of z = 0.89, we use the standard normal distribution table provided below. We go down 0.8 and move over to the right 0.09 to get the area of 0.8133. This area (probability) represents the area to the left of z=0.89.

Table 6.2.1: Normal Distribution Table For Positive Z-values

Based on the z-table above, the Left Area will be 0.8133.

Right Area = 1 - Left Area

Right Area = 1 - 0.8133 = 0.1867

$$P(z>0.89)= 0.1867 = 18.67\%$$

Using the table, P(x >280) = 18.67%. This means that the probability that a pregnancy will last more than 280 days is 18.67%.

The answer using the z-table is slightly off due to rounding our z-value to two decimal places. The calculator uses more decimal places for z, hence it has a more accurate answer.

c. First, translate the statement into a mathematical statement.

$$P\ (x<250)$$

Now, draw a picture. Remember that the center of this normal curve is 272.

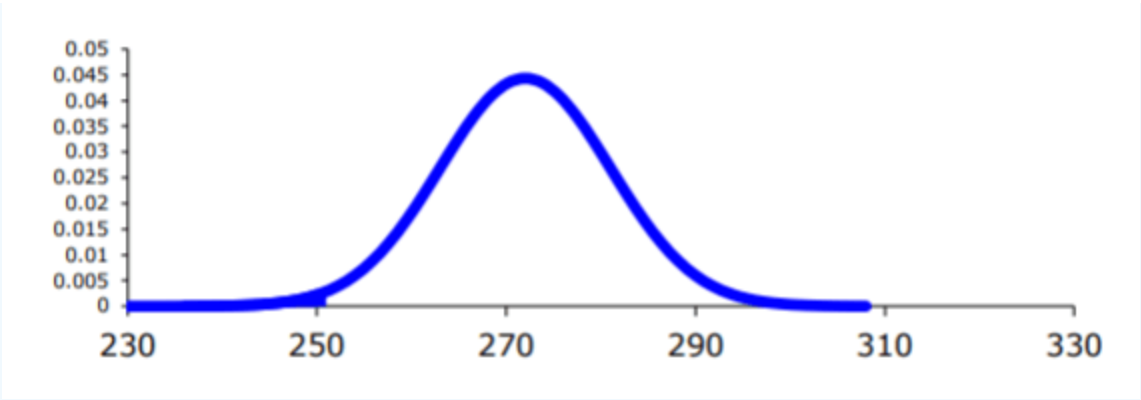


Figure 6.2.5: Graph of Normal Distribution with Area to the Left of 250.

To find the probability on the TI-83/84, looking at the picture, though it is hard to see in this case, the lower limit is negative infinity. Again, the calculator doesn't have this on it, put in a really small number, such as $-1 \times 10^{99} = -1E99$ on the calculator.

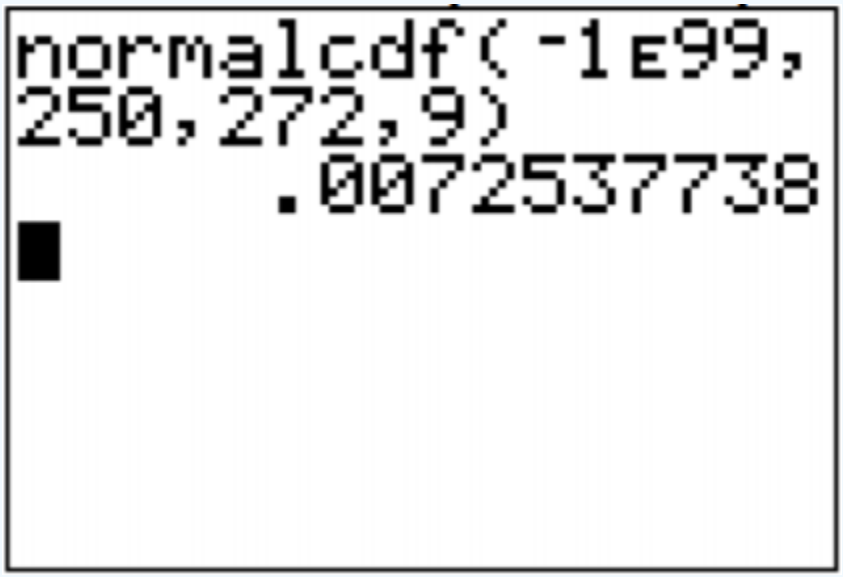


Figure 6.2.6: TI-83/84 Output for Area to the Left of 250.

$$P(x < 250) = \text{ normalcdf } (-1E99, 250, 272, 9) = 0.0073$$

Thus, 0.73% of all pregnancies last less than 250 days. This is unusual since the probability is less than 5%.

To find the probability using a z-table, we need to first compute the z-score and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of a pregnancy lasting less than 250 days, so x our random variable x will be 250 days.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{250 - 272}{9}$$

$$z = \frac{-22}{9}$$

$z = -2.44$ (rounded to two decimal places)

As a result, *P* (x < 250) = P(z < -2.44). Since z is less than -2.44, our shaded area will be on the left of z = -2.44.

We can use the z-table to find our probability to the left. If we convert the normal distribution to the standard normal distribution, we get the following:

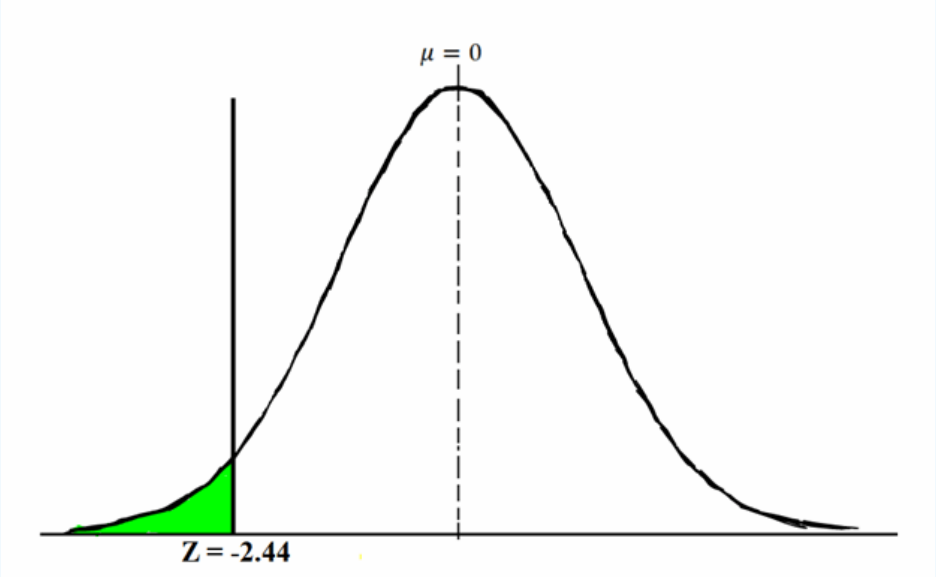


Figure 6.2.7: Normal Distribution, Z = -2.44, Shaded on the Left

We use the standard normal distribution table provided below to find the area or probability to the left of z = -2.44. We go down -2.4 and move over to the right 0.04 to get the area of 0.0073. This area (probability) represents the area to the left of z=-2.44.

Table 6.2.2: Normal Distribution Table For Negative Z-values

Based on the z-table above, the Left Area will be 0.0073, which is the probability that z is less than -2.44. If we express the answer as a percentage, we get the following answer:

$$P(z < -2.44) = 0.73\%$$

Using the table, P(x <250) = 0.73%. This means the probability that a pregnancy will last less than 250 days is 0.73%.

The answer using the z-table is slightly off due to rounding our z-value to two decimal places. The calculator uses more decimal places for z, hence it has a more accurate answer.

d. First translate the statement into a mathematical statement.

$$P (265$$

Now, draw a picture. Remember that the center of this normal curve is 272.

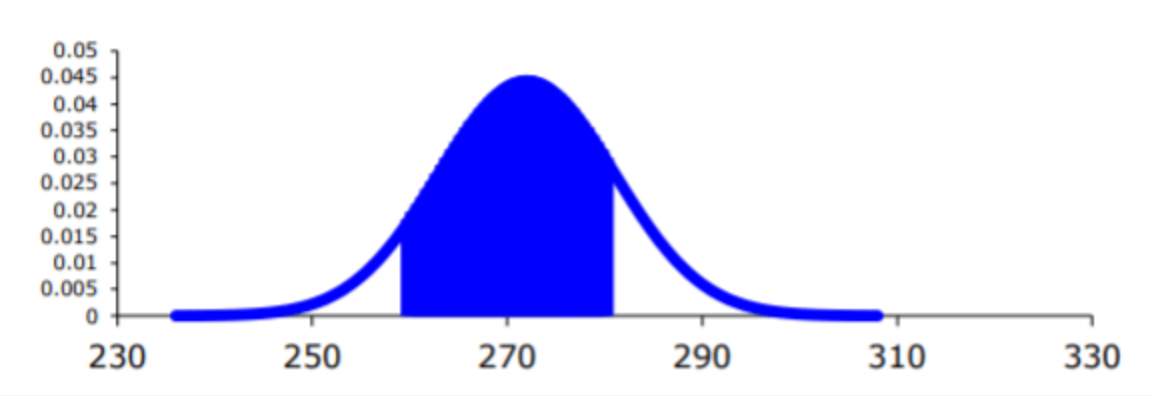


Figure 6.2.8: Normal Distribution with Area Between 265 and 280.

In this case, the lower limit is 265 and the upper limit is 280.

Using the calculator

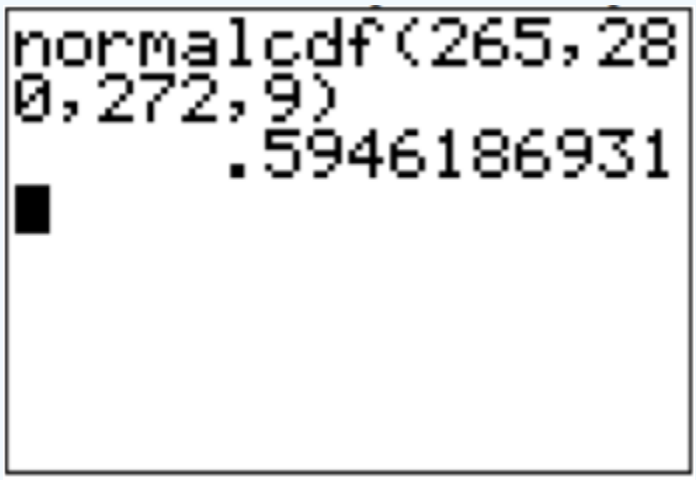


Figure 6.2.9: TI-83/84 Output for Area Between 265 and 280.

$$P(265 < x < 280) = =\text{"" normalcdf=""} (265, 280, 272, 9) = \prime\prime 0.595\text{</p"> </x<280)=\text}$$

Thus 59.5% of all pregnancies last between 265 and 280 days.

To find the probability using a z-table, we need to first compute the z-scores and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of a pregnancy lasting between 265 and 280 days, so x our random variable will be 265 days and 280 days. For each x-value, we need to find their corresponding z-value.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{265 - 272}{9}$$

$$z = \frac{-7}{9}$$

$z = -0.78$ (rounded to two decimal places)

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{280 - 272}{9}$$

$$z = \frac{8}{9}$$

$z = 0.89$ (rounded to two decimal places)

As a result, P(265 < x < 280) = P(-0.78 < z < 0.89).

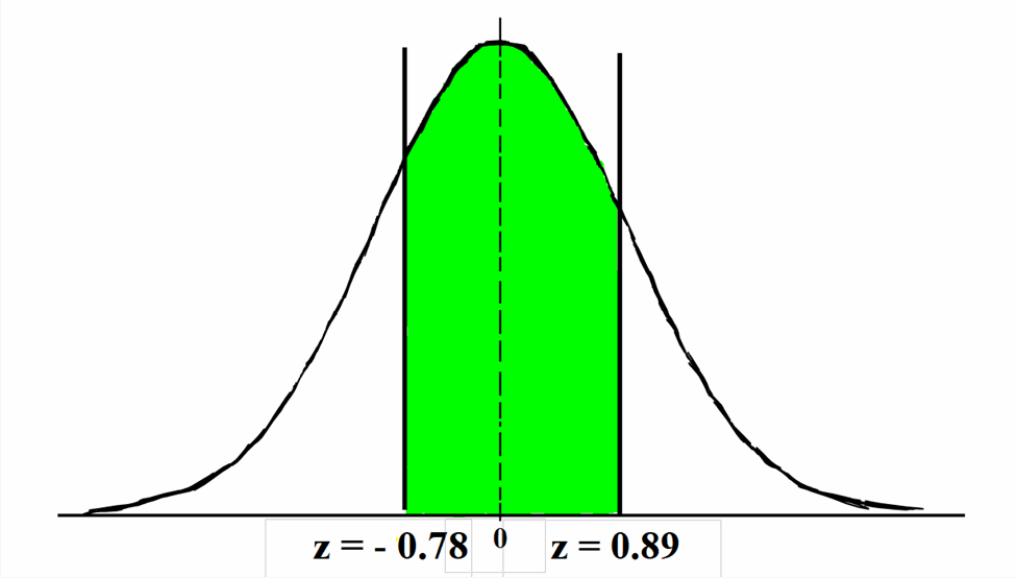


Figure 6.2.10: Normal Distribution Where Shaded Area is Between z = -0.78 and z = 0.89

We can find our probability to the left of each z-score by using the z-table. If we subtract the higher probability value with the lower probability value, we will find the area (probability) between z = -0.78 and z = 0.89.

$$P\left(-0.78 < z < 0.89\right) = p\left(z < 0.89\right) - p\left(z < -0.78\right)$$< p=""> </z<0.89\right)=p\left(z<0.89\right)-p\left(z<-0.78\right)\)<>

To find $P\left(z < 0.89\right)$ and $P\left(z < -0.78\right)$ we use the z-table.

z = 0.89

To find the area or probability to the left of z = 0.89, we use the standard normal distribution table provided below. We go down 0.8 and move over to the right 0.09 to get the area of 0.8133. This area (probability) represents the area to the left of z=0.89.

$$P\left(z < 0.89\right) = 0.8133$$

Table 6.2.1: Normal Distribution Table For Positive Z-values

z = -0.78

To find the area or probability to the left of z = -0.78, we use the standard normal distribution table provided below. We go down -0.7 and move over to the right 0.08 to get the area of 0.2177. This area (probability) represents the area to the left of z=-0.78.

$$P\left(z < -0.78\right) = 0.2177$$

Table 6.2.2: Normal Distribution Table For Negative Z-values

Now that we found $P\left(z < 0.89\right) = 0.8133$ and $P\left(z < -0.78\right) = 0.2177$, we can use the following formula:

$$P(265 < x < 280) = p(-0.78 < z < 0.89) = p(z < 0.89) - p(z < -0.78)$$< p=""> </x<280\right)=p\left(-0.78<z<0.89\right)=p\left(z<0.89\right)-p\left(z<-0.78\right)\)<>

$$P(265 < x < 280) = p(-0.78 < z < 0.89) = 0.8133 - 0.2177$$< p=""> </x<280\right)=p\left(-0.78<z<0.89\right)=0.8133-0.2177\)<>

$$P(265 < x < 280) = p(-0.78 < z < 0.89) = 0.5956$$< p=""> </x<280\right)=p\left(-0.78<z<0.89\right)=0.5956\)<>

Thus, the probability of a pregnancy lasting between 265 and 280 days will be 59.56%.

e. This problem asks you to find an *x* value from a probability. You want to find the *x* value that has 10% of the length of pregnancies to the left of it. On the TI-83/84, the command is in the DISTR menu and is called invNorm(. The invNorm(command needs the area to the left. In this case, that is the area you are given. For the command on the calculator, once you have invNorm(on the main screen you type in the probability to the left, mean, standard deviation, in that order with the commas.

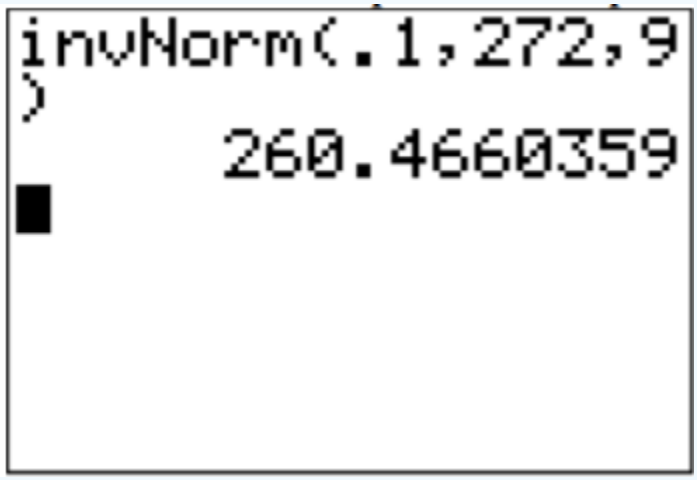


Figure 6.2.11: TI-83/84 Output for Inverse Norm Function

Thus, 10% of all pregnancies last less than approximately 260 days.

f. From part (c), you found the probability that a pregnancy lasts less than 250 days is 0.73%. Since this is less than 5%, it is very unusual. You would think that either the woman had a premature baby, or that she may be wrong about when she became pregnant.

Example 6.2.2 general normal distribution

The mean mathematics SAT score in 2012 was 514 with a standard deviation of 117 ("Total group profile," 2012). Assume the mathematics SAT score is normally distributed.

Click on blue links to access the normal distribution tables for negative z-values and positive z-values. After the value is found, press "back" to return to the page.

a. State the random variable.
b. Find the probability that a person has a mathematics SAT score over 700.
c. Find the probability that a person has a mathematics SAT score of less than 400.
d. Find the probability that a person has a mathematics SAT score between 500 and 650.
e. Find the mathematics SAT score that represents the top 1% of all scores.

Solution

a. *x* = mathematics SAT score

b. First, translate the statement into a mathematical statement.

$$P(x>700)$$

Now, draw a picture. Remember that the center of this normal curve is 514.

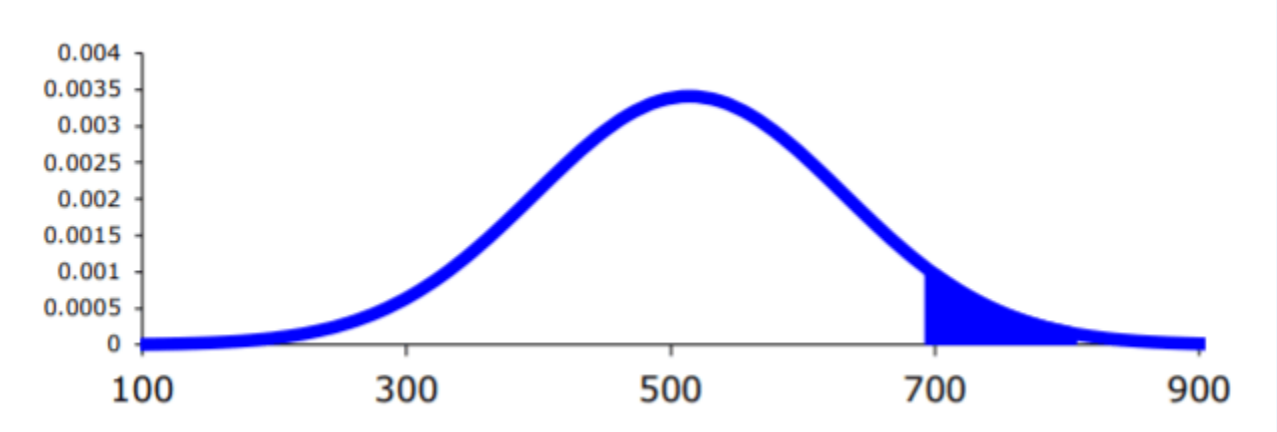


Figure 6.2.12: Normal Distribution with Area to the Right of 700

On TI-83/84: $P(x > 700) =$ normalcdf $(700, 1E99, 514, 117) \approx 0.056$

There is a 5.6% chance that a person scored above 700 on the mathematics SAT. This is not unusual.

c. First, translate the statement into a mathematical statement.

$$P(x<400)$$

Now, draw a picture. Remember that the center of this normal curve is 514.

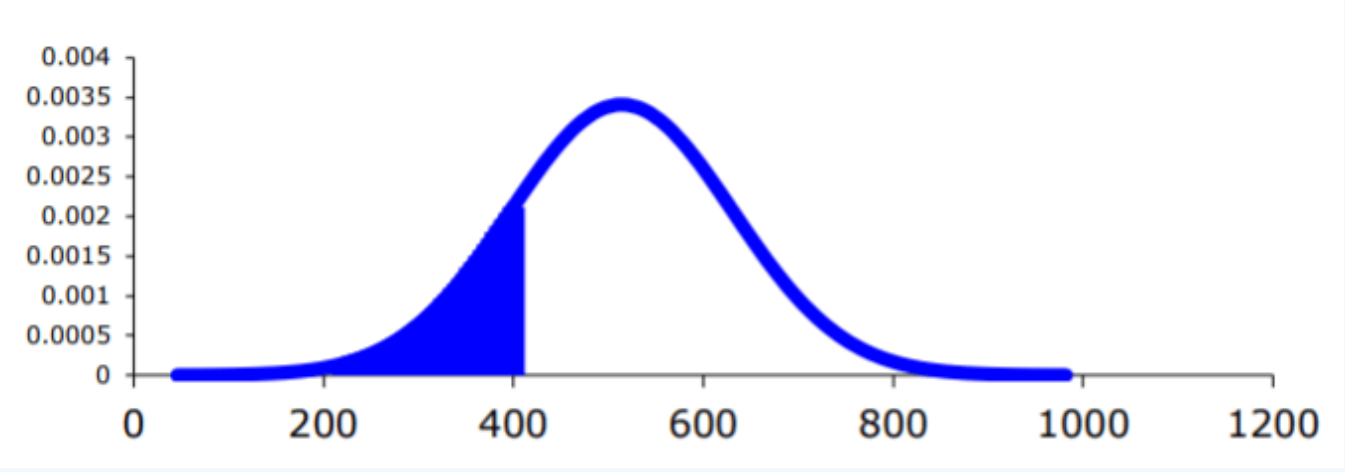


Figure 6.2.13: Normal Distribution with Area to the Left of 400.

On TI-83/84: $P(x < 400) =$ normalcdf $(-1E99, 400, 514, 117) \approx 0.165$

So, there is a 16.5% chance that a person scores less than 400 on the mathematics part of the SAT.

d. First translate the statement into a mathematical statement.

$$P(500$$

Now, draw a picture. Remember that the center of this normal curve is 514.

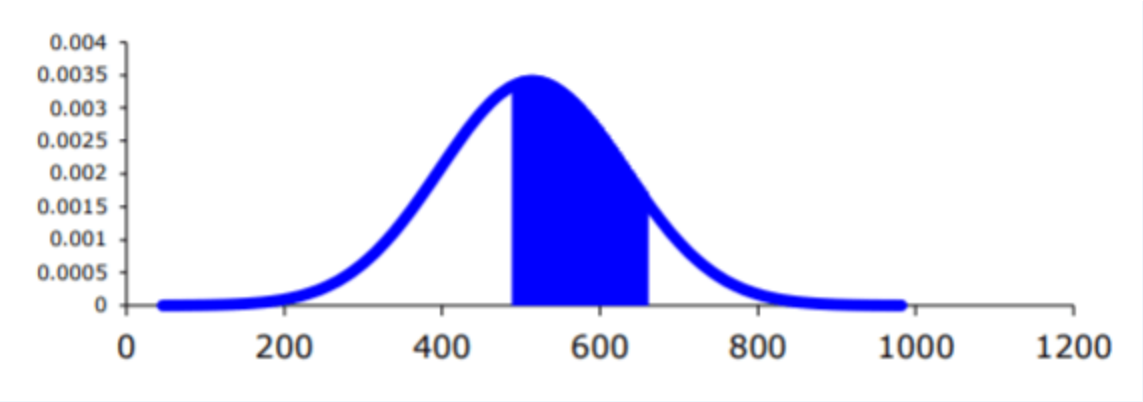


Figure 6.2.14: Normal Distribution of Area Between 500 and 650.

On TI-83/84: $P(500 < x < 650) =$"" normalcdf="" $(500, 650, 514, 117) =$ $''$ $''$ $\approx$= $''$ $''0.425$<="" p=""> </x<650)=\text>

So, there is a 42.5% chance that a person has a mathematical SAT score between 500 and 650.

e. This problem asks you to find an x value from a probability. You want to find the *x* value that has 1% of the mathematics SAT scores to the right of it. Remember, the calculator and R always need the area to the left; you need to find the area to the left by 1 - 0.01 = 0.99.

On TI-83/84: $\text{invNorm}(.99, 514, 117) \approx 786$

So, 1% of all people who took the SAT scored over 786 points on the mathematics SAT.

Finding the Minimum and Maximum of a Random Variable for a Normal Distribution

When working with a normal distribution, if we want to know the minimum value to be part of the top range, the maximum value to be in the bottom range, or the minimum or maximum value for the middle range, then we can use the invNorm() function to obtain the answers.

Example 6.2.1 Find the Minimum and Maximum of the Random Variable

The ages of students at Citrus College are normally distributed with a mean of 24 years and a standard deviation of 3 years.

Click on blue links to access the normal distribution tables for negative z-values and positive z-values. After the value is found, press "back" to return to the page.

a. What is the maximum age a student can be to fall in the bottom 20% of the age distribution at Citrus College?
b. What is the minimum age a student must be to fall in the top 20% of the age distribution at Citrus College?
c. What are the minimum and maximum ages that encompass the middle 50% of the age distribution at Citrus College?

Solution

a. To find the maximum age to be part of the bottom 20%, we draw a bell-shaped curve and shade the bottom 20% of the graph.

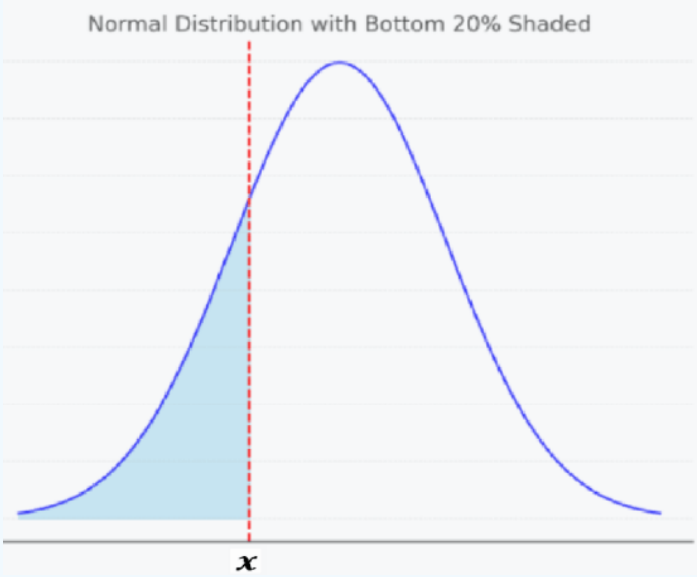


Figure 6.2.15: Normal distribution with the bottom 20% shaded, where the fence is the maximum value.

Calculating z Using invNorm() function:

To find the z-value with a graphing calculator, use the invNorm() function.

1. Press the **2nd** button.
2. Press the **VARS** button to open the "DISTR" menu.
3. Select **2: invNorm()**.
4. Use the invNorm(area to the left of z, mean, standard deviation) function and press the **ENTER** button. If done correctly, the screen will show the following:

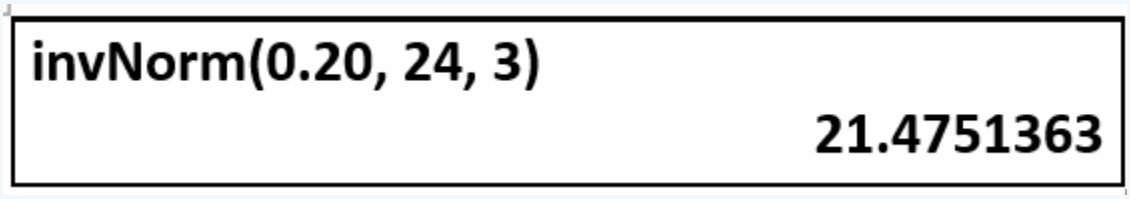


Figure 6.2.16: On graphing calculator screen, displays invNorm(0.20, 24, 3) and its calculation of 21.4751363

As a result, the maximum age to be part of the bottom 20% of Citrus College students will be roughly 21.5 years old.

b. To find the maximum age to be part of the top 20%, we draw a bell-shaped curve and shade the top 20% of the graph. This will be on the right side.

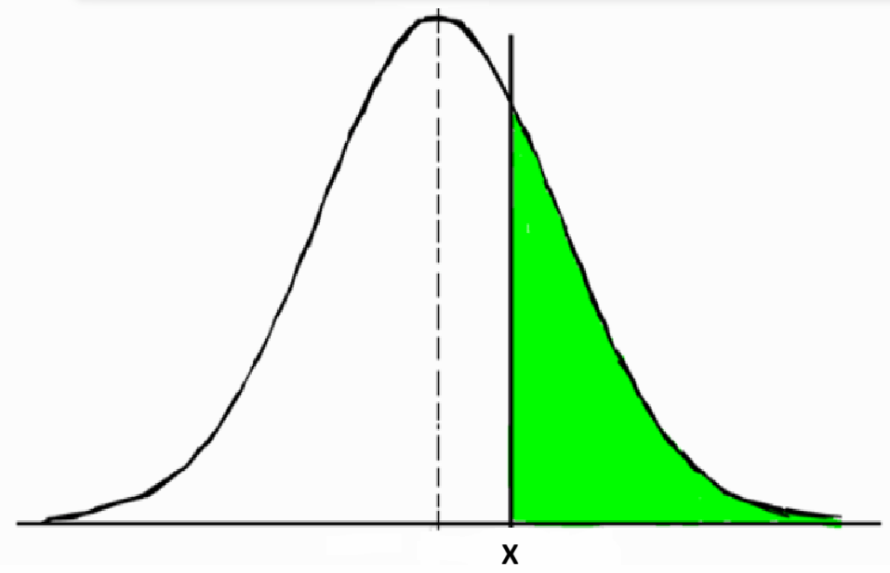


Figure 6.2.17: Normal distribution with top 20% shaded where the fence is the minimum value.

Calculating z Using invNorm() function:

To find the z-value with a graphing calculator, use the invNorm() function.

1. Press the **2ND** button.
2. Press the **VARS** button to open the "DISTR" menu.
3. Select **2: invNorm()**.
4. Use the invNorm(area to the left of z, mean, standard deviation) function and press the **ENTER** button. In this case, the right is 20% so the area to the left of z will be 80%. If done correctly, the screen will show the following:

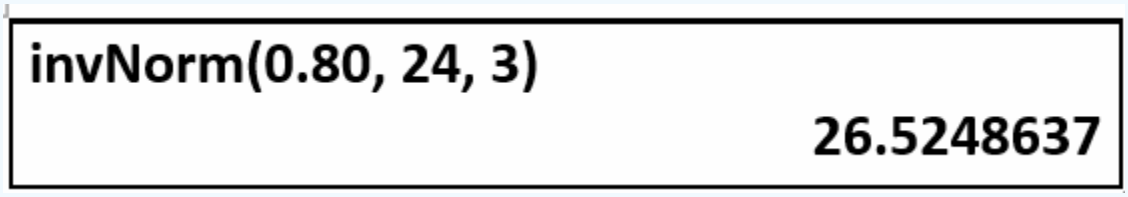


Figure 6.2.18: On graphing calculator screen, displays invNorm(0.80, 24, 3) and its calculation of 26.5248637

As a result, the minimum age to be part of the top 20% of Citrus College students will be roughly 26.5 years old.

c. To find the minimum and maximum ages that encompass the middle 50% of the age distribution at Citrus College, we draw a bell-shaped curve and shade the middle 50% of the graph.

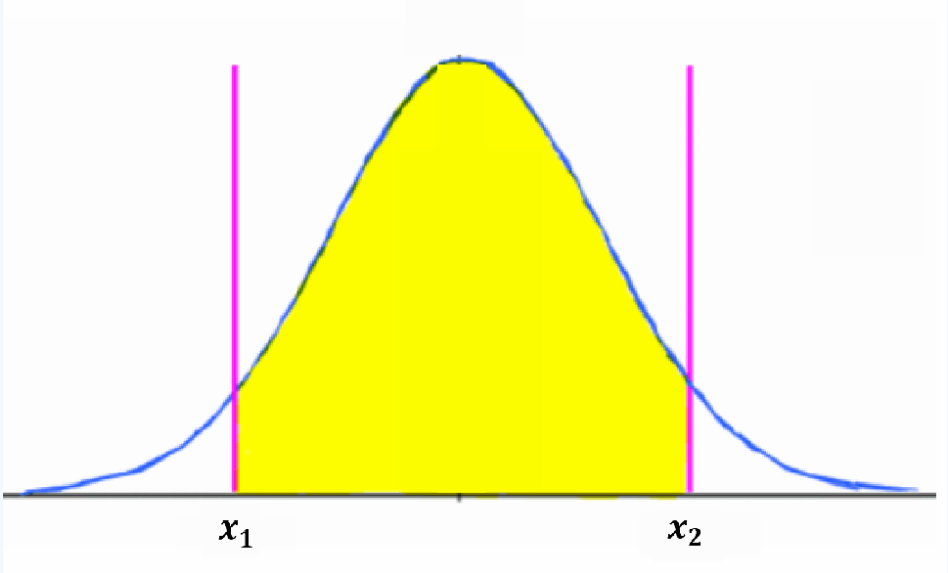


Figure 6.2.19: Normal distribution with middle 50% shaded, where we have a minimum and maximum vertical fences on the sides.

Calculating z Using invNorm() function:

To find the z-value with a graphing calculator, use the invNorm() function.

1. Press the **2ND** button.
2. Press the **VARS** button to open the "DISTR" menu.
3. Select **2: invNorm()**.
4. Use the invNorm(area to the left of z, mean, standard deviation) function to find the minimum value. In order to find the minimum value, the area to the left of the fence is what is needed. The whole graph should add up to 100%. The area to the left and right is symmetric. As a result, the outside areas have to be 25%. Here is an image to illustrate this idea.

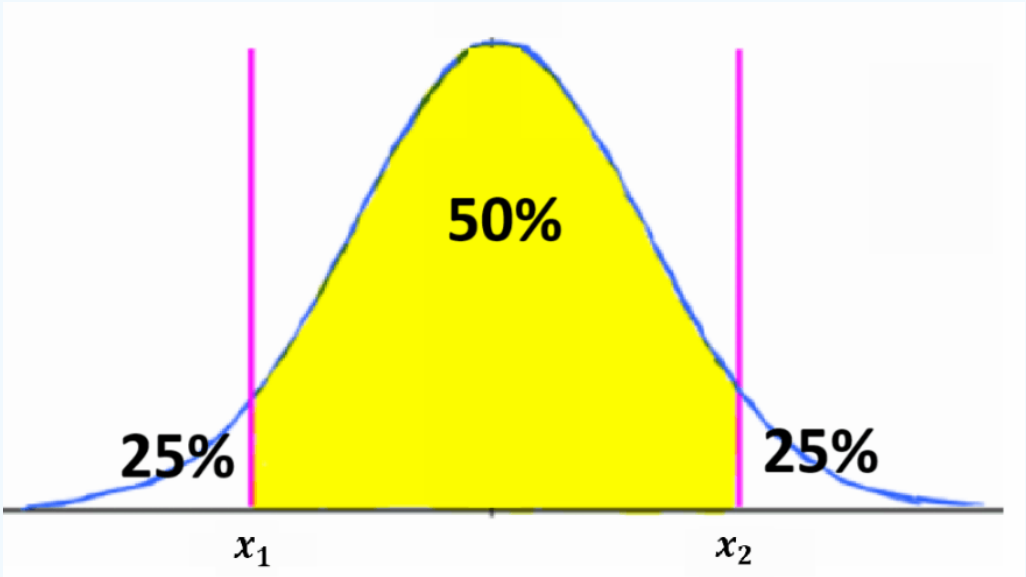


Figure 6.2.20: Normal distribution graph where the middle region is 50% and the side regions are 25%.

As a result, the area to the left of the minimum fence (x_1) will be 25%. The area to the left of the maximum fence (x_2) will be the middle region (50%) plus the left region (25%) which will be 75%. These are the values we put for the area in decimal form inside invNorm() function. The mean for both will still be 24 and standard deviation will be 3. When done inputting the invNorm() function on the calculator, press the **ENTER** button. If done correctly, the screen will show the following:

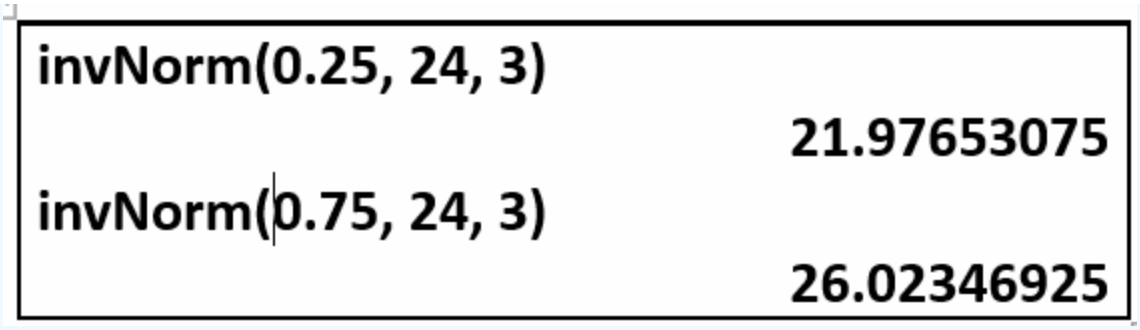


Figure 6.2.21: On graphing calculator screen, displays shows invNorm(0.25, 24, 3) and its calculation of 21.97653075. It also shows invNorm(0.75, 24, 3) and its calculation of 26.02346925.

As a result, the middle 50% of the age distribution at Citrus College is roughly between 22 and 26 years old.

Authors

"6.2: Finding Probabilities for the Normal Distribution" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"6.3: Finding Probabilities for the Normal Distribution" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 6.2: Finding Probabilities for the Normal Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 6.3: Finding Probabilities for the Normal Distribution by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

6.3: The Central Limit Theorem

Learning Objectives

- Understand the Central Limit Theorem, stating that the distribution of sample means approaches a normal distribution as sample size increases, regardless of the. population's shape
- Use the Central Limit Theorem to compute probabilities about sample means using z-scores.
- Apply this theorem in inferential statistics, including confidence intervals and hypothesis testing.

The central limit theorem (CLT) states that, regardless of the original population distribution, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases, typically when the sample size is 30 or more. This theorem is fundamental in statistics because it allows for inferential analysis, meaning we can make predictions about a population using sample data. It is widely used in hypothesis testing, confidence intervals, and quality control, helping researchers and analysts conclude large populations from smaller, manageable samples.

Definition 6.3.1:Statistical Inference

Statistical Inference: to make accurate decisions about parameters from statistics.

When it says "accurate decision," you want to be able to measure how accurate. You measure how accurate using probability. In both binomial and normal distributions, you needed to know that the random variable followed either distribution. You need to know how the statistic is distributed, and then you can find probabilities. In other words, you need to know the shape of the sample mean or whatever statistic you want to decide about.

How is the statistic distributed? This is answered with a sampling distribution.

Definition 6.3.2: Sampling Distribution

Sampling Distribution: how a sample statistic is distributed when repeated trials of size *n* are taken.

This depends on how the original distribution is distributed. However, using the Central Limit Theorem, we do not need to know the shape of the population distribution. The sampling distribution will be normal as long as n is at least 30.

Definition: Central Limit Theorem

The central limit theorem (CLT) states that, for a sufficiently large sample size, the distribution of the sample mean will be approximately normal, regardless of the shape of the original population distribution. This means that as the sample size increases, the sampling distribution of the mean becomes more normally distributed, making it useful for statistical inference, hypothesis testing, and confidence interval estimation.

- The mean of the distribution will μ
- The standard deviation is called the standard error and is $\frac{\sigma}{\sqrt{n}}$.

What this says is that no matter what *x* looks like, $\overline{x}$ would look normal if *n* is large enough. Now, what size of *n* is large enough? That depends on how *x* is distributed in the first place. If the original random variable is normally distributed, then *n* just needs to be 2 or more data points. If the original random variable is somewhat mound-shaped and symmetrical, then *n* needs to be greater than or equal to 30. Sometimes the sample size can be smaller, but this is a good rule of thumb. The sample size may have to be much larger if the original random variable is skewed one way or another.

Now that you know what the sample mean will look like in a normal distribution, you can find the probability related to the sample mean. Remember that the mean of the sample mean is just the mean of the original data ($\mu_{\overline{x}} = \mu$), but the standard deviation of the sample mean, $\sigma_{\overline{x}}$, also known as the standard error of the mean, is actually $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. Make sure you use this in all calculations.

If you are using the z-score, the formula when working with $\overline{x}$ is $z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$. If you are using the TI-83/84 calculator, then the input would be normalcdf(lower limit, upper limit, μ, $\sigma/\sqrt{n}$). If you are using R, then the input would be pnorm($\overline{x}, \mu, \sigma/\operatorname{sqrt}(n)$) to find the area to the left of $\overline{x}$. Remember to subtract pnorm($\overline{x}, \mu, \sigma/\operatorname{sqrt}(n)$) from 1 if you want the area to the right of $\overline{x}$.

Example 6.3.1 general normal distribution

The length of a human pregnancy is normally distributed with a mean of 272 days with a standard deviation of 9 days (Bhat & Kushtagi, 2006).

Click on blue links to access the normal distribution tables for negative z-values and positive z-values. After the value is found, press "back" to return to the page.

a. State the random variable.
b. Find the probability that the mean of 30 pregnancies lasts more than 273.46 days.
c. Find the probability that the mean of 30 pregnancies lasts less than 267.99 days.
d. Find the probability that the mean of 30 pregnancies lasts between 267.99 and 273.46 days.

Solution

a. $\overline{x}$ = mean length of 30 human pregnancies

b. First, translate the statement into a mathematical statement.

$$P(\overline{x} > 273.46)$$

Now, draw a picture. Remember that the center of this normal curve is 272.

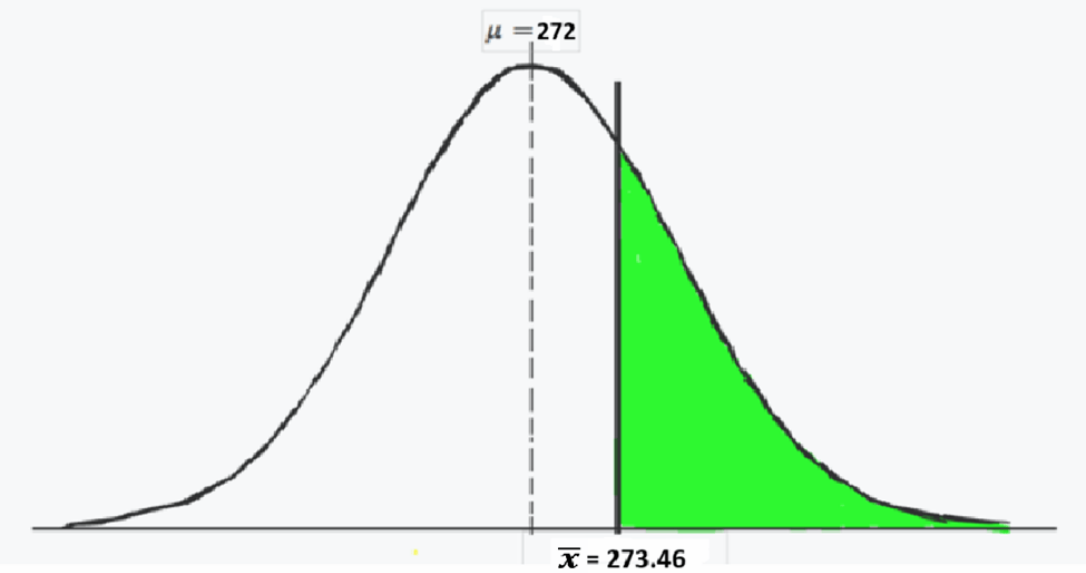


Figure 6.3.1: Normal distribution where μ of 272 is located in the middle where $\overline{x}$ = 273.46 is a fence on the left of the center. The shaded area is on the right.

To find the probability on the TI-83/84, looking at the picture, you realize the lower limit is 273.46. The upper limit is infinity. The calculator doesn't have infinity on it, so you need to put in a really big number. Some people like to put in 1000, but if you are working with numbers that are bigger than 1000, then you would have to remember to change the upper limit. The safest number to use is 1×10^{99}, which you put in the calculator as 1E99 (where E is the EE button on the calculator). The μ (population mean) will be 272. The population standard deviation of our sample means ($\sigma_{\overline{x}}$) will be $\frac{\sigma}{\sqrt{n}}$.

The work to find $\sigma_{\overline{x}}$ is provided below.

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\overline{x}} = \frac{9}{\sqrt{30}}$$

$$\sigma_{\overline{x}} = 1.64$$

As a result, to find the probability, we will use the normalcdf() command in our TI 83/84 calculator where the lower limit will be 273.46, the upper limit will be 1 E 99, μ (population mean) will be 272, and population standard deviation of our sample means ($\sigma_{\overline{x}}$) will be 1.64.

The command looks like:

$\text{normalcdf}(273.46, 1E99, 272, 1.64)$

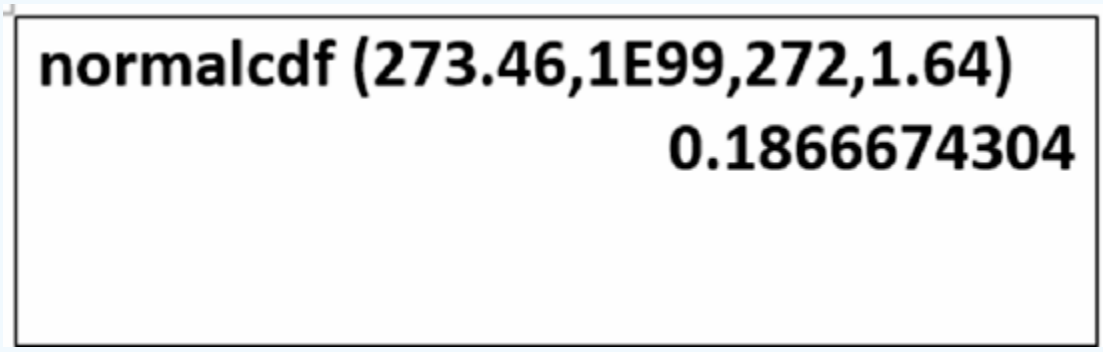


Figure 6.3.2: TI-83/84 Output for Area to the Right of 273.46.

Thus, $P(\overline{x} > 273.46) \approx 0.1867$

Thus, 18.67% of the mean pregnancies of 30 women last more than 273.46 days. This is not unusual since the probability is greater than 5%.

To find the probability using a z-table, we need to first compute the z-score and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of the mean of 30 pregnancies lasting more than 273.46 days, so $\overline{x}$ will be 273.46 days.

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{(\overline{x} - \mu)}{\sigma/\sqrt{n}}$$

$$z = \frac{(273.46 - 272)}{9/\sqrt{30}}$$

$$z = \frac{1.46}{1.64}$$

$z = 0.89$ (rounded to two decimal places)

As a result, $P(\overline{x} > 273.46) = P(z > 0.89)$. Since z is more than 0.89, our shaded area will be on the right of z = 0.89.

We can use the z-table to find our probability to the left. If we convert the normal distribution to the standard normal distribution, we get the following:

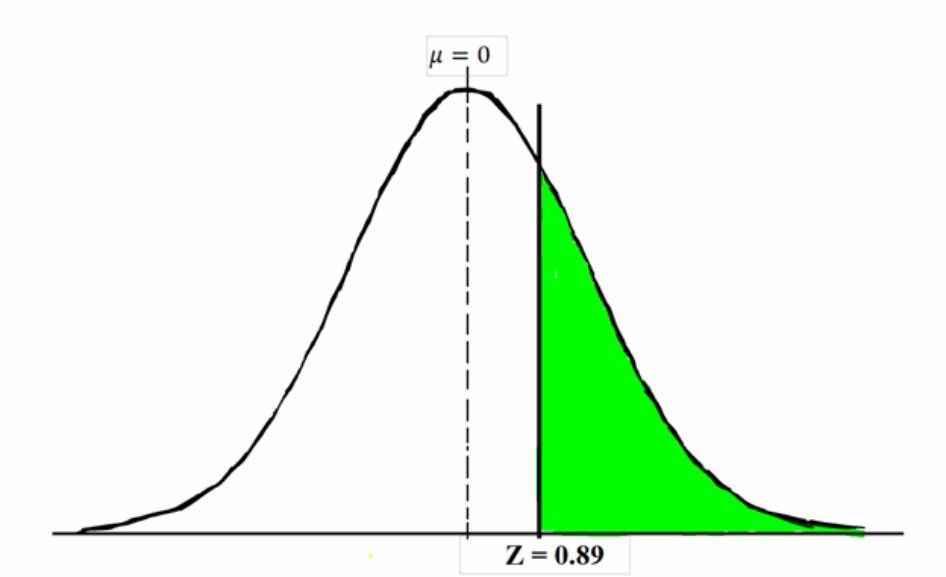


Figure 6.3.3: Normal Distribution, Z = 0.89, Shaded on the Right

From the z-table, it only gives us the probability to the left of the z-score. Since we know the total area of the normal distribution is 1. To get the area to the right, we need to take 1 minus the area to the left. The work is provided below.

P(z > 0.89) = 1 - P(z < 0.89)

To find the area or probability to the left of z = 0.89, we use the standard normal distribution table provided below. We go down 0.8 and move over to the right 0.09 to get the area of 0.8133. This area (probability) represents the area to the left of z=0.89.

Table 6.3.5: Normal Distribution Table For Positive Z-values

Based on the z-table above, the Left Area will be 0.8133.

Right Area = 1 - Left Area

Right Area = 1 - 0.8133 = 0.1867

P(z>0.89)= 0.1867 = 18.67%

Using the table, $P(\overline{x} > 273.46)$ = 18.67%. This means that the probability that the mean of 30 pregnancies will last more than 273.46 days is 18.67%.

The answer using the z-table is slightly off due to rounding our z-value to two decimal places. The calculator uses more decimal places for z, hence it has a more accurate answer.

c. First, translate the statement into a mathematical statement.

$$P(\overline{x} < 267.99)$$

Now, draw a picture. Remember that the center of this normal curve is 272.

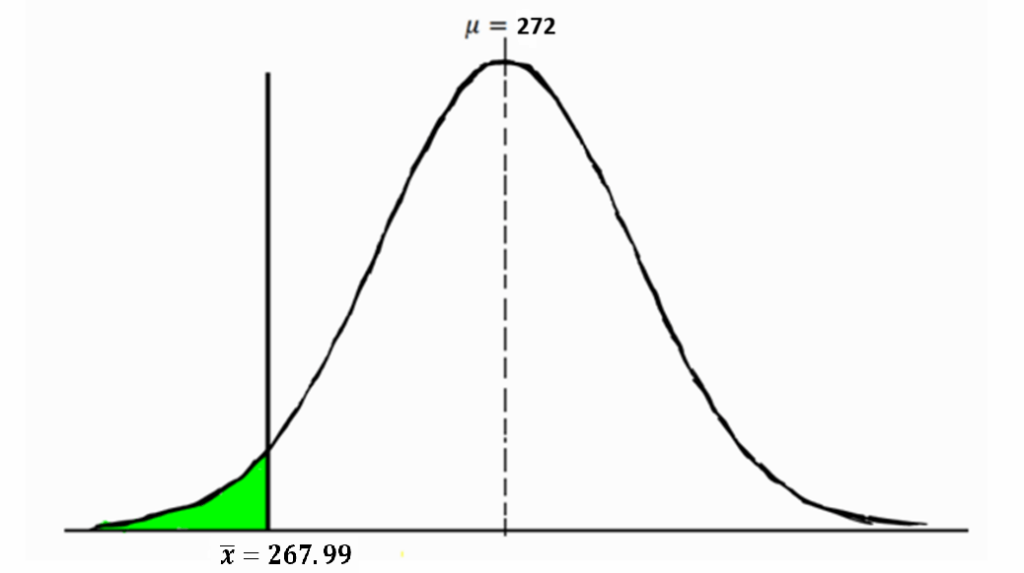


Figure 6.3.4: Normal Distribution with Area to the Left of 267.99

To find the probability on the TI-83/84, looking at the picture, though it is hard to see in this case, the lower limit is negative infinity. Again, the calculator doesn't have this on it, put in a really small number, such as $-1 \times 10^{99} = -1E99$ on the calculator.

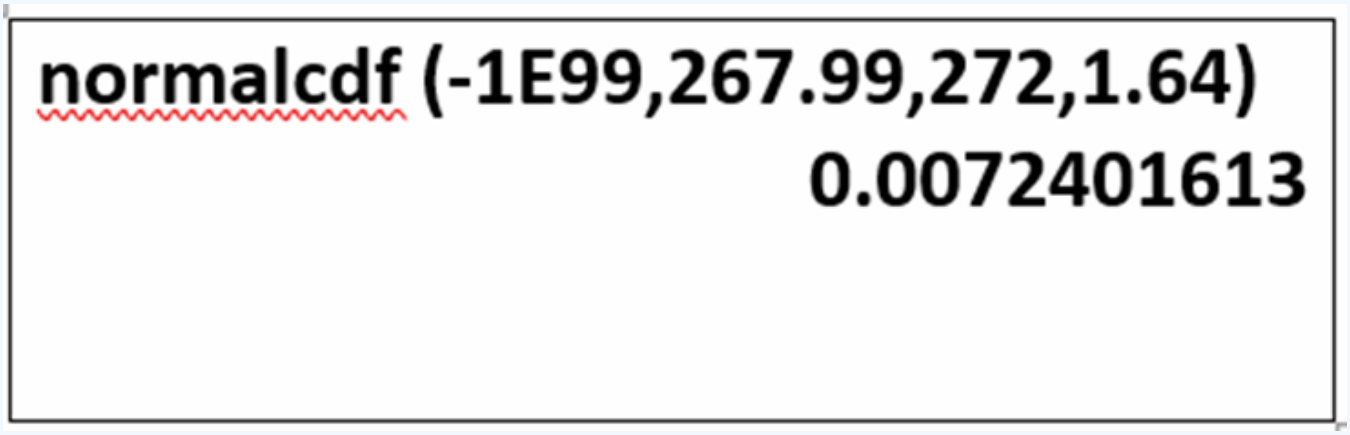


Figure 6.3.5: TI-83/84 Output for the Area to the Left of 267.99

$$P(\overline{x} < 267.99) = \text{normalcdf } (-1E99, 267.99, 272, 1.64) = 0.0072$$

Thus, the probability that the mean of 30 pregnancies lasts less than 267.99 days is 0.72%. This is unusual since the probability is less than 5%.

To find the probability using a z-table, we need to first compute the z-score and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of the mean of 30 pregnancies that last less than 267.99 days, so $\overline{x}$ will be 267.99 days.

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{(\overline{x} - \mu)}{\sigma/\sqrt{n}}$$

$$z = \frac{(267.99 - 272)}{9/\sqrt{30}}$$

$$z = \frac{-4.01}{1.64}$$

$z = -2.45$ (rounded to two decimal places)

As a result, $P(\overline{x} < 267.99)$ = P(z < -2.45). Since z is less than -2.45, our shaded area will be on the left of z = -2.45.

We can use the z-table to find our probability to the left. If we convert the normal distribution to the standard normal distribution, we get the following:

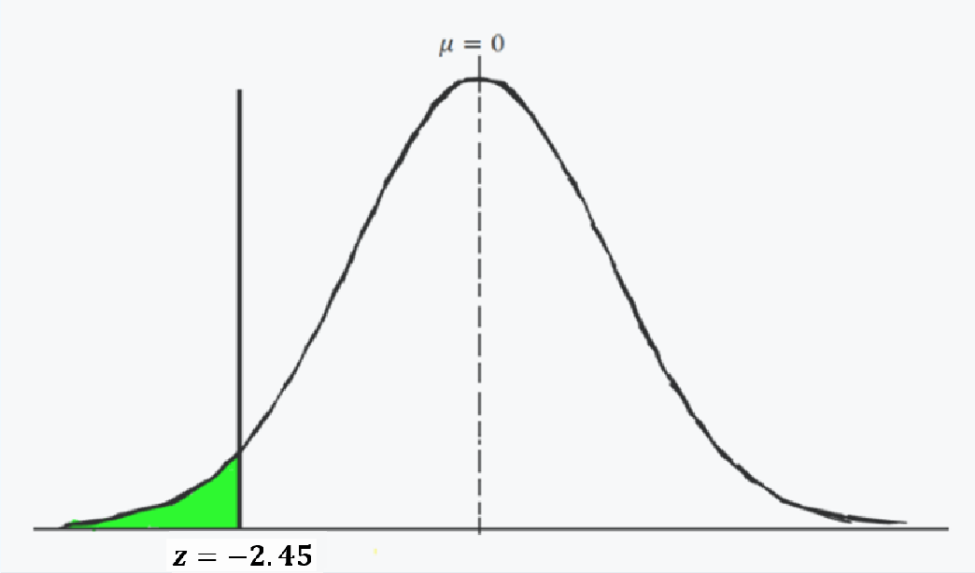


Figure 6.3.6: Normal Distribution, Z = -2.45, Shaded on the Left

We use the standard normal distribution table provided below to find the area or probability to the left of z = -2.45. We go down -2.4 and move over to the right 0.05 to get the area of 0.0071. This area (probability) represents the area to the left of z=-2.45.

Table 6.3.6: Normal Distribution Table For Negative Z-values

Based on the z-table above, the Left Area will be 0.0071, which is the probability that z is less than -2.45. If we express the answer as a percentage, we get the following answer:

P(z < -2.45) = 0.71%

Using the table, $P(\overline{x} < 267.99)$ = 0.71%. This means the probability that the mean of 30 pregnancies will last less than 267.99 days is 0.71%.

The answer using the z-table is slightly off due to rounding our z-value to two decimal places. The calculator uses more decimal places for z, hence it has a more accurate answer.

d. First translate the statement into a mathematical statement.

$$P(267.99 < \overline{x} < 273.46)$$

Now, draw a picture. Remember that the center of this normal curve is 272.

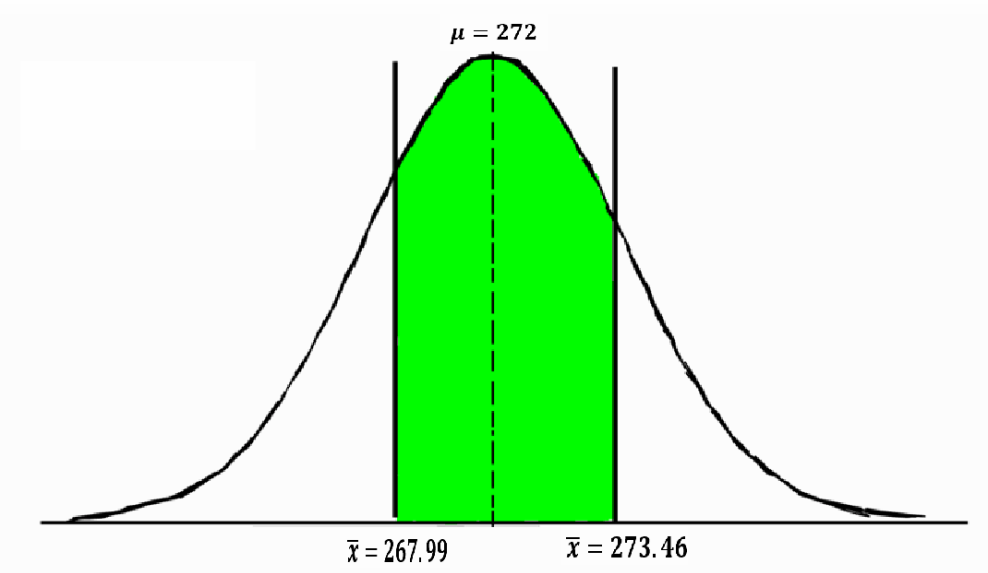


Figure 6.3.7: Normal Distribution with an Area Between 267.99 and 273.46

In this case, the lower limit is 267.99 and the upper limit is 273.46.

Using the calculator

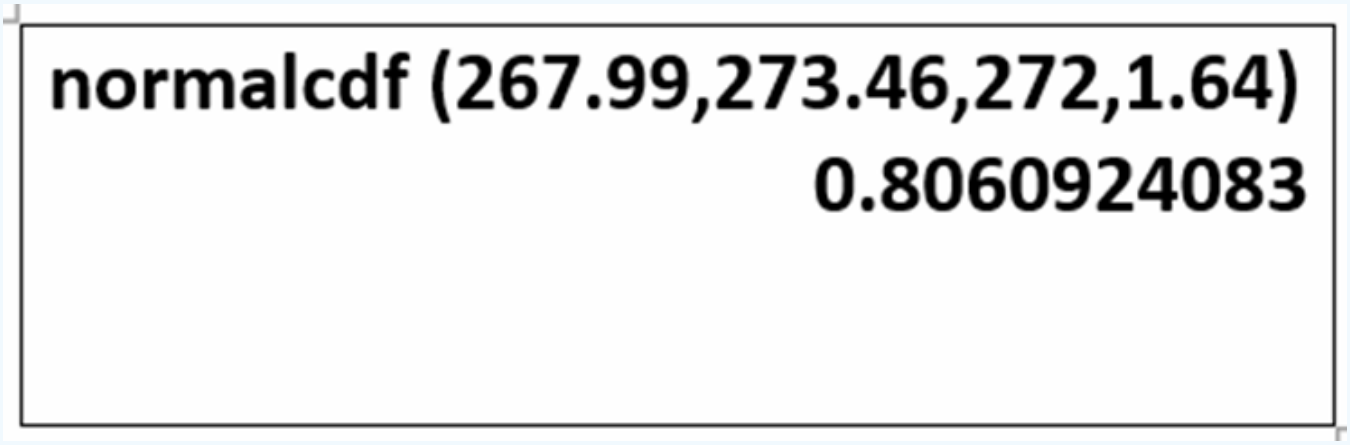


Figure 6.3.8: TI-83/84 Output for Area Between 267.99 and 273.46

$$P(267.99 < \overline{x} < 273.46) = \text{normalcdf } (267.99, 273.46, 272, 1.64) = 0.8061$$

Thus, the probability that the mean of 30 pregnancies will last between 267.99 and 273.46 days is 80.61%.

To find the probability using a z-table, we need to first compute the z-scores and then use the table to find the probability. Based on the problem, we know that μ (population mean) is 272 days and σ (population standard deviation) is 9 days. We know that we are finding the probability of the mean of 30 pregnancies lasting between 267.99 and 273.46 days, so $\overline{x}$ will be 267.99 days and 273.46 days. For each $\overline{x}$, we need to find its corresponding z-value.

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{(\overline{x} - \mu)}{\sigma/\sqrt{n}}$$

$$z = \frac{(267.99 - 272)}{9/\sqrt{30}}$$

$$z = \frac{-4.01}{1.64}$$

$z = -2.45$ (rounded to two decimal places)

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{(\overline{x} - \mu)}{\sigma/\sqrt{n}}$$

$$z = \frac{(273.46 - 272)}{9/\sqrt{30}}$$

$$z = \frac{1.46}{1.64}$$

$z = 0.89$ (rounded to two decimal places)

As a result, $P(267.99 < \overline{x} < 273.46)$ = P(-2.45 < z < 0.89).

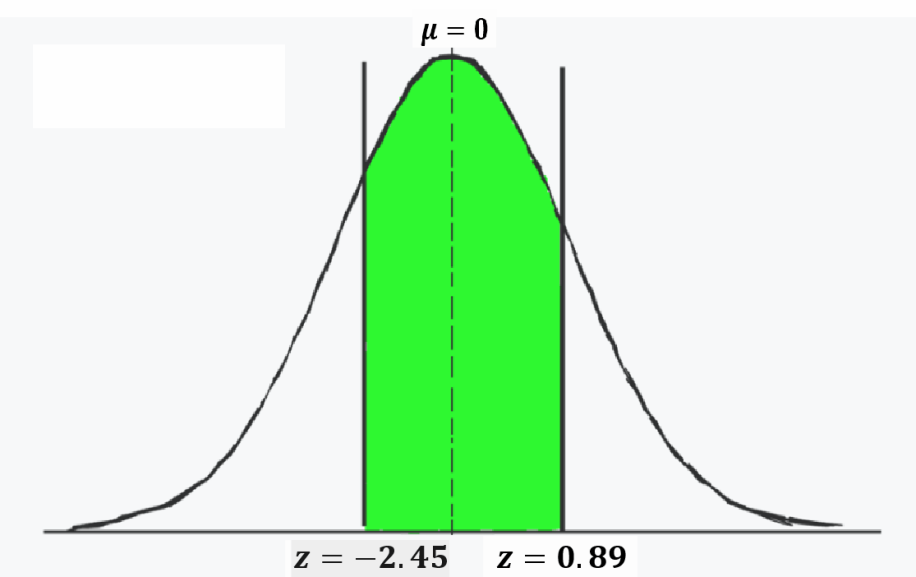


Figure 6.3.9: Normal Distribution Where Shaded Area is Between z = -2.45 and z = 0.89

We can find our probability to the left of each z-score by using the z-table. If we subtract the higher probability value from the lower probability value, we will find the area (probability) between z = -2.45 and z = 0.89.

$$P(-2.45 < z < 0.89) = p(z < 0.89) - p(z < -2.45)$$< p=""> </z<0.89\right)=p\left(z<0.89\right)-p\left(z<-2.45\right)\)<>

To find $P(z < 0.89)$ and $P(z < -2.45)$ we use the z-table.

z = 0.89

To find the area or probability to the left of z = 0.89, we use the standard normal distribution table provided below. We go down 0.8 and move over to the right 0.09 to get the area of 0.8133. This area (probability) represents the area to the left of z=0.89.

$$P(z < 0.89) = 0.8133$$

Table 6.3.5: Normal Distribution Table For Positive Z-values

z = -2.45

To find the area or probability to the left of z = -2.45, we use the standard normal distribution table provided below. We go down -2.4 and move over to the right by 0.05 to get an area of 0.0071. This area (probability) represents the area to the left of z=-2.45.

$$P\left(z<-2.45\right)=0.0071$$

Table 6.3.6: Normal Distribution Table For Negative Z-values

Now that we found $P\left(z<0.89\right)=0.8133$ and $P\left(z<-2.45\right)=0.0071$, we can use the following formula:

$$P(267.99<\overline{x}<273.46)=P\left(-2.45<z<0.89\right)=p\left(z<0.89\right)-p\left(z<-2.45\right)<\ p=""\ >\ </z<0.89\right)=p\left(z<0.89\right)-p\left(z<-2.45\right)\)<>$$

$$P(267.99<\overline{x}<273.46)=P\left(-2.45<z<0.89\right)=0.8133-0.0071<\ p=""\ >\ </z<0.89\right)=0.8133-0.0071\)<>$$

$$P(267.99<\overline{x}<273.46)=P\left(-2.45<z<0.89\right)=0.8062<\ p=""\ >\ </z<0.89\right)=0.8062\)<>$$

Thus, the probability that the mean of 30 pregnancies will last between 267.99 and 273.46 days will be 80.62%.

Authors

"6.3: The Central Limit Theorem" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"6.5: Sampling Distribution and the Central Limit Theorem" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 6.3: The Central Limit Theorem is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 6.5: Sampling Distribution and the Central Limit Theorem by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.
- 6.3: Finding Probabilities for the Normal Distribution by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

6.4: Formulas for Chapter 6

Chapter 6 Formulas

Cutoff Score

The cutoff score in a normal distribution is the value that separates a specified proportion of the data from the rest, based on a given probability. It is used to identify thresholds, such as the top 10% or bottom 5% of values. The formula to find the cutoff score x is provided directly below.

$$x = \mu + z \cdot \sigma$$

Where:

- μ is the mean
- σ is the standard deviation
- z is the z-score corresponding to the desired cumulative probability from the standard normal distribution table

z-Score for Normal Distribution

The z-score is used to convert a value from a normal distribution to the standard normal distribution. It measures how many standard deviations a value is from the mean. This allows comparisons across different normal distributions.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- x the raw score or observed value from the original normal distribution.
- μ is the mean
- σ is the standard deviation

z-Score for Distribution of Sample Means

The z-score for the distribution of sample means measures how many standard errors a sample mean $\bar{x}$ is from the population mean μ. It is used to standardize a sample mean when the sampling distribution is approximately normal.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the the sample mean
- μ is the population mean
- σ is the population standard deviation
- n is the sample size
- $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean (i.e., the standard deviation of the sampling distribution of the sample mean)

Authors

"6.4: Formulas for Chapter 6" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 6.4: Formulas for Chapter 6 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

6.5: Chapter 6 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Area under the curve: In a normal distribution, the area under the curve represents probability. The total area under the entire curve is equal to 1 (or 100%).

Central Limit Theorem: A statistical theory stating that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's distribution.

Cut-off score: A specific value that separates a certain percentage of the data from the rest, often used in probability and confidence interval calculations.

Normal distribution: A bell-shaped, symmetric probability distribution that describes how the values of a variable are distributed. Most of the observations cluster around the central peak, and probabilities for values taper off equally in both directions.

Population distribution: The distribution of a variable for the entire population. It can take any shape depending on the population data.

Sampling distribution: The distribution of a statistic (like the mean) based on all possible random samples of a given size from a population.

Standard deviation: A measure of how spread out the values in a data set are from the mean. In normal distributions, it helps determine the width of the bell curve.

Standard normal distribution: A normal distribution with a mean of 0 and a standard deviation of 1. It is used to compute probabilities and z-scores.

Z-score: A value that indicates how many standard deviations a data point is from the mean. It is used to compare values from different distributions or to find probabilities.

Key Symbols

μ - population mean.

σ - population standard deviation.

$\bar{x}$ - sample mean.

s - sample standard deviation.

z - z-score (standard score).

P - probability.

x - individual data value.

n - sample size.

N- population size.

$\sigma_{\bar{x}}$ - standard deviation of the sampling distribution of the sample mean (standard error).

x_c -cut-off score or value that separates a certain portion of the distribution

Authors

"6.5: Chapter 6 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 6.5: Chapter 6 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

7: CONFIDENCE INTERVAL FOR ONE SAMPLE

7.1: Introduction to Confidence Intervals
7.2: Confidence Interval for the Mean Using z-values
7.3: Confidence Interval for the Mean Using t-values
7.4: Confidence Interval and Sample Size for the Proportion
7.5: Formulas for Chapter 7
7.6: Chapter 7 - Key Terms and Symbols

This page titled 7: Confidence Interval for One Sample is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

7.1: Introduction to Confidence Intervals

Learning Objectives

- Estimate population parameters (e.g., mean or proportion) using confidence intervals based on sample statistics.
- Provide a range that likely contains the true population value at a specified confidence level (e.g., 95%).
- Measure the reliability of estimates and account for sampling variability in concluding data.

Point Estimate:

A point estimator is a statistic that is calculated from a sample. As an example, when you want to estimate the population mean, μ, the point estimator is the sample mean, $\overline{x}$. To estimate the population proportion, *p*, you use the sample proportion, $\hat{p}$.

Point estimators are easy to calculate, but they have some drawbacks. First, if the sample size is large, then the estimate is better. However, the sample size is not always given for a point estimator. Also, it is challenging to measure the accuracy of the forecast. Both of these problems are solved with a confidence interval.

Definition 7.1.1

Confidence interval: This is where you have an interval surrounding your parameter, and the interval has a chance of being a true statement. In general, a confidence interval looks like this: The standard formula is *Point Estimate* $\pm E$, where *E* is the margin of error term that is added and subtracted from the point estimator. Thus making an interval.

Interpreting a confidence interval:

The statistical interpretation is that the confidence interval has a probability (1 - α, where α is the complement of the confidence level) of containing the population parameter. For example, if there is a 95% confidence interval of 0.65 < p < 0.73, then it can be said that "there is a 95% chance that the interval 0.65 to 0.73 contains the true population proportion." This means that for 100 intervals, 95 of them will contain the true proportion, and 5% will not. The wrong interpretation is that there is a 95% chance that the true value of *p* will fall between 0.65 and 0.73. This interpretation is inaccurate because the true value (called the parameter) is fixed. Fixed values are not random and do not have probabilities associated with it. Only the interval is random. In the end, the intervals are randomly computed to capture the parameter with the interval. So the probability is the chance that each interval captures a parameter, not that the parameter falls in the interval.

The common probabilities utilized to compute the confidence intervals are 90%, 95%, and 99%. These are known as the confidence levels. The confidence level and the alpha level are related. For a two-tailed test, the confidence level is *C* = 1 - α. This is because the α is both tails and the confidence level is the area between the two tails. As an example, for a two-tailed test (H_1 is not equal to) with α equal to 0.10, the confidence level would be 0.90 or 90%. If you have a one-tailed test, then your α is only one tail. Because of symmetry, the other tail is also α. So you have 2α with both tails. So the confidence level, which is the area between the two tails, is *C* = 1 - 2α. A visual of the probability distribution using a normal distribution is provided below.

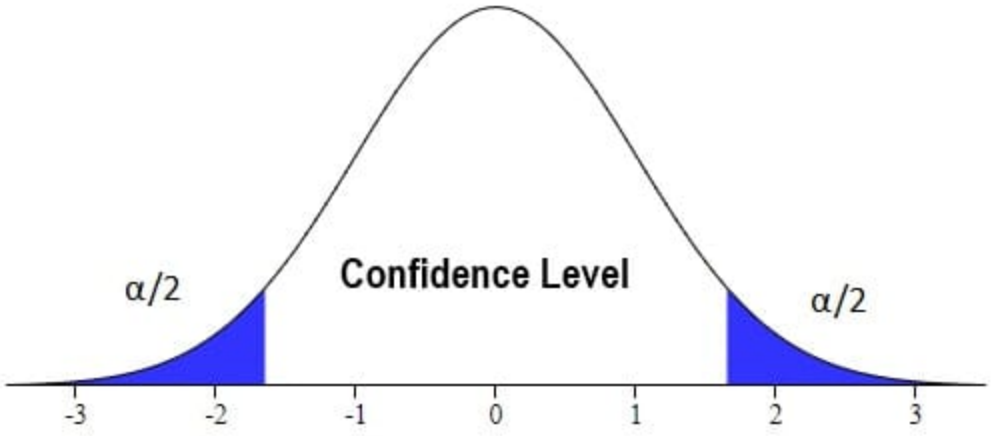


Figure 7.1.1: Confidence Level and α Over 2 Areas.

Example 7.1.1 stating the statistical and real world interpretations for a confidence interval

a. Suppose the confidence level is 95%. The confidence interval for the mean age a woman gets married in 2013 is $26 < \mu < 28$. State the statistical and real-world interpretations of this statement.
b. Suppose a 99% confidence interval for the proportion of Americans who have tried marijuana as of 2013 is \(0.35

Solution

a. Statistical Interpretation: There is a 95% chance that the interval $26 < \mu < 28$ contains the mean age a woman gets married in 2013.
Real World Interpretation: The mean age that a woman married in 2013 is between 26 and 28 years of age.
b. Statistical Interpretation: There is a 99% chance that the interval \(0.35

One final thing to understand about confidence is how the sample size and confidence level affect how wide the interval is. The following discussion demonstrates what happens to the width of the interval as you get more confident.

Think about shooting an arrow into the target. Suppose a person is good at shooting arrows and has a 90% chance of hitting the bull's eye. Keep in mind that the bull's eye is small. Since the person hits the bull's eye approximately 90% of the time, the person will probably hit inside the next ring out 95% of the time. The person will have a better chance if the bull's eye can be increased. In such a situation, the person may have a 99% chance of hitting the target, but that is a much bigger circle to hit. As your confidence in hitting the target increases, the circle you hit gets bigger. The same is true for confidence intervals. This is demonstrated in *Figure* 7.1.1.

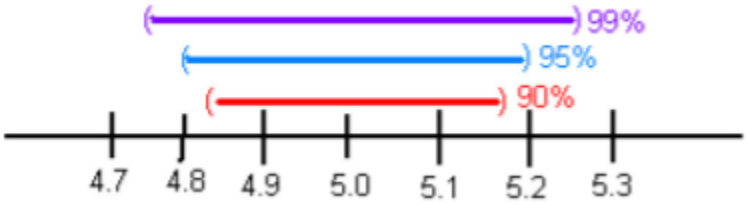


Figure 7.1.2: Affect of Confidence Level on Width

The higher level of confidence makes a wider interval. There's a trade-off between width and confidence level. You can be confident about your answer but the answer will not be precise. Or you can have a precise answer (small margin of error) but not be very confident about your answer.

Now look at how the sample size affects the size of the interval. Suppose *Figure* 7.1.2 represents confidence intervals calculated on a 95% interval. A larger sample size from a representative sample makes the width of the interval narrower. This makes sense. Large samples are closer to the true population so the point estimate is close to the true value.

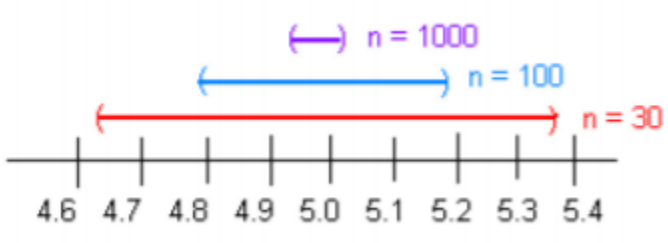


Figure 7.1.3: Affect of Sample Size on Width

Now that the foundations underlying confidence intervals have been introduced. The next step is to demonstrate how to calculate them using a formula. The formula depends on which parameter is being estimated, the given confidence level, and other information.

Authors

"7.1: Introduction to Confidence Intervals" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.1: Basics of Confidence Intervals" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 7.1: Introduction to Confidence Intervals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: Basics of Confidence Intervals by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

7.2: Confidence Interval for the Mean Using z-values

Learning Objectives

- Compute confidence intervals for the mean using z-values when the population standard deviation is known.
- Ensure the sample is randomly selected to support valid inference.
- Estimate the range where the true population mean is likely to lie at a specified confidence level.
- Determine and meet minimum sample size requirements to ensure accuracy and reliability.

One of the parameters that can be estimated using a confidence interval is the mean μ. If the population is normally distributed, the population standard deviation σ is known or $n \geq 30$, then the sampling distribution is normal or approximately normal. Thus, the confidence level will be represented as a middle area under a standard normal distribution, and the confidence interval can be derived as follows.

1. Construct a standard normal distribution such that the confidence level is the middle area in the distribution.

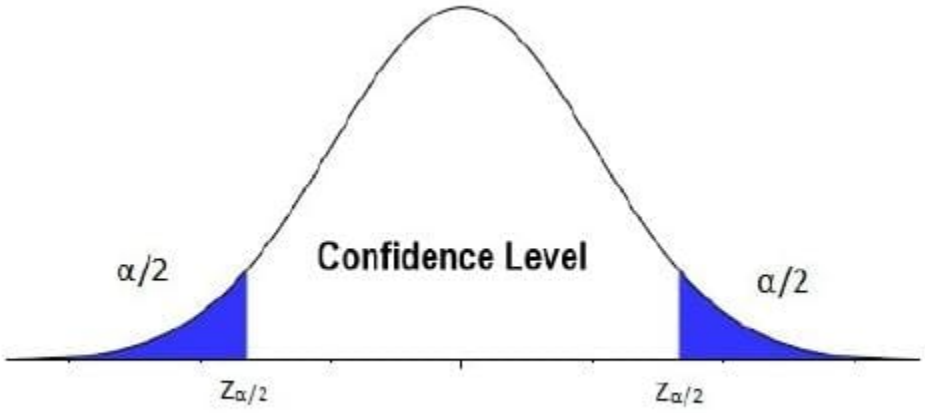


Figure 7.2.1: Standard Normal Distribution with Critical Values

2. The Z-statistic can fall into the range $Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$. Using the central limit theorem, Z-statistic can be rewritten as $Z_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}$. The formula can be solved for μ using algebra to derive the confidence interval formula.

Definition: Confidence Interval for μ when σ is Known

$$\overline{X} - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- $\overline{X}$ is the sample mean.
- σ is the population standard deviation.
- n is the sample size.
- $Z_{\frac{\alpha}{2}}$ is the z-value that can be found using the Z-Table and the given confidence level.
- μ is the population mean that is being estimated.

Examples

Example 7.2.1

A preventative healthcare researcher wishes to estimate the average number of daily steps people take in a Southern California region. The population standard deviation is around 750 steps per day and the distribution is approximately normal. She collects a random sample of 50 people who reside in the region and computes the sample mean, which is 3485 steps per day. Compute a confidence interval for the mean using a confidence level of 98%.

Solution

1. List all the given information.

- $\overline{X}$ = 3485
- σ = 750
- n = 50
- $Z_{\frac{\alpha}{2}}$ = 2.33. See Table A.1 below

Confidence Level	Critical Value
99%	2.575
98%	2.33
95%	1.96
90%	1.645
80%	1.28

Figure 7.2.2: Common Critical Values for the Confidence Level

2. Compute the confidence interval by plugging in given information into the formula and calculating the endpoints using the order of operations.

$$\overline{X} - Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$3485 - 2.33 \cdot \left(\frac{750}{\sqrt{50}}\right) < \mu < 3485 + 2.33 \cdot \left(\frac{750}{\sqrt{50}}\right)$$

$$3885 - 247.13 < \mu < 3885 + 247.13$$

$$3637.87 < \mu < 4132.13$$

Example 7.2.2

An educational researcher wishes to estimate the average age of students who attend school in the evenings at a local college. The ages of students are normally distributed and the population standard deviation is 5.2 years. The researcher collected a random sample of 37 students who attend evening classes at the college and determined the sample mean to be 28.8. Use a 95% confidence level to estimate the average age of evening students who attend the college.

Solution

1. List all the given information.

- $\overline{X}$ = 28.8
- σ = 5.2
- n = 37
- $Z_{\frac{\alpha}{2}}$ = 1.96. See Table A.1.

2. Compute the confidence interval by plugging in given information into the formula and calculating the endpoints using the order of operations.

$$\overline{X} - Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$28.8 - 1.96 \cdot \left(\frac{5.2}{\sqrt{37}}\right) < \mu < 28.8 + 1.96 \cdot \left(\frac{5.2}{\sqrt{37}}\right)$$

$$28.8 - 1.68 < \mu < 28.8 + 1.68$$

$$27.12 < \mu < 30.48$$

Example 7.2.3

An arborist who works for a large city in Southern California wants to estimate the mean age of Jacaranda trees in the city. The arborist collected a random sample of 40 Jacaranda trees and determined the average age to be 32.3 years. The ages of the Jacaranda trees are normally distributed with a population standard deviation of 6.5 years. Use the TI 83/84 calculator to estimate the mean age of the Jacaranda trees using a confidence level of 90%.

Solution

1. Go into the **[STAT]** menu and use the right arrows to select **[TESTS]** then choose **[7:ZInterval]**. Select **[STATS**] and type in the information shown in (7.2.4). Finally, select **[CALCULATE]** and hit enter.

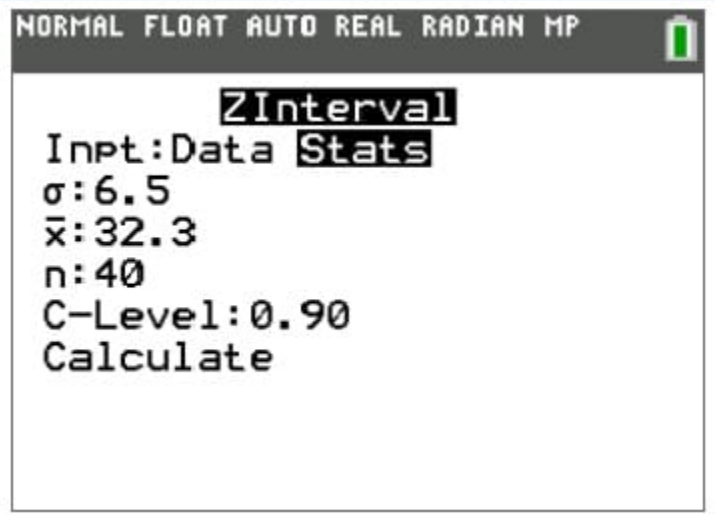


Figure 7.2.3: Input for Confidence Interval in TI-84+

2. The output displayed in the TI 83/84 calculator screen should mirror the one displayed in the figure below.

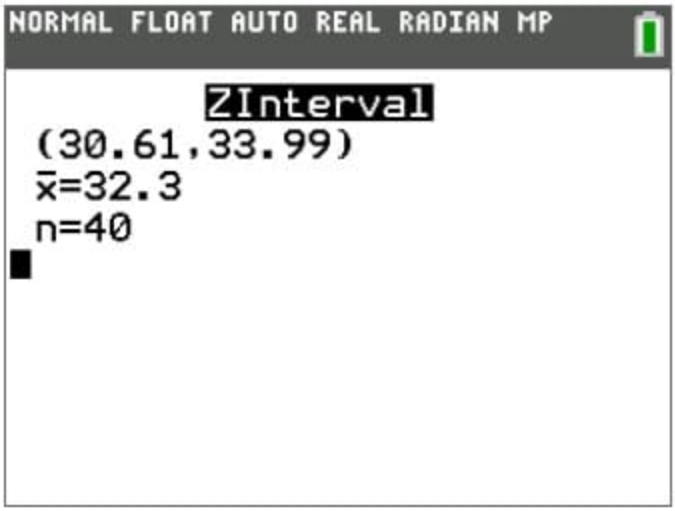


Figure 7.2.4: Calculator Output for TI-84+

3. The confidence interval for the mean age of Jacaranda trees is built using the endpoints displayed between the parenthesis in the figure above.

$30.61 < \mu < 33.99$

Minimum Sample Size

The minimum size needed to construct a confidence interval for the mean μ can be derived from the confidence interval formula. The margin of error E represents the range of variability around the estimated value. To find the minimum sample size, set $E = Z_C \left(\frac{\sigma}{\sqrt{n}} \right)$ and solve for n. The formula is presented below.

Definition: Minimum Sample Size for a Confidence Interval of the Mean

The minimum sample size needed to compute a confidence interval for the mean is defined using the formula below. Always round the final answer upward to the nearest whole number.

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

- n is the minimum sample size.
- σ is the population standard deviation.
- E is the margin of error.
- $Z_{\frac{\alpha}{2}}$ is the z-value that can be found using Table A.1 using the confidence level.

Example of Minimum Sample Size

Example 7.2.4

An entomologist is studying the life span in years of the flam skipper dragonfly (Libellula saturate). It has been determined that the population standard deviation in years is 0.85 years. The entomologist will use a confidence level of 98% and a margin of error of 0.25 years. Compute the minimum sample size needed for n.

Solution

1. List the information.

- σ =0.85
- E =0.25
- $Z_{\frac{\alpha}{2}}$ = 2.33 using table A.1.

2. Compute n by plugging information into the formula.

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

$$n = \left(\frac{2.33 \cdot 0.85}{0.25} \right)^2$$

$$n = 62.758$$

3. Round n upwards,

$$n \approx 63$$

The minimum sample size is 63.

Author

"7.2: Confidence Interval for the Mean Using z-values" by Alfie Swan is licensed under CC BY 4.0

This page titled 7.2: Confidence Interval for the Mean Using z-values is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

7.3: Confidence Interval for the Mean Using t-values

Learning Objectives

- Compute confidence intervals for the mean using t-values when the population standard deviation is unknown.
- Use the sample standard deviation to estimate variability.
- Understand that the t-distribution accounts for additional uncertainty, especially with small sample sizes.
- Recognize that the t-distribution approaches the normal distribution as sample size increases.

t-distribution:

For most confidence intervals, the population standard deviation will not be known. Thus, the distribution used to estimate the population mean will most likely not be normal. The t-distribution will be used instead of the normal distribution when σ is unknown.

The t-distribution was discovered by William S. Gosset in 1908. He was a statistician employed by Guinness Brewery who was not allowed to use his name when publishing his discovery. Thus, he used the pseudonym "student", and his distribution was known as Student's t- distribution. A graph of the t-distribution is listed below

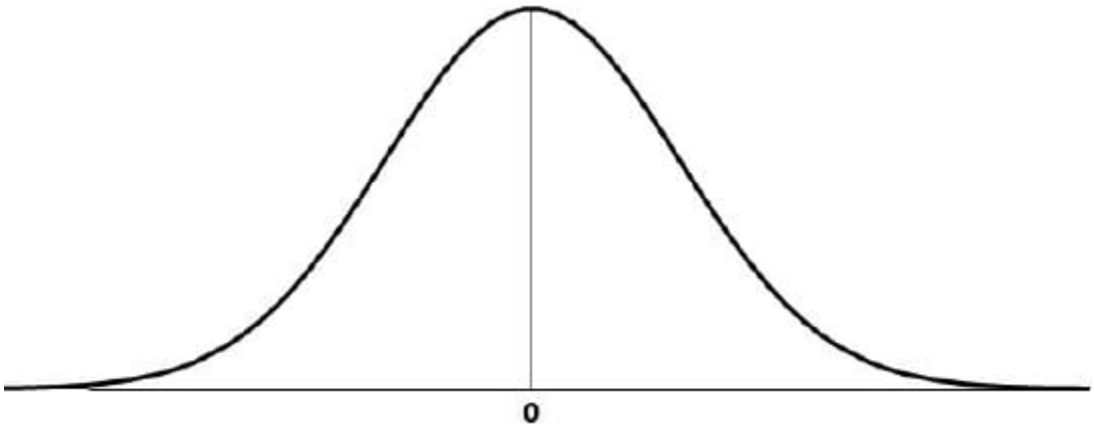


Figure 7.3.1: Graph of t-distribution. (Copyright; author via source)

Key Properties of the t-distribution

- It is symmetric.
- The mean is 0.
- The mean, median, and mode are all equal and located at the center of the distribution.
- The graph never touches the horizontal axis.
- The variance is greater than 1.
- The graph is a family of curves based on the concept of degrees of freedom (d.f. = n - 1).
- As the sample size n gets larger, the shape of the t-distribution becomes normal.

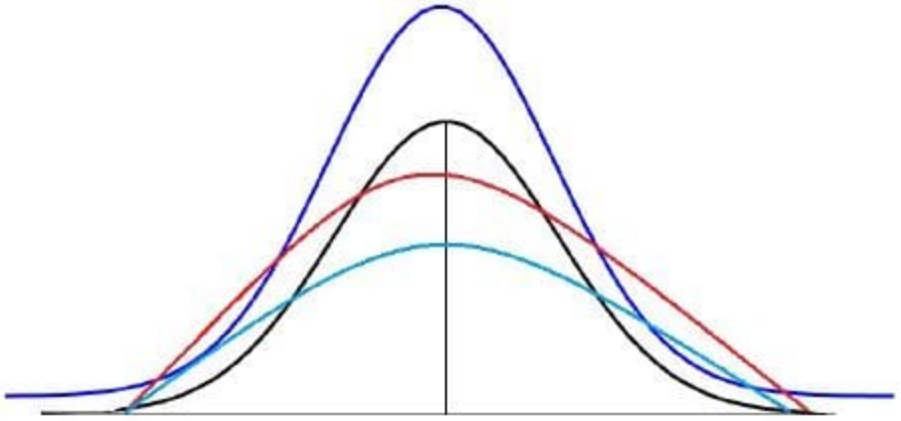


Figure 7.3.3: Graph of the t-distribution with Different Degrees of Freedom (Approaches the Normal Distribution)

Confidence Interval for One Population Mean (t-Interval) Formula:

The confidence interval formula can be found by using a confidence level as the middle area of a t-distribution and then using algebra to solve for μ using the corresponding t-values associated with the confidence level. A graph of the confidence level in the t-distribution along with corresponding t-values is presented in the graph below.

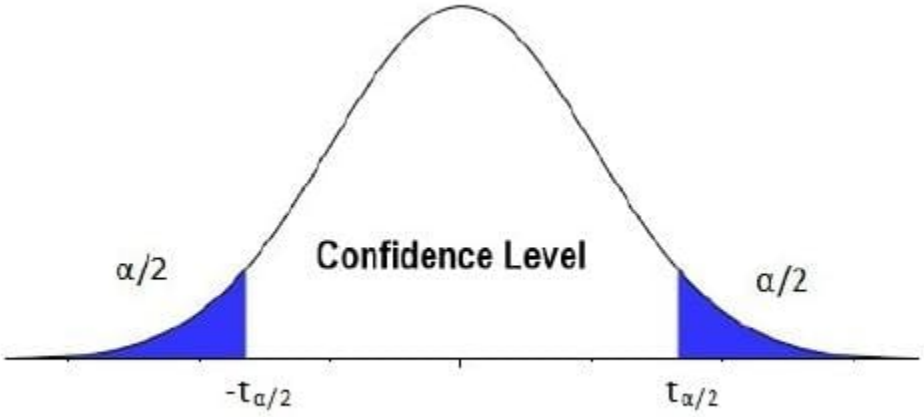


Figure 7.3.4: Graph of t-distribution with Critical Values.

The range of values on the horizontal axis leads to the following inequality \(-t_ \dfrac{\alpha}{2}

Definition: Confidence Interval for μ when σ is unknown (t-values)

$$\overline{X} - t\frac{\alpha}{2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t\frac{\alpha}{2}\left(\frac{s}{\sqrt{n}}\right)$$

- $\overline{X}$ is the sample mean.
- s is the sample standard deviation.
- n is the sample size.
- $T\frac{\alpha}{2}$ is the t-value found in Table A2 using the confidence level and d.f. = n - 1.
- μ is the population mean that is being estimated.

Examples:

Example 7.3.1

Suppose a researcher wishes to estimate the age of a student at a California community college who passes an elementary statistics class. The researcher collects a sample of 11 random students. The sample mean age is 22.6 and the sample standard deviation is 1.88. The researcher decides to use a confidence level of 95%.

Solution:

Step 1) Label the given statistics.

- $\overline{X}$ = 22.6 (sample mean)
- s = 1.88 (sample standard deviation)
- n = 11 (sample size)
- Confidence level = 95%

Step 2) Look up the $t_{\frac{\alpha}{2}}$ value in Table A.2 using the confidence level and d.f. = n - 1 = 11 - 1 = 10. In Table A.2, go down the first column to 10 degrees of freedom. Then go over to the column headed with 95%. Thus $t_{\frac{\alpha}{2}} = 2.228$. (See Example 7.3.1.)

- $t_{\frac{\alpha}{2}}$ = 2.228

Degrees of Freedom (df)	80%	89%	95%	98%	99%
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
10	1.382	1.812	**2.228**	2.764	3.169

Figure 7.3.4: Table of Critical Values for the t-distribution at Degrees of Freedom of 10

Step 3) Plug the information into the formula and use the order of operations to calculate the two endpoints.

$$\overline{X} - t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

$$22.6 - 2.228\left(\frac{1.88}{\sqrt{11}}\right) < \mu < 22.6 + 2.228\left(\frac{1.88}{\sqrt{11}}\right)$$

$$22.6 - 1.26 < \mu < 22.6 + 1.26$$

$$21.34 < \mu < 23.86$$

Example 7.3.2 confidence interval for the population mean using the formula

A random sample of 20 IQ scores of famous people was taken information from the website of IQ of Famous People ("IQ of famous," 2013) and then using a random number generator to pick 20 of them. The data are in Example 7.3.2 (this is the same data set that was used in Example 7.3.2). Find a 98% confidence interval for the average IQ of a famous person.

Table 7.3.1: IQ Scores of Famous People

158	180	150	137	109
225	122	138	145	180
118	118	126	140	165
150	170	105	154	118

1. State the random variable and the parameter in words.
2. State and check the assumptions for a confidence interval.
3. Find the sample statistic and confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. x = IQ score of a famous person

μ = mean IQ score of a famous person

2.

a. A random sample of 20 IQ scores was taken. This was stated in the problem.
b. The population of IQ scores is normally distributed. This was shown in Example 7.3.2.

3. Sample Statistic:

$\overline{x} = 145.4$

$s = 29.27$

Calculate the degrees of freedom, *df* = *n* - 1 = 20 - 1 = 19, and use the confidence level, which is 98%, to look up the t-value. In table A.2, go down the first column to 19 degrees of freedom. Then go over to the column headed with 98%. Thus $t_{\frac{\alpha}{2}} = 2.539$. (See Example 7.3.2.)

Degrees of Freedom (*df*)	80%	90%	95%	98%	99%
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
19	1.328	1.729	2.093	**2.539**	2.861

Figure 7.3.5: Table of Critical Values for the t-distribution at Degrees of Freedom of 19

$$\overline{X} - t\frac{\alpha}{2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + t\frac{\alpha}{2}\left(\frac{s}{\sqrt{n}}\right)$$

$$145.4 - 2.539\left(\frac{29.27}{\sqrt{20}}\right) < \mu < 145.4 + 2.539\left(\frac{29.27}{\sqrt{20}}\right)$$

$$145.4 - 16.6 < \mu < 145.4 + 16.6$$

$$128.8 < \mu < 162$$

4. There is a 98% chance that $128.8 < \mu < 162$ contains the mean IQ score of a famous person.

5. The mean IQ score of a famous person is between 128.8 and 162.

Example 7.3.3 confidence interval for the population mean using technology

For this example, the TI-83/84 is used to compute a confidence interval for the mean. In this problem, a random sample of 10 final exam scores for a statistics class is collected from multiple sections over the past 20 years. The scores are provided in the data set below. Construct a confidence interval for the mean with a confidence level of 99%.

Table 7.3.2: Final Exam Scores for Statistics Class

85	98	74	56	62
95	87	77	60	84

Solution

1. On the TI-83/84: Go into the **[STAT]** menu. Select **[EDIT]** and make sure **[1:EDIT]** is highlighted. Hit enter, and then type the data into L_1.

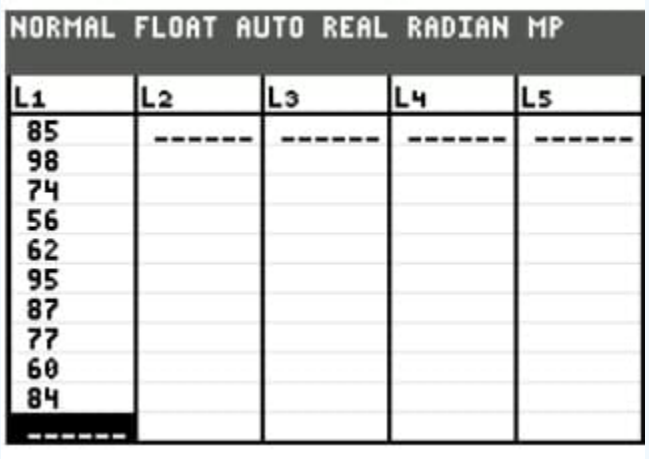


Figure 7.3.6: Enter Data into List 1

2. Go into the **[STAT]** menu and use the right arrows to select **[TESTS]** then choose **[8:TInterval]**. Select **[DATA]** and type in the information shown in (7.3.4). Finally, select **[CALCULATE]** and hit enter.

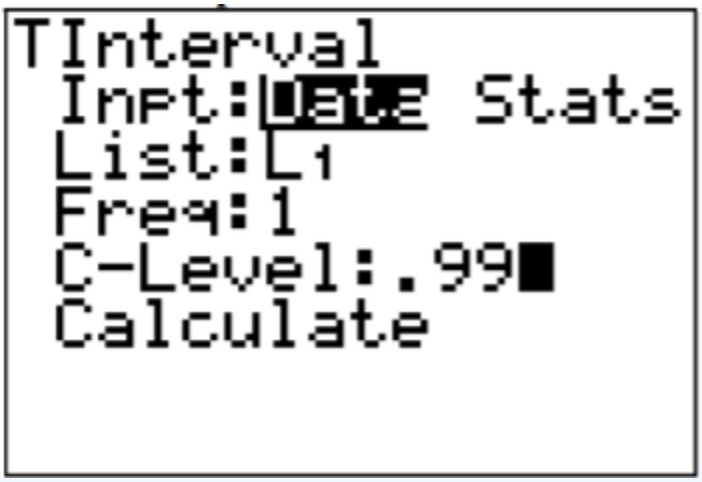


Figure 7.3.7: Setup for the t-Interval (Confidence Interval for t- Distribution)

3. The output should look as follows.

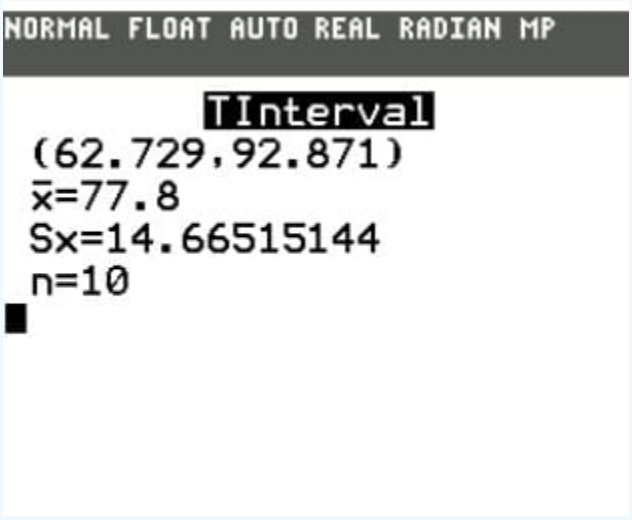


Figure 7.3.8: Output of the Endpoints of the Confidence Interval

4. There is a 99% chance that 62.7 points < μ < 92.9 points contains the mean score of the final exam.

5. The mean final exam score for the statistics class is between 62.7 and 92.9 years.

Authors

"7.3: Confidence Interval for the Mean Using t-values" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.3: One-Sample Interval for the Mean" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 7.3: Confidence Interval for the Mean Using t-values is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.3: One-Sample Interval for the Mean by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

7.4: Confidence Interval and Sample Size for the Proportion

Learning Objectives

- Compute confidence intervals for a population proportion using z-values based on data from a random sample.
- Estimate the range in which the true population proportion is likely to fall at a specified confidence level.
- Ensure the sample size is large enough to justify the normal approximation.
- Meet minimum sample size requirements to support accurate and valid z-score calculations.

Introduction

A confidence interval for a proportion is a range of values used to estimate the true population proportion based on a sample. It provides an interval within which we expect the actual proportion to fall, with a certain level of confidence (such as 90%, 95%, or 99%). This type of interval is commonly used in surveys, polls, and studies involving categorical data, such as estimating the percentage of voters who support a candidate or the proportion of defective products in a shipment. The interval is calculated using the sample proportion, the sample size, and a critical value from the standard normal (z) distribution.

This section will discuss how to estimate the population proportion, *p*. For example, a researcher may be curious about what proportion of students at a local college smoke regularly. To estimate the proportion *p*, a sampling distribution of sample proportions must be formed. To ensure that the distribution is normal the following inequalities must be verified.

$n \cdot \hat{p} \geq 5$ and $n \cdot \hat{q} \geq 5$

- $\hat{p}$ is the proportion of success. It is computed as $\hat{p} = \frac{X}{n}$
- $\hat{q}$ is the proportion of failure. It is computed as $\hat{q} = 1 - \hat{p}$
- n is the sample size.
- X is the number of items in the sample that possess a characteristic being studied.

After normality is established a sampling distribution is formed for $\hat{p}$. This sampling distribution has a mean of p. Also, the standard deviation of the distribution is $\sqrt{\frac{pq}{n}}$. A similar process to that described in a previous section can be used to establish a confidence interval to estimate *p*. The formula is provided below.

Confidence Interval for One Population Proportion (1-Prop Interval)

Definition: Confidence Interval to Estimate Population Proportion *p*

$$\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Examples of Confidence Intervals for *p*

Example 7.4.1

Suppose an educational researcher wishes to estimate the proportion of students who smoke regularly at a local college. She randomly samples 50 students and finds that 9 of them smoke regularly. To estimate the proportion she will use a confidence interval with a confidence level of 95%. Compute the confidence level using the formula.

Solution

Step 1) Compute the sample proportion $\hat{p}$.

$\hat{p} = \frac{9}{50}$ = 0.18

Step 2) Compute $\hat{q}$.

$\hat{q}$ = 1 - 0.18 = 0.82

Step 3) Look up $Z_{\frac{\alpha}{2}}$ using Table A.1. The value is 1.96.

Confidence Level	Critical Value
99%	2.575
98%	2.33
95%	1.96
90%	1.645
80%	1.28

Figure 7.4.1: Critical Values for the Given Confidence Levels in the Normal Distribution

Step 4) Calculate the confidence interval.

$\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$

$0.18 - 1.96 \cdot \sqrt{\frac{0.18 \cdot 0.82}{50}} < p < 0.18 + 1.96 \cdot \sqrt{\frac{0.18 \cdot 0.82}{50}}$

$0.18 - 0.106 < p < 0.18 + 0.106$

0.074 < *p* < 0.286

Example 7.4.2 confidence interval for the population proportion using the formula

A concern was raised in Australia that the percentage of deaths of Aboriginal prisoners was higher than the percentage of deaths of non-Aboriginal prisoners, which is 0.27%. A sample of six years (1990-1995) of data was collected, and it was found that out of 14,495 Aboriginal prisoners, 51 died ("Indigenous deaths in," 1996). Find a 95% confidence interval for the proportion of Aboriginal prisoners who died.

1. State the random variable and the parameter in words.
2. State and check the assumptions for a confidence interval.
3. Find the sample statistic and the confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. x = number of Aboriginal prisoners who die

p = proportion of Aboriginal prisoners who die

2.

a. A simple random sample of 14,495 Aboriginal prisoners was taken. However, the sample was not random since it was data from six years. The numbers are for all prisoners in these six years, but the six years were not picked at random. Unless there was something special about the chosen six years, the sample is probably representative. This assumption is probably met.
b. There are 14,495 prisoners in this case. The prisoners are all Aboriginals, so you are not mixing Aboriginal with non-Aboriginal prisoners. There are only two outcomes, either the prisoner dies or doesn't. The chance that one prisoner dies over another may not be constant, but if you consider all prisoners the same, then it may be close to the same probability. Thus the assumptions for the binomial distribution are satisfied
c. In this case, x = 51 and n - x = 14495 - 51 = 14444; both are greater than or equal to 5. The sampling distribution for $\hat{p}$ is normal.

3. Compute Sample Proportions:

$$\hat{p} = \frac{x}{n} = \frac{51}{14495} \approx 0.003518$$

$$\hat{q} = 1 - 0.003518 = 0.996482$$

Confidence Interval:

$z_{\frac{\alpha}{2}} = 1.96$, since 95% confidence level. This value is found in Table A.1.

$$\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.003518 - 1.96 \cdot \sqrt{\frac{0.003518 \cdot 0.996482}{14444}} < p < 0.003518 + 1.96 \cdot \sqrt{\frac{0.003518 \cdot 0.996482}{14444}}$$

\(0.003518-0.000964

\(0.002554

4. There is a 95% chance that \(0.002554

5. The proportion of Aboriginal prisoners who died is between 0.0026 and 0.0045.

Confidence intervals for p can also be done using technology. The following example shows the process on the TI-83/84.

Example 7.4.3 confidence interval for the population proportion using technology

A researcher studying the effects of income levels on the breastfeeding of infants hypothesizes that countries where the income level is lower have a higher rate of infant breastfeeding than higher-income countries. It is known that in Germany, considered a high-income country by the World Bank, 22% of all babies are breastfed. In Tajikistan, considered a low-income country by the World Bank, researchers found that in a random sample of 500 new mothers 125 were breastfeeding their infants. Find a 90% confidence interval of the proportion of mothers in low-income countries who breastfeed their infants.

1. State your random variable and the parameter in words.
2. State and check the assumptions for a confidence interval.
3. Find the sample statistic and the confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. x = number of women who breastfeed in a low-income country

p = proportion of women who breastfeed in a low-income country

2.

a. A simple random sample of 500 breastfeeding habits of women in a low-income country was taken as was stated in the problem.
b. There were 500 women in the study. The women are considered identical, though they probably have some differences. There are only two outcomes, either the woman breastfeeds or she doesn't. The probability of a woman breastfeeding is probably not the same for each woman, but it is probably not very different for each woman. The assumptions for the binomial distribution are satisfied
c. x = 125 and n - x = 500 - 125 = 375 and both are greater than or equal to 5, so the sampling distribution of $\hat{p}$ is well approximated by a normal curve.

3. On the TI-83/84: Go into the **[STAT]** menu. Move over to **[TESTS]** and choose **[1-PropZInt]**. Type in X, n, and the confidence level, then select **[CALCULATE]** and press enter.

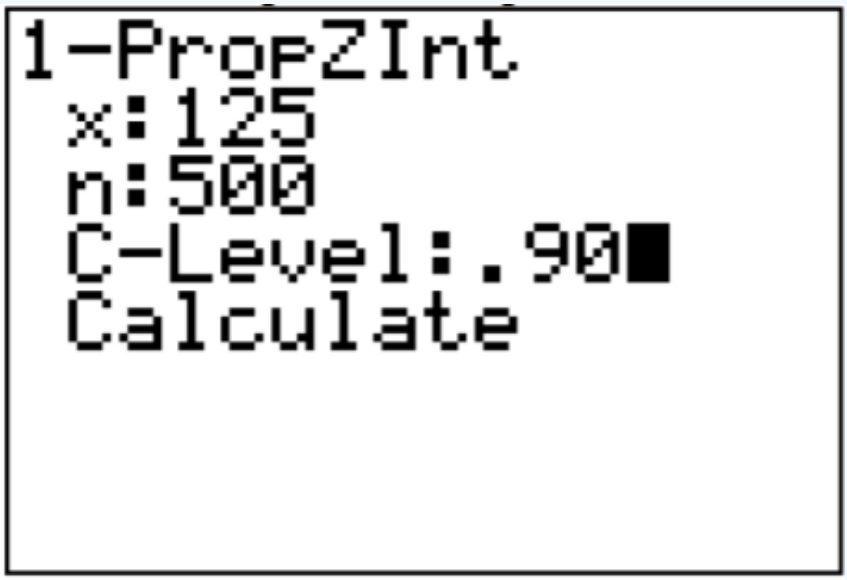


Figure 7.4.2: Setup for 1-Proportion z-Interval

Once you press Calculate, you will see the results in the figure below.

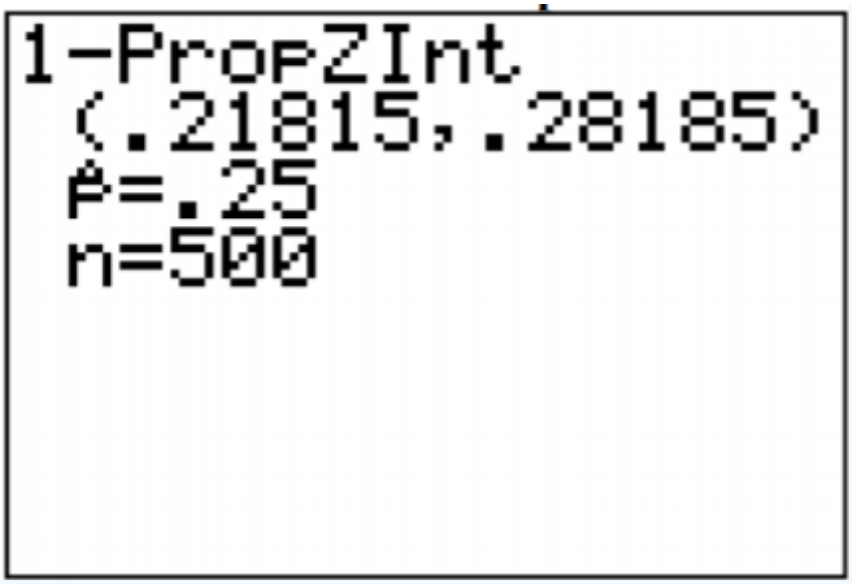


Figure 7.4.3: Results for 1-Proportion z-Interval

The answer is computed using the TI-83/84 calculator and is written below after rounding each endpoint to three decimal places.

0.219 < *p* < 0.284

4. There is a 90% chance that 0.219 < *p* < 0.284 contains the proportion of women in low-income countries who breastfeed their infants.

5. The proportion of women in low-income countries who breastfeed their infants is between 0.219 and 0.284.

Minimum Sample Size

The minimum sample needed to construct a confidence interval for *p* can also be computed. The formula is found by using the margin of error. The formula is $E = Z\frac{\alpha}{2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$. Using algebra, the equation is used to isolate *n*. The formula is provided below.

Definition: Minimum Sample Size Formula

$$n = \hat{p} \cdot \hat{q} \left(\frac{Z\frac{\alpha}{2}}{E} \right)^2$$

- E is the margin of error.
- Round *n* up to the next whole number.

Example 7.4.4

A pollster wishes to construct a new confidence interval for the proportion of voters who are in favor of voting for a new proposition that reduces taxes on energy. Using previous data, she estimates the sample proportion to be 61.5%. What is the minimum sample size needed to construct an updated confidence interval for *p* with a confidence level of 90% and a margin of error of 10%?

Solution

Step 1) Write down the given information and look up $Z_{\frac{\alpha}{2}}$ in Table A.1.

- $\hat{p} = 0.615$
- $\hat{q} = 1 - 0.615 = 0.385$
- E = 0.10
- $Z_{\frac{\alpha}{2}} = 1.645$

Step 2) Compute *n*.

$$n = \hat{p} \cdot \hat{q} \left(\frac{Z_{\frac{\alpha}{2}}}{E} \right)^2$$

$$n = 0.615 \cdot 0.385 \left(\frac{1.645}{0.10} \right)^2$$

$$n = 64.07190694$$

Step 3) Round n upwards to the next whole number.

$$n \approx 65$$

Step 4) State the result.

The minimum sample size needed is 65 people.

Authors

"7.4: Confidence Interval and Sample Size for the Proportion" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.2: One-Sample Interval for the Proportion - Statistics" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 7.4: Confidence Interval and Sample Size for the Proportion is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: One-Sample Interval for the Proportion by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

7.5: Formulas for Chapter 7

Formulas for Chapter 7

Confidence Interval for the Mean When σ is Known

A confidence interval of the mean using z-values is a range of values, based on sample data, that is likely to contain the true population mean with a certain level of confidence, such as 95%. This approach is used when the population standard deviation is known and the sample size is sufficiently large (typically n ≥ 30). The confidence interval is constructed using the sample mean, the z-score corresponding to the desired confidence level, and the standard error of the mean.

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Where:

- $\bar{x}$ is the sample mean
- $z_{\alpha/2}$ is the z-score corresponding to the desired confidence level
- σ is the population standard deviation
- n is the sample size
- μ is the population mean
- $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean

Confidence Interval for the Mean When σ is Not Known

A confidence interval for the mean using t-values is used when the population standard deviation is unknown and the sample size is relatively small, typically less than 30. In this case, the sample standard deviation is used instead of the population standard deviation, and the t-distribution is used instead of the normal distribution. This method provides a range of values that is likely to contain the true population mean with a specified level of confidence.

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Where:

- $\bar{x}$ is the sample mean
- $t_{\alpha/2}$ is the t-score for the desired confidence level and degrees of freedom
- s is the sample standard deviation
- n is the sample size
- μ is the population mean
- $\frac{s}{\sqrt{n}}$ is the standard error of the mean

Sample Proportion and Complement

The sample proportion of success, denoted as $\hat{p}$, is the ratio of the number of successes in the sample to the total number of observations. It is calculated using the formula $\hat{p} = \frac{X}{n}$, where X is the number of observed successes and n is the sample size. The complement of the sample proportion, denoted as $\hat{q}$, represents the proportion of failures and is given by $\hat{q} = 1 - \hat{p}$.

Confidence Interval for Proportion

A confidence interval for proportions is used to estimate the true population proportion based on a sample proportion. It provides a range of values within which the population proportion is likely to fall, with a certain level of confidence. This method

is appropriate when the sample size is large enough such that both $n\hat{p}$ and $n\hat{q}$ are at least 5, where $\hat{q} = 1 - \hat{p}$. The normal approximation to the binomial distribution is used to construct the interval.

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Where:

- $\hat{p}$is the sample proportion
- $\hat{q} = 1 - \hat{p}$ is the complement of the sample proportion
- $z_{\alpha/2}$ is the z-score corresponding to the desired confidence level
- n is the sample size
- p is the population proportion
- $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is the standard error of the proportion

Minimum Sample Size for a Confidence Interval for the Mean

The formula for minimum sample size is used to determine how many observations are needed to estimate a population mean within a specified margin of error at a certain confidence level. This ensures that the resulting confidence interval is both accurate and reliable. The formula takes into account the desired precision, the variability in the population, and the confidence level.

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Where:

- n is the minimum required sample size
- $z_{\alpha/2}$ is the z-score corresponding to the desired confidence level
- σ is the population standard deviation
- E is the desired margin of error

Minimum Sample Size for a Confidence Interval for a Proportion

The formula for the minimum sample size needed to estimate a population proportion is used to calculate how many observations are required to achieve a desired margin of error at a specific confidence level. It helps ensure the estimate of the population proportion is precise and reliable.

$$n = \hat{p} \cdot \hat{q} \cdot \left(\frac{z_{\alpha/2}}{E}\right)^2$$

Where:

- n is the minimum required sample size
- $\hat{p}$ is the sample proportion (estimated proportion of success)
- $\hat{q} = 1 - \hat{p}$ is the complement of the sample proportion (estimated proportion of failure)
- $z_{\alpha/2}$ is the z-score corresponding to the desired confidence level
- E is the desired margin of error

Authors

"7.5: Formulas for Chapter 7" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 7.5: Formulas for Chapter 7 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

7.6: Chapter 7 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Confidence Interval: An interval estimate that gives a range of values likely to contain the population parameter with a certain level of confidence.

Confidence Level: The probability that the confidence interval contains the true population parameter; common levels are 90%, 95%, and 99%.

Margin of Error: The maximum expected difference between the point estimate and the actual population parameter; it reflects the precision of the estimate.

Point Estimate: A single value derived from sample data used to estimate a population parameter.

Population Mean: The average value of a characteristic for the entire population.

Population Proportion: The ratio of individuals in a population with a particular characteristic, expressed as a decimal or percent.

Proportion Complement: The value obtained by subtracting the sample or population proportion from 1; represents the opposite category.

Sample Mean: The average value of a characteristic in a sample, used as a point estimate for the population mean.

Sample Proportion: The ratio of individuals in a sample with a particular characteristic.

Sample Size: The number of observations or data points in a sample affects the accuracy and reliability of confidence intervals.

Sampling Distribution: The probability distribution of a statistic (like the sample mean or proportion) based on all possible random samples from a population.

Sampling Error: The difference between the sample statistic and the corresponding population parameter due to chance variation in sample selection.

Standard Error: A measure of the variability of a sampling distribution; it is used to construct confidence intervals.

t-Distribution: A probability distribution used instead of the normal distribution when estimating the population mean with a small sample size or when the population standard deviation is unknown.

z-Value (or z-Score): A value from the standard normal distribution used in constructing confidence intervals when the population standard deviation is known and the sample size is large.

Key Symbols

$\bar{x}$ - sample mean.

μ - population mean.

$\hat{p}$ - sample proportion.

p - population proportion.

q - complement of the population proportion (i.e., $q = 1 - p$).

$\hat{q}$ - complement of the sample proportion (i.e., $\hat{q} = 1 - \hat{p}$).

n - sample size.

E - margin of error.

α - significance level (i.e., $1 -$ confidence level).

$z_{\alpha/2}$ - critical value from the standard normal distribution.

$t_{\alpha/2}$ - critical value from the t-distribution.

σ - population standard deviation.

s - sample standard deviation.

\(SE \) - standard error.

\(CI \) - confidence interval.

\(CL \) - confidence level.

Authors

"7.6: Chapter 7 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 7.6: Chapter 7 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

8: HYPOTHESIS TESTING FOR ONE SAMPLE

8.1: Introduction to Hypothesis Testing
8.2: Type I and II Errors
8.3: z-Test for a Mean
8.4: t-Test for a Mean
8.5: z-Test for a Proportion
8.6: Formulas for Chapter 8
8.7: Chapter 8 Key Terms and Symbols

This page titled 8: Hypothesis Testing for One Sample is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

8.1: Introduction to Hypothesis Testing

Learning Objectives

- Understand the purpose and process of hypothesis testing in statistics
- Learn to state null and alternative hypotheses clearly
- Identify and select an appropriate significance level (α)
- Choose the correct statistical test based on the hypothesis type
- Calculate the test statistic using sample data
- Make informed decisions to reject or fail to reject the null hypothesis
- Distinguish between left-tailed, right-tailed, and two-tailed tests

Hypothesis testing is a statistical method used to make decisions or draw conclusions about a population based on sample data. It involves making an initial assumption (called the null hypothesis), collecting data, and then determining whether the evidence supports rejecting that assumption in favor of an alternative hypothesis. This process helps researchers test claims or ideas using probability and data.

Key Definition for Hypothesis Testing

- A statistic is a characteristic or measure from a sample. A parameter is a characteristic or measure of a population. We use statistics to generalize about parameters, known as estimations. Every time we take a sample statistic, we would expect that estimate to be close to the parameter, but not necessarily exactly equal to the unknown population parameter. How close would depend on how large a sample we took, who was sampled, how they were sampled, and other factors. Hypothesis testing is a scientific method used to evaluate claims about population parameters.

- A statistical hypothesis is an educated conjecture about a population parameter. This conjecture may or may not be true. We will take sample data and infer from the sample if there is evidence to support our claim about the unknown population parameter.

- The null hypothesis (H_0, pronounced “H-naught” or “H-zero”) is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters. The null hypothesis is assumed true until there is sufficient evidence otherwise.

- The alternative hypothesis (H_1 or H_a, pronounced “H-one” or “H-ā”) is a statistical hypothesis that states that there is a difference between a parameter and a specific value, or that there is a difference between two parameters. H_1 is always the complement of H_0.

- The level of significance is also called the significance level. We use the Greek letter α, pronounced “alpha,” to represent the significance level. The level of significance is the probability that the null hypothesis is rejected when it is true. Note: like in the previous chapter, 1 – α is the confidence level.

Initial Steps of Hypothesis Testing

When doing your research, you should set up your hypotheses and choose the significance level before analyzing the sample data.

When reading a word problem, your first step is to identify the parameter(s), for example, μ, you are testing, and which direction (left, right, or two-tailed) test you are being asked to perform. For this course, the homework problems will state the researcher’s claim; usually, this is the alternative hypothesis.

The null hypothesis is always set up as a parameter equal to some value (called the test value) or equal to another parameter. The null hypothesis is assumed true unless there is strong evidence from the sample to suggest otherwise. Similar to our judicial system, a person is innocent until the prosecutor shows enough evidence that they are not innocent.

For example, an investment company wants to build a new food cart. They know from experience that food carts are successful if they have, on average, more than 100 people a day walk by the location. They have a potential site to build on, but before they begin, they want to see if they have enough foot traffic. They observe how many people walk by the site every day over a month. The investors want to be very careful about setting up in a bad location where the food cart will fail, rather than miss the opportunity in a prime location. We have two hypotheses. For an average of more than 100 people, we would write this in symbols as $\mu > 100$. This claim needs to go into the alternative hypothesis since there is no equality, just strictly greater than 100. The complement of greater than is $\mu \leq 100$. This has a form of equality ($\leq$), so it needs to go in the null hypothesis.

We then would set up the hypotheses as:

- H_0: $\mu \leq 100$ (Do not build)
- H_1: $\mu > 100$ (Build).

When performing the hypothesis test, the test statistic assumes that the parameter is the null hypothesis is equal to some value. This still implies that the parameter could be any value less than or equal to 100, but our hypothesis test should be written as:

- H_0: $\mu = 100$
- H_1: $\mu > 100$

Either notation is fine, but most textbooks will always have the = sign in the null hypothesis. The null hypothesis is based on historical value, a claim, or a product specification.

Determining the Direction of a Hypothesis Test

When there is a greater than sign (>) in the alternative hypothesis, we call this a right-tailed test. If we had a less-than sign (<) in the alternative hypothesis, we would have a left-tailed test. If there were a no equal sign ($\neq$) in the alternative hypothesis, we would have a two-tailed test. The tails will determine which side the critical region will fall on the sampling distribution. Note that you should never have an =, $\leq$, or $\geq$ sign appear in the alternative hypothesis.

Definition: Ways to Set Up Hypotheses for the Population Mean μ

There are three ways to set up the hypotheses for a population mean μ, where k is a placeholder for the numeric test value.

Format for a Two-tailed test:

$H_0 : \mu = \text{k}$

$H_1 : \mu \neq \text{k}$

Format for a Right-tailed test:

$H_0 : \mu = \text{k}$

$H_1 : \mu > \text{k}$

Format for a Left-tailed test:

$H_0 : \mu = \text{k}$

$H_1 : \mu < \text{k}$

Guidelines For Writing The Correct Hypothesis

- The null hypothesis of a two-tailed test states that the mean μ is equal to some value k.
- The null hypothesis of a right-tailed test implies that the mean μ is less than or equal to some value k.
- The null hypothesis of a left-tailed test implies that the mean μ is greater than or equal to some value k.

Look for key phrases in the research question to help you set up the hypotheses. Make sure that the =, ≤, and ≥ signs always go in the null hypothesis. The ≠,> and < signs always go in the alternative hypothesis. Look for these phrases in Figure 8-1 to help you decide if you are setting up a two-tailed test (first column), a right-tailed test (second column), or a left-tailed test (third column).

When you read a question, you must identify the parameter of interest. The parameter determines which distribution to use. Make sure that you can recognize and distinguish which parameter you are making a conjecture about: mean = μ, proportion = p, variance = σ^2, standard deviation = σ. There will be more parameters in later chapters.

Do not use the sample statistics, like $\overline{x}$ or $\hat{p}$, in the hypotheses. We are not making any inferences about the sample statistics. We know the value of the sample statistic. We use the sample statistics to infer if a change has occurred in the population.

For example, if we were making a conjecture about the percent or proportion in a population, we would have the following hypotheses:

H_0: p = k

H_1: p ≠ k.

Setting up the hypotheses correctly is the most important step in hypothesis testing. Here are some example research questions and how to set up the null and alternative hypotheses correctly; in a later section, we will perform the entire hypothesis test.

Use Figure 8-1 as a guide in setting up your hypotheses. The first column shows the hypotheses and how to shade in the distribution for a two-tailed test, with common phrases in the claim. The two-tailed test will always have a not equal ≠ sign in the alternative hypothesis, and both tails shaded. The second column is for a right-tailed test. Note that the greater than > sign will always be in the alternative hypothesis, and the right tail is shaded. The third column is for a left-tailed test. The left-tailed test will always have a less than < sign in the alternative hypothesis, and the left tail shaded in.

Hypothesis Testing Common Phrases

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
		
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 8.1.1: Table that provides several tails per test and keywords to translate to mathematics.

Example 8.1.1

State the hypotheses in both words and symbols for the following claims.

The key phrase in the claim is “less than.” The less than sign < is only allowed in the alternative hypothesis, and we are testing against the national average.

H_0: The national mean salary is $61,420.

H_1: The GTEP director believes the mean salary in Oregon is less than $61,420.

Solution

H_0: $\mu = 61420$

H_1: $\mu < 61420$

Example 8.1.2

State the hypotheses in both words and symbols for the following claims.

The key phrase in the claim is “more than.” The greater than sign > is only allowed in the alternative hypothesis. This is about a proportion, not a mean, so use the parameter p.

H_0: The principal will not assign parking spaces if 30% or less of the students own a car.

H_1: The principal will assign parking spaces if more than 30% of students own a car.

Solution

H_0: $p = 0.3$

H_1: $p > 0.3$

Example 8.1.3

State the hypotheses in both words and symbols for the following claims.

The key word in the claim is “different.” The not equal sign ≠ is only allowed in the alternative hypothesis.

H_0: The population mean age is 26 years old.

H_1: The evening students’ mean age is believed to be different from 26 years old.

Solution

H_0: $\mu = 26$

H_1: $\mu \neq 26$

To understand the process of a hypothesis test, you need to first understand what a hypothesis is, which is an educated guess about a parameter. Once you have the alternative hypothesis, you collect data and use the data to decide if there is enough evidence to show that the alternative hypothesis is true. However, in hypothesis testing you assume something else is true, the null hypothesis, and then you look at your data to see how likely it is to get an event that your data demonstrates with that assumption. If the event is very unusual, then you might think that your assumption is false. If you can say this assumption is false, then your alternative hypothesis could be true. You assume the opposite of your alternative hypothesis is true and show that it cannot be true. If this happens, then your alternative hypothesis is probably true. All hypothesis tests go through the same process. Once you have the process down, then the concept is much easier.

When setting up your hypotheses make sure the parameter, not the statistic, is used in the hypotheses. The equality always goes

in the null hypothesis H_0 and the alternative hypothesis H_1 will be a left-tailed test with a less than sign <, a two-tailed test with a not equal sign $\neq$, or a right-tailed test with a greater than sign >.

Authors

"8.1: Introduction to Hypothesis Testing" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.1: Introduction" by Rachel Webb is licensed under CC BY-SA 4.0

"9.E: Hypothesis Testing with One Sample (Exercises)" by OpenStax is licensed under CC BY 4.0

This page titled 8.1: Introduction to Hypothesis Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: Introduction by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

8.2: Type I and II Errors

Learning Objectives

- Understand the concepts of Type I and Type II errors in hypothesis testing.
- Identify a Type I error as rejecting a true null hypothesis (false positive).
- Identify a Type II error as failing to reject a false null hypothesis (false negative).
- Recognize how significance levels (α) influence the risk of Type I errors.
- Understand how sample size and study design affect the likelihood of Type II errors.
- Learn strategies to manage and minimize the risks of both error types in hypothesis testing.

How do you quantify small? Is 5%, 10%, or 15% really small? How do you decide? That depends on your field of study and the importance of the situation. Is this a pilot study? Is someone's life at risk? Would you lose your job? Most industry standards use 5% as the cutoff point for how small is small enough, but 1%, 5%, and 10% are frequently used depending on what the situation calls for.

Now, how small is small enough? To answer that you want to know the types of errors you can make in hypothesis testing.

The first error is if you say that H_0 is false when in fact it is true. This means you reject H_0 when H_0 is true. The second error is if you say that H_0 is true when in fact it is false. This means you fail to reject H_0 when H_0 is false.

Figure 8-4 shows that if we "Reject H_0 " when H_0 is true, we are committing a Type I error. The probability of committing a Type I error is the Greek letter α, pronounced alpha. This can be controlled by the researcher by choosing a specific level of significance α.

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Decision
Do Not Reject H_0	Correct Decision	Type II Error

Figure 8.2.1: Matrix for Types of Errors

Figure 8-4 shows that if we "Do Not Reject H_0 " when H_0 is false, we are committing a Type II error. The probability of committing a Type II error is denoted by the Greek letter β, pronounced beta. When we increase the sample size, this will reduce β. The power of a test is 1 – β.

A jury trial is about to take place to decide if a person is guilty of committing murder. The hypotheses for this situation would be:

- H_0: The defendant is innocent
- H_1: The defendant is not innocent

The jury has two possible decisions to make: either acquit or convict the person on trial, based on the evidence that is presented. There are two possible ways that the jury could make a mistake. They could convict an innocent person or they could let a guilty person go free. Both are bad news, but if the death penalty is imposed on the convicted person, the justice system could be killing an innocent person. If a murderer is let go without enough evidence to convict them, then they could murder again. In statistics, we call these two types of mistakes Type I and Type II errors.

Figure 8-5 is a diagram showing the four possible jury decisions and two errors.

Jury's Decision	What really happened? H_0 True (the person is innocent)	H_0 False (the person is guilty)
Reject H_0 (Convict)	Type I Error α = P(Type I error) (Level of Significance)	Correct Decision $1 - \beta$ (Power)
Do Not Reject H_0 (Acquit)	Correct Decision $1 - \alpha$ (Confidence Level)	Type II Error β = P(Type II error)

Figure 8.2.2: Metaphorical Example of Court Case Results using Type I and II Errors

Definition: Type I and II Errors

Type I Error is rejecting H_0 when H_0 is true, and **Type II Error** is failing to reject H_0 when H_0 is false.

Since these are the only two possible errors, one can define the probabilities attached to each error.

α = P(Type I Error) = P(Rejecting H_0 | H_0 is true)

β = P(Type II Error) = P(Failing to reject H_0 | H_0 is false)

Example 8.2.1

An investment company wants to build a new food cart. They know from experience that food carts are successful if they have on average more than 100 people a day walk by the location. They have a potential site to build on, but before they begin, they want to see if they have enough foot traffic. They observe how many people walk by the site every day over a month. They will build if there is more than an average of 100 people who walk by the site each day. In simple terms, explain what the type I & II errors would be using context from the problem.

Solution

The hypotheses are: H_0: $\mu = 100$ and H_1: $\mu > 100$.

Sometimes it is helpful to use words next to your hypotheses instead of formal symbols

- H_0: $\mu \leq 100$ (Do not build)
- H_1: $\mu > 100$ (Build).

A Type I error would be to reject the null hypothesis when in fact it is true. Take your finger and cover up the null hypothesis (we decide to reject the null), then what is showing? The alternative hypothesis is what action we take.

If we reject H_0, then we will build the new food cart. However, H_0 was true, which means that the mean was less than or equal to 100 people walking by.

In simpler terms, this would mean that our evidence showed that we have enough foot traffic to support the food cart. Once we built, though, there were not, on average, more than 100 people who walked by, and the food cart may fail.

A type II error would be to fail to reject the null when in fact the null is false. Evidence shows that we should not build on the site, but this actually would have been a prime location to build on.

The missed opportunity of a type II error is not as bad as possibly losing thousands of dollars on a bad investment.

What is more severe of an error is dependent on what side of the desk you are sitting on. For instance, if a hypothesis is about miles per gallon for a new car the hypotheses may be set up differently depending on if you are buying the car or selling the car. For this course, the claim will be stated in the problem and always set up the hypotheses to match the stated claim. In general, the research question should be set up as some type of change in the alternative

Controlling for Type I Error

The significance level used by the researcher should be picked before the collection and analysis of data. This is called "a priori," versus picking α after you have done your analysis, which is called "post hoc." When deciding on what significance level to pick, one needs to look at the severity of the consequences of the type I and type II errors. For example, if the type I error may cause the loss of life or large amounts of money, the researcher would want to set α low.

Controlling for Type II Error

The power of a test is the complement of a Type II error or correctly rejecting a false null hypothesis. You can increase the power of the test and hence decrease the type II error by increasing the sample size. Similar to confidence intervals, we can reduce our margin of error when we increase the sample size. In general, we would like to have a high confidence level and a high power for our hypothesis tests. When you increase your confidence level, then in turn the power of the test will decrease. Calculating the probability of a Type II error is a little more difficult, and it is a conditional probability based on the researcher's hypotheses and is not discussed in this course.

8.2.1 Finding Critical Values

A researcher decides they want to have a 5% chance of making a Type I error, so they set α = 0.05. What z-score would represent that 5% area? It would depend on if the hypotheses were a left-tailed, two-tailed, or right-tailed test. This z-score is called a critical value. Figure 8-8 shows examples of critical values for the three possible sets of hypotheses.

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
		

Figure 8.2.3: Critical Regions for Different Types of Tests.

Two-tailed Test

If we are doing a two-tailed test then the α = 5% area gets divided into both tails. We denote these critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$. When the sample data finds a z-score (*test statistic*) that is either less than or equal to $z_{\alpha/2}$ or greater than or equal to $z_{1-\alpha/2}$ then we would reject H_0. The area to the left of the critical value $z_{\alpha/2}$ and to the right of the critical value $z_{1-\alpha/2}$ is called the critical or rejection region. See Figure 8-9.

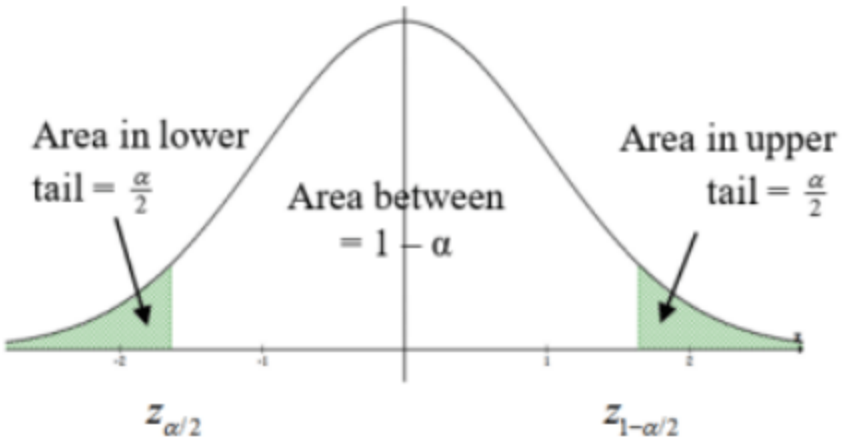


Figure 8.2.4: Standard Normal Distribution with Critical Regions and Critical Values.

When α = 0.05 then the critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are found using the following technology.

Excel: $z_{\alpha/2}$ =NORM.S.INV(0.025) = –1.96 and $z_{1-\alpha/2}$ =NORM.S.INV(0.975) = 1.96

TI-Calculator: $z_{\alpha/2}$ = invNorm(0.025,0,1) = –1.96 and $z_{1-\alpha/2}$ = invNorm(0.975,0,1) = 1.96

Since the normal distribution is symmetric, you only need to find one side's z-score and we usually represent the critical values as ± $z_{\alpha/2}$.

Most of the time, we will be finding a probability (p-value) instead of the critical values. The p-value and critical values are related and tell the same information, so it is important to know what a critical value represents.

Right-tailed Test

If we are doing a right-tailed test then the α = 5% area goes into the right tail. We denote this critical value $z_{1-\alpha}$. When the sample data finds a z-score more than $z_{1-\alpha}$, then we would reject H_0; reject H_0 if the test statistic is ≥ $z_{1-\alpha}$. The area to the right of the critical value $z_{1-\alpha}$ is called the critical region. See Figure 8-10.

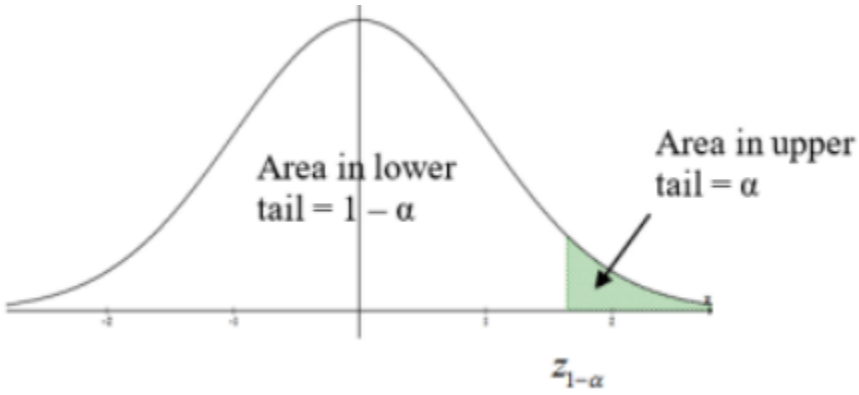


Figure 8.2.5: Standard Normal Distribution with a Right-Tailed Critical Region

When α = 0.05 then the critical value $z_{1-\alpha}$ is found using the following technology.

Excel: $z_{1-\alpha}$ =NORM.S.INV(0.95) = 1.645 Figure 8-10

TI-Calculator: $z_{1-\alpha}$ = invNorm(0.95,0,1) = 1.645

Left-tailed Test

If we are doing a left-tailed test then the α = 5% area goes into the left tail. If the sampling distribution is normal, then we can use the inverse normal function in Excel or a calculator to find the corresponding z-score. We denote this critical value z_α.

When the sample data finds a z-score less than z_α, then we would reject H0, and reject Ho if the test statistic is ≤ z_α. The area to the left of the critical value z_α is called the critical region. See Figure 8-11.

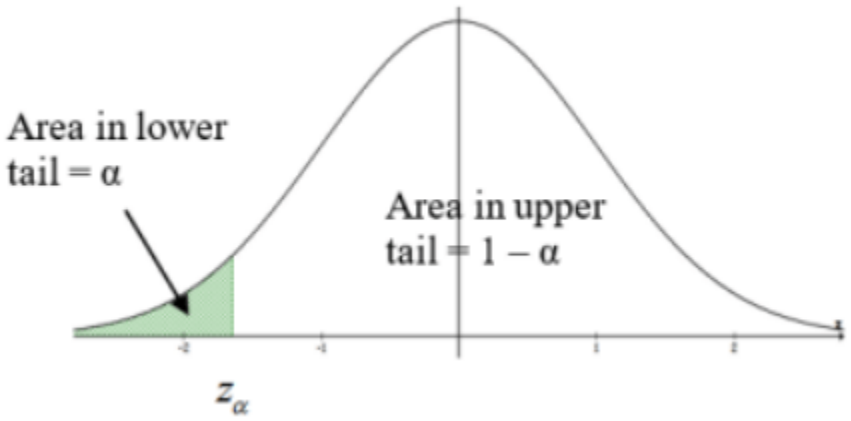


Figure 8.2.6: Standard Normal Distribution with a Left-Tailed Critical Region

When α = 0.05 then the critical value z_α is found using the following technology.

Excel: z_α =NORM.S.INV(0.05) = –1.645

TI-Calculator: z_α = invNorm(0.05,0,1) = –1.645

The Claim and Summary

The wording of the summary statement changes depending on which hypothesis the researcher claims to be true. We really should always be setting up the claim in the alternative hypothesis, since most of the time we are collecting evidence to show that a change has occurred, but occasionally a textbook will have the claim in the null hypothesis. **Do not use the phrase "accept H_0" since this implies that H_0 is true.** The lack of evidence is not evidence of anything.

> There were only two possible correct answers for the decision step.
>
> i. Reject H_0
>
> ii. Fail to reject H_0

Caution! If we fail to reject the null, this does not mean that there was no change; we just do not have any evidence that change has occurred. The absence of evidence is not evidence of absence. On the other hand, we need to be careful when we reject the null hypothesis, we have not proved that there is a change.

When we reject the null hypothesis, there is only evidence that a change has occurred. Our evidence could have been false and led to an incorrect decision. If we use the phrase, "accept H_0," this implies that H_0 was true, but we just do not have evidence that it is false. Hence, you will be marked incorrect for your decision if you use accept H_0, use instead "fail to reject H_0," or "do not reject H_0."

Authors

"8.2: Type I and II Errors" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.2: Type I and II Errors" by Rachel Webb is licensed under CC BY-SA 4.0

"9.E: Hypothesis Testing with One Sample (Exercises)" by OpenStax is licensed under CC BY 4.0

This page titled 8.2: Type I and II Errors is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: Type I and II Errors by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

8.3: z-Test for a Mean

Learning Objectives

- Understand when to use a Z-test in hypothesis testing.
- Recognize that a z-test is appropriate when the population standard deviation is known and the sample size is large or the population is normally distributed.
- Learn to apply the critical value method by comparing the test statistic to the critical cutoff value.
- Learn to apply the p-value method by comparing the p-value to the significance level (α) to make a decision.
- Develop the ability to interpret results using both the critical value and p-value methods.

The z-test for a mean is a statistical method used to determine whether a sample mean significantly differs from a known population mean. It is typically used when the population standard deviation is known and the sample size is sufficiently large (usually $n \geq 30$). The test calculates a z-score, which measures how many standard deviations the sample mean is from the population mean under the null hypothesis. This z-score is then compared to a critical value based on the chosen significance level to decide whether to reject the null hypothesis. The z-test helps researchers make data-driven conclusions about population means.

The Traditional Method (Critical Value Method)

There are five steps in hypothesis testing when using the traditional method:

1. Identify the claim and formulate the hypotheses.
2. Find the critical value(s) and state the rejection rule (the rule by which you will reject the null hypothesis (H_0).
3. Compute the test statistic.
4. Decide to reject or not reject the null hypothesis by comparing the test statistic to the critical value(s). Reject H_0 when the test statistic is in the critical tail(s).
5. Summarize the results and address the claim using context and units from the research question.

Note: The test statistic and the critical value(s) come from the same distribution and will usually have the same letter, such as z, t, or F. The critical value(s) will have a subscript with the lower tail area $(z_\alpha, z_{1-\alpha}, z_{\alpha/2})$ or an asterisk next to it (z*) to distinguish it from the test statistic.

You can find the critical value(s) or test statistic in any order, but make sure you know the difference when you compare the two. The critical value is found from α and is the start of the shaded area called the critical region (also called the rejection region or area). The test statistic is computed using sample data and may or may not be in the critical region.

The critical value(s) is set before you begin (a priori) by the level of significance you are using for your test. This critical value(s) defines the shaded area known as the rejection area. The test statistic for this example is the z-score we find using the sample data that is then compared to the shaded tail(s). When the test statistic is in the shaded rejection area, you reject the null hypothesis. When your test statistic is not in the shaded rejection area, then you fail to reject the null hypothesis. Depending on whether your claim is null or the alternative, the sample data may or may not support your claim.

8.3.1 z-Test

When the population standard deviation is known and stated in the problem, we will use the **z-test**.

Definition: z-Test When Population Standard Deviation Is Given

The z-test is a statistical test for the mean of a population. It can be used when σ is known. The population should be approximately normally distributed when n < 30.

When using this model, the *test statistic* is $z = \dfrac{\bar{x} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$ where μ is the test value from the H_0.

Example 8.3.1 z-Test Using Traditional (Critical Value) Method

M&Ms candies advertise a mean weight of 0.8535 grams. A sample of 50 M&M candies is randomly selected from a bag of M&Ms and the mean is found to be $\overline{x}$ = 0.8472 grams. The standard deviation of the weights of all M&Ms is (somehow) known to be σ = 0.06 grams. A skeptic M&M consumer claims that the mean weight is less than what is advertised. Test this claim using the **traditional method** of hypothesis testing. Use a 5% level of significance.

Solution

By letting α = 0.05, we are allowing a 5% chance that the null hypothesis (average weight that is at least 0.8535 grams) is rejected when in actuality it is true.

1. **State the Hypothesis and Identify the Claim:** The claim is "M&Ms candies have a mean weight that is less than 0.8535 grams." This translates mathematically to μ < 0.8535 grams. Therefore, the null and alternative hypotheses are:

H_0: μ = 0.8535

H_1: μ < 0.8535 (claim)

This is a left-tailed test since the alternative hypothesis has a "less than" sign.

We are performing a test of the population mean. We can use the z-test because we were given a population standard deviation σ (not a sample standard deviation s). In practice, σ is rarely known and usually comes from a similar study or previous year's data.

2. **Find the Critical Value:** The critical value for a left-tailed test with α = 0.05 can be found by using the Standard Normal Distribution Table. Look up α = 0.05 in the one-tail row, and locate the critical value at the very bottom of the table in the z-row. Since the test is left-tailed, the critical value will be a negative value from what appears in the table. It is –1.65.

Table F – t-values for specific one or tailed distributions.

	Confidence Intervals	80%	90%	95%	98%	99%
	1- tail	0.10	0.05	0.025	0.01	0.005
df	2-tails	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.665	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
Z-value		1.28	1.65	1.96	2.33	2.58

Figure 8.3.1: The Table for Z and T Values

3. **Compute the Test Statistic:** The formula for the test statistic is the z-score that we used back in the Central Limit Theorem section $z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{0.8472 - 0.8535}{\left(\frac{0.06}{\sqrt{50}}\right)} = -0.74$.

4. **Make the Decision:** The test point –0.74 falls in the non-critical region. Therefore, do not reject the null hypothesis (H_0).

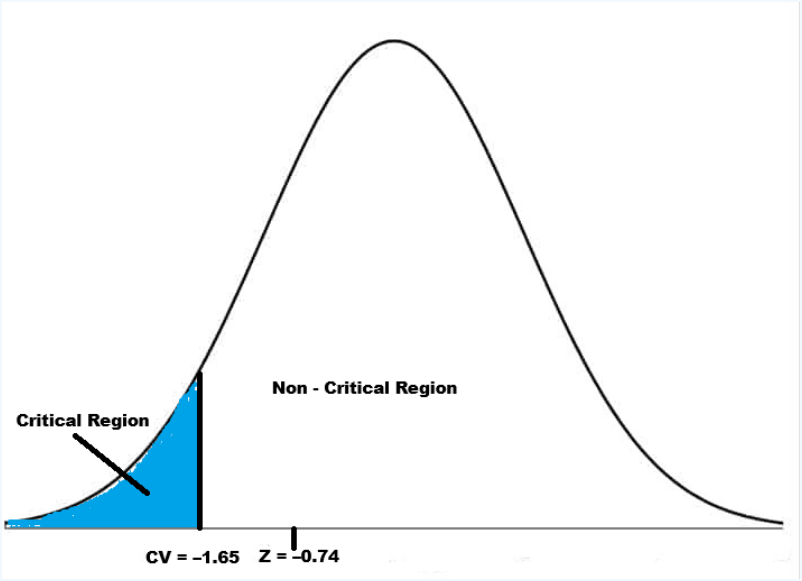


Figure 8.3.2: Standard Normal Distribution Graph. C.V. line is -1.65, and the Test Value z is -0.74 and Lands in the Non-Critical Region

5. **Summarize the Results:** At α = 0.05, there is not enough evidence to support the claim that the mean weight is less than 0.8535 grams.

Example 8-5 used the traditional critical value method. With the onset of computers, this method is outdated, and the p-value and confidence interval methods are becoming more popular.

Most statistical software packages will give a p-value and confidence interval, but not the critical value.

TI-84: Press the **[STAT]** key, go to the **[TESTS]** menu, arrow down to the **[Z-Test]** option, and press the **[ENTER]** key. Arrow over to the **[Stats]** menu and press the **[ENTER]** key. Then type in the value for the hypothesized mean (μ_0), standard deviation, sample mean, sample size, arrow over to the $\neq$, <, > sign that is in the alternative hypothesis statement then press the **[ENTER]** key, arrow down to **[Calculate]** and press the **[ENTER]** key. *Alternatively* (If you have raw data in a list) Select the **[Data]** menu and press the **[ENTER]** key. Then type in the value for the hypothesized mean (μ_0), and type in your list name (TI-84 L_1 is above the 1 key).

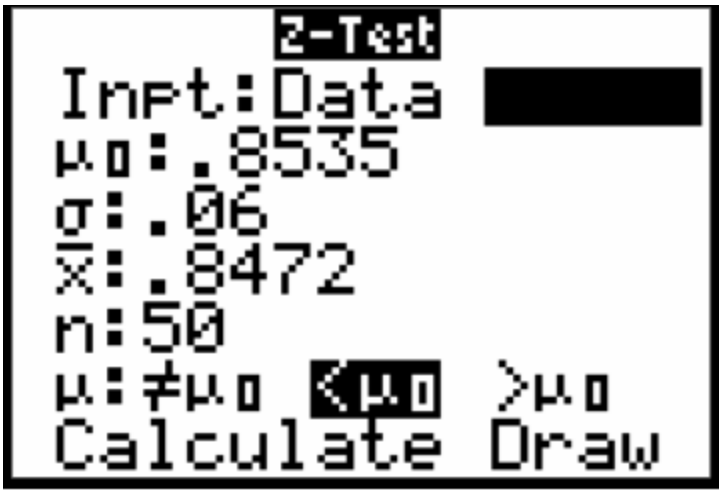


Figure 8.3.3: Input for z-Test in TI-84+ Calculator.

Press the **[STAT]** key, go to the **[TESTS]** menu, arrow down to either the **[Z-Test]** option, and press the **[ENTER]** key. Arrow over to the **[Stats]** menu and press the **[ENTER]** key. Then type in the value for the hypothesized mean (μ_0), standard deviation, sample mean, sample size, arrow over to the $\neq$, <, > sign that is in the alternative hypothesis statement then press the **[ENTER]** key, arrow down to **[Calculate]** and press the **[ENTER]** key. Alternatively, if you have raw data in a list) select the **[Data]** menu and press the **[ENTER]** key. Then type in the value for the hypothesized mean (μ_0), and type in your list name (TI-84 L_1 is above the 1 key).

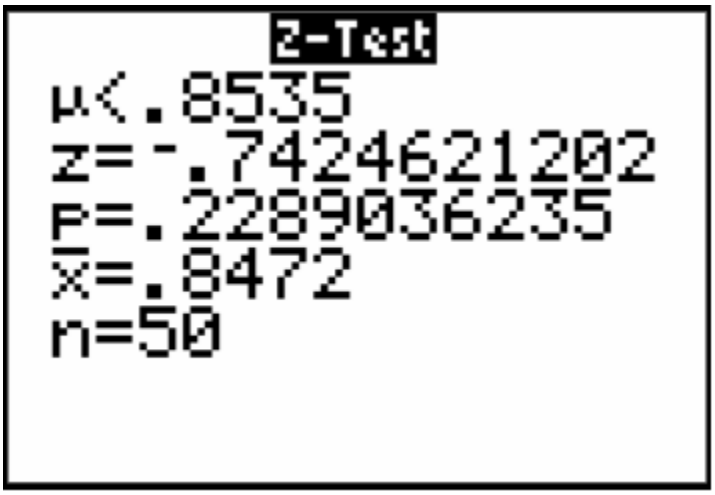


Figure 8.3.4: Output for z-Test in TI-84+ Calculator.

The calculator returns the alternative hypothesis (check and make sure you selected the correct sign), the test statistic, the p-value, the sample mean, and the sample size.

Example 8.3.2 z-Test Using a Table (Traditional Method)

An educational researcher wants to determine if the average cost of textbooks for Citrus College students is more than $500 per semester. A random sample of 36 students is selected, and their average textbook cost is found to be $520, with a population standard deviation of $60. Test the researcher's claim at α = 0.05.

Solution

1. **State the Hypothesis and Identify the Claim:** "The average cost of textbooks for Citrus College students is more than $500 per semester." This translates mathematically to μ > $500. Therefore, the null and alternative hypotheses are:

H_0: μ = $500

H_1: μ > $500 (claim)

This is a right-tailed test since the alternative hypothesis has a “greater than” sign.

We are performing a test of the population mean. We can use the z-test because we were given a population standard deviation σ (not a sample standard deviation s). In practice, σ is rarely known and usually comes from a similar study or the previous year’s data.

2. **Find the Critical Value:** To find the critical value for a right-tailed test, use the table below.

Table F – t-values for specific one or tailed distributions.

	Confidence Intervals	80%	90%	95%	98%	99%
	1- tail	0.10	0.05	0.025	0.01	0.005
df	2-tails	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.666	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
Z value		1.28	1.65	1.96	2.33	2.58

Figure 8.3.8: The Table for Z and T Values

Since α = 0.05 and the test is right-tailed, the critical value will be positive, and it will be 1.65.

3. **Compute the Test Statistic:** The formula for the test statistic is the z-score that we used back in the Central Limit Theorem section $z = \dfrac{\bar{x} - \mu_0}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} = \dfrac{520 - 500}{\left(\dfrac{60}{\sqrt{36}}\right)} = 2.$

4. **Make the Decision:** Since the test point of 2.00 falls in the critical region, reject the null hypothesis (H_0).

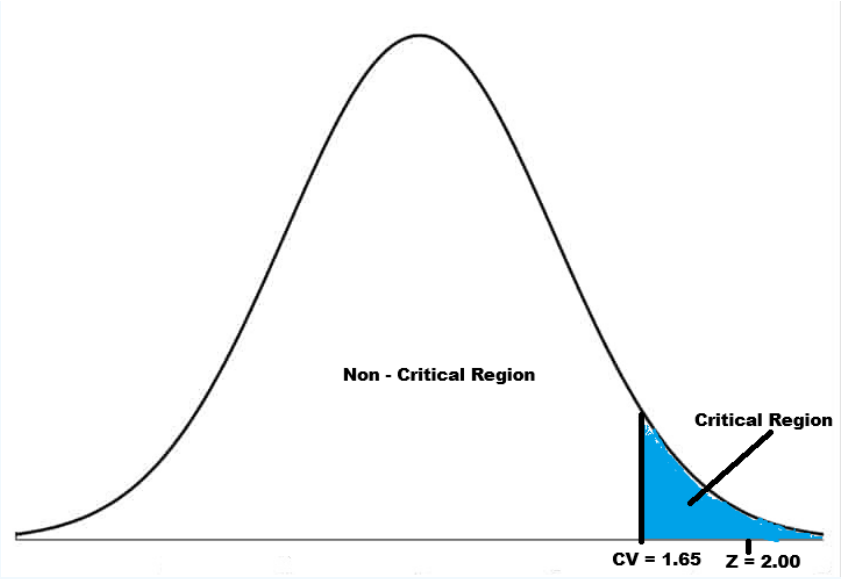


Figure 8.3.9: Right-Tailed Standard Normal Distribution Curve Where C.V. =1.65 and Test Value z =2.00 Lands in the Critical Region.

5. **Summarize the Results:** There is enough evidence to support the claim that the average cost of textbooks for Citrus College students is more than $500 per semester.

The P-Value Method

Most modern statistics and research methods utilize this method with the advent of computers and graphing calculators.

There are five steps in hypothesis testing when using the p-value method:

i. Identify the claim and formulate the hypotheses.
ii. Compute the test statistic.
iii. Compute the p-value.
iv. Decide to reject or not reject the null hypothesis by comparing the p-value with α. Reject H_0 when the p-value ≤ α.
v. Summarize the results and address the claim.

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors the alternative hypothesis.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative hypothesis.
- The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether to reject H_0. We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language and use context and units so non-statisticians can also understand the results.

What is the p-value?

The p-value is the probability of observing an effect as extreme as that in your sample data, assuming that the null hypothesis is true. The p-value is calculated based on the assumption that the null hypothesis is true for the population and that the difference in the sample is caused entirely by random chance.

Recall the example at the beginning of the chapter.

Example 8.3.3 z-Test Using TI-84 Graphing Calculator (P-Value Method)

Suppose a manufacturer of a new laptop battery claims the mean life of the battery is 900 days with a standard deviation of 40 days. You are the buyer of this battery, and you think this claim is inflated. You would like to test your belief because, without a good reason, you cannot get out of your contract. You take a random sample of 35 batteries and find that the mean battery life is 890 days. Test the claim using the p-value method. Let α = 0.05.

Solution

We had the following hypotheses:

1. **State the Hypothesis and Identify the Claim:** State the claim and determine H_0 and H_1.

   H_0: μ = 900, since the manufacturer says the mean life of a battery is 900 days.

   H_1: μ < 900, since you believe the mean life of the battery is less than 900 days.

2. **Compute the Test Statistic:** Compute the test statistic using sample data.

The test statistic was found to be: $z = \dfrac{\bar{x} - \mu_0}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} = \dfrac{890 - 900}{\left(\dfrac{40}{\sqrt{35}}\right)} = -1.479.$

3. **Find the P-Value:** The p-value is computed using the tail area of the test point. If the test has one tail, as in this example, the tail area equals the p-value.

The p-value is P($\overline{x}$ < 890 | H_0 is true) = P($\overline{x}$< 890 | μ = 900) = P(Z < –1.48).

Looking up the left tail area for z = –1.48 in the standard normal distribution is 0.0694. Thus, the p-value = 0.0694.

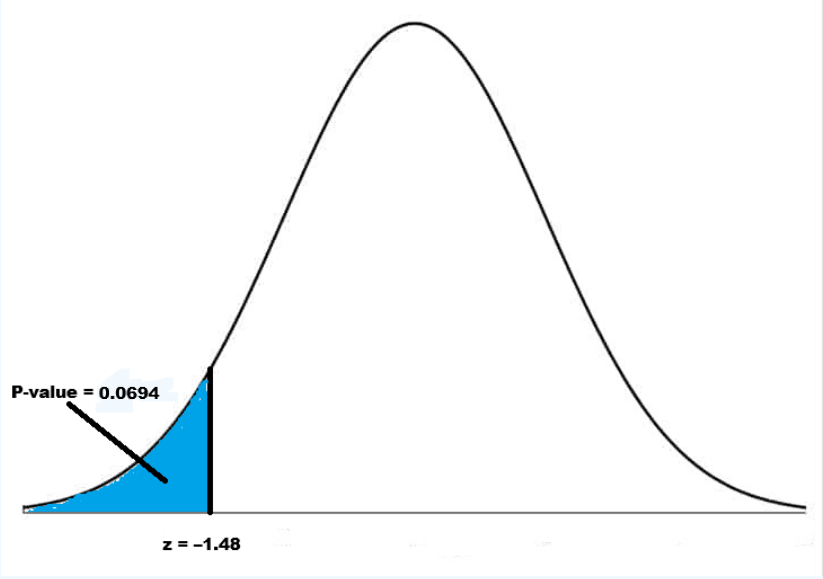


Figure 8.3.5: The Test Point with z = -1.48 and Left Tail Area of 0.0694.

Alternatively, the p-value can be computed using a TI-84+ graphing calculator. Input the data using the steps mentioned in the previous example.

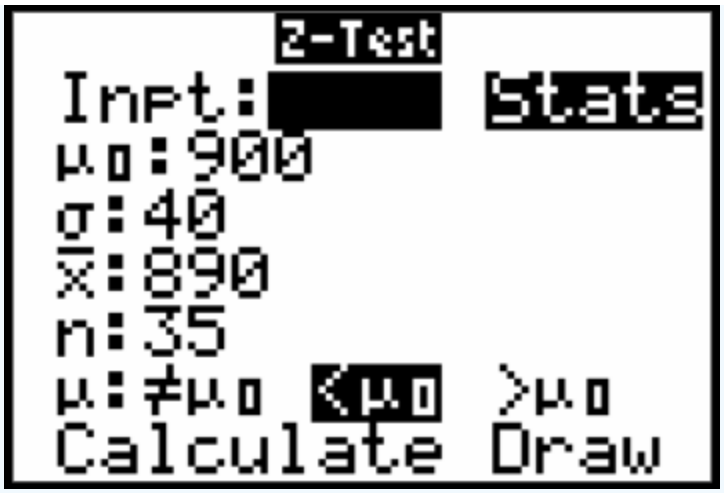


Figure 8.3.6: Input for z-Test in TI-84+ Calculator.

The TI calculators will easily find the p-value for you. This value may differ slightly from the table due to rounding on the z-value.

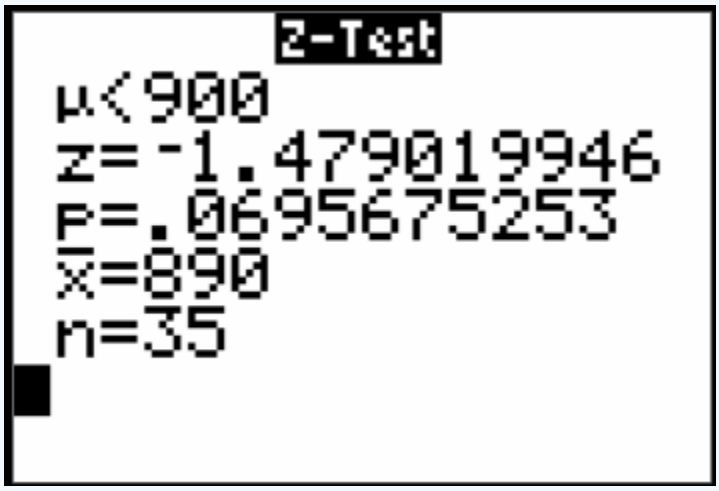


Figure 8.3.7: Output for z-Test in TI-84+ Calculator.

4. **Make the Decision:** Reject or not reject the null hypothesis by comparing the p-value with α. Now compare the p-value = 0.0696 (TI-84+) or 0.0694 (From Standard Normal Distribution Table) to α = 0.05. Reject H_0 when the p-value ≤ α, and do not reject H_0 when the p-value > α. The p-value for this example is larger than alpha 0.0696 > 0.05, therefore, the decision is not to reject H_0.
5. **Summarize the Results:** Since the null hypothesis is not rejected, there is not enough evidence to indicate that the mean life of the battery is less than 900 days.

Authors

"8.3: z-Test for a Mean" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.3: Hypothesis Test for One Mean" by Rachel Webb is licensed under CC BY-SA 4.0

"9.E: Hypothesis Testing with One Sample (Exercises)" by OpenStax is licensed under CC BY 4.0

This page titled 8.3: z-Test for a Mean is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.3: Hypothesis Test for One Mean by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

8.4: t-Test for a Mean

Learning Objectives

- Understand when to use a t-test in hypothesis testing.
- Recognize that a t-test is appropriate when the population standard deviation is unknown and the sample size is small.
- Learn that the T-test relies on the T-distribution to account for additional variability in small samples.
- Apply the critical value method by comparing the test statistic to a t-value from the t-distribution.
- Use the result of the comparison to decide whether to reject the null hypothesis.

t-Test for the Mean

The one-sample t-test is a statistical method used to determine whether the average of a single sample is significantly different from a known or assumed population mean. It's especially useful when the population standard deviation is unknown and the sample size is relatively small.

This test is commonly applied in a variety of fields. For example, in education, researchers might test whether the average test score of a new teaching method differs from the national average. In healthcare, a doctor may want to check if a new treatment changes average blood pressure levels compared to a standard value. In manufacturing, a company might use the t-test to see if the average weight of a product batch matches the target specification.

There are two main approaches to making a decision with the t-test:

- **The traditional method** compares the calculated test statistic to a critical value from the t-distribution.
- **The p-value method** assesses how likely the observed result would be if the null hypothesis were true.

Both approaches lead to the same conclusion—either rejecting or not rejecting the null hypothesis based on the evidence provided by the sample.

When the population standard deviation is unknown, we will use the **t-test**.

Definition: t-Test for the Mean

The t-test is a statistical test for the mean of a population. It will be used when σ is unknown. The population should be approximately normally distributed when n < 30.

When using this model, the *test statistic* is $t = \dfrac{\bar{X} - \mu}{\left(\dfrac{s}{\sqrt{n}}\right)}$ where μ is the test value from the H_0. The degrees of freedom are *d.f.* = *n* – 1.

The z and t-tests are easy to mix up. Sometimes, a standard deviation will be stated in the problem without specifying if it is a population's standard deviation σ or the sample standard deviation s. If the standard deviation is in the same sentence that describes the sample or only raw data is given, then this would be s. When you only have sample data, use the t-test.

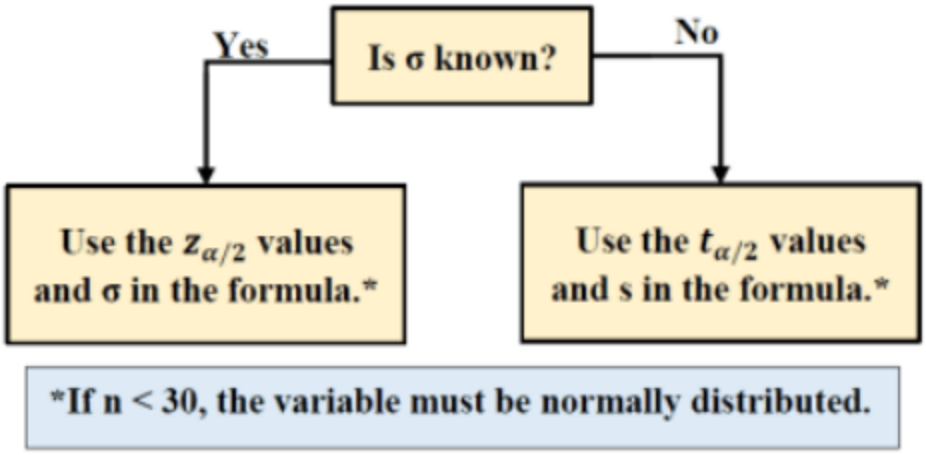


Figure 8.4.1: Flow Chart that demonstrates when to use the z or t-test.

Use the figure below as a guide in setting up your hypotheses. The two-tailed test will always have a not equal ≠ sign in H_1 and both tails shaded. The right-tailed test will always have the greater than > sign in H_1 and the right tail shaded. The left-tailed test will always have a less than < sign in H_1 and the left tail shaded.

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$
		
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 8.4.2: Types of Test Based on Hypothesis Test and Key Words.

Example 8.4.1 Traditional t-test

A real estate researcher is analyzing data on frequent movers. The researcher believes that people are becoming less mobile and settling down earlier in life. She claims that the average person used to live in is now less than 6 different homes by the age of 30. But she suspects that number is dropping.

To test this, she randomly surveyed 15 people under the age of 30 and recorded the number of different homes each person had lived in:

Data (number of homes):
4, 5, 3, 6, 5, 4, 5, 6, 3, 4, 4, 5, 4, 3, 4

Test the researcher's claim using $\alpha = 0.05$.

Solution

1. State the hypotheses and identify the claim: The statement “the average is less than (<) than 6” must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

   H_0: μ = 6

   H_1: μ < 6 (claim)

   This is a left-tailed test with the claim in the alternative hypothesis.

2. Find the critical value using the t-distribution table. To find the critical value, use the degrees of freedom and $\alpha = 0.05$..

   The degrees of freedom is d.f. = 15 - 1 = 14. There are 15 data values.

   The critical value is –1.761.

3. Compute the test statistic: We are using the t-test because we are performing a test about a population mean. We must use the t-test (instead of the z-test) because the population standard deviation σ is unknown. Also, we must compute the sample mean and sample standard deviation using the data.

   Using a calculator to compute $\bar{X}$ = 4.333 and s = 0.961 after being rounded to three place values.

   $$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{4.333 - 6}{\left(\frac{0.961}{\sqrt{15}}\right)} = -6.73$$

4. State the decision. Decision: Since the test statistic t =-6.73 is in the critical region, we should reject H_0.

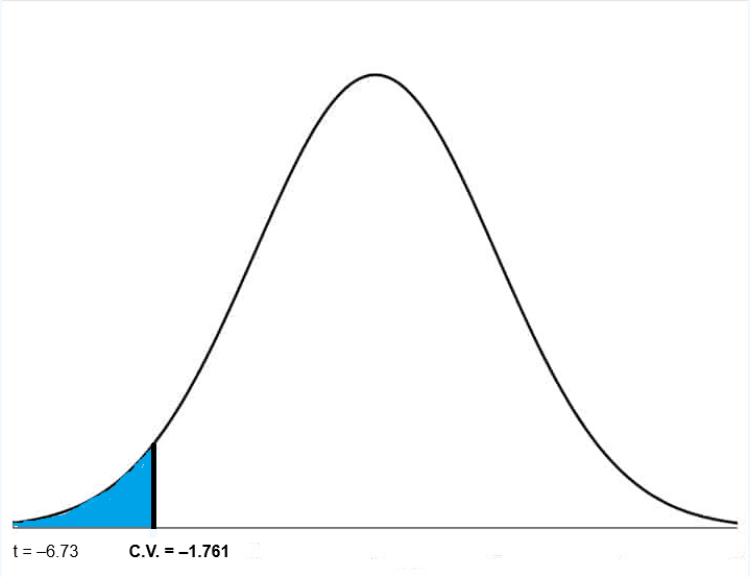


Figure 8.4.3: The t-value = –6.73 Falls in the Critical Region.

5. State the summary. There is enough evidence to support the researcher's claim that the average number of homes people have lived in is now less than 6.

Example 8.4.2 Traditional t-test

A career research analyst claims that recent graduates in STEM (Science, Technology, Engineering, and Mathematics) fields earn an average starting income of $68,000. However, a university career center believes that this figure might be off and decides to conduct an investigation. They randomly surveyed a sample of 25 recent STEM graduates from their university. The results show that the average income of the sample is $70,200, with a sample standard deviation of $6,000. At $\alpha = 0.10$, test the career center's belief that the average starting income differs from $68,000. Use the traditional method.

Solution

1. **State the Hypotheses and Identify the Claim:** The statement "the income differs (≠) than $68,000" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

   H_0: μ = 68,000

   H_1: μ ≠ 68,000 (claim)

   This is a two-tailed test with the claim in the alternative hypothesis.

2. **Find the Critical Values:** Use the t-distribution table to find the critical value with the degrees of freedom and $\alpha = 0.10$.

   The degrees of freedom is d.f. = 25 - 1 = 24. There are 25 data values.

   The critical value is ±1.711.

3. **Compute the Test Statistic:** We are using the t-test because we are performing a test about a population mean. We must use the t-test (instead of the z-test) because the population standard deviation σ is unknown. Also, we must compute the sample mean and sample standard deviation using the data.

   Using a calculator to compute $\bar{X}$ = 70,200 and s = 6,000.

   $$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{70,200 - 68,000}{\left(\frac{6,000}{\sqrt{25}}\right)} = 1.83$$

4. **Make the Decision:** Since the test statistic t =1.83 is in the critical region, we should reject H_0.

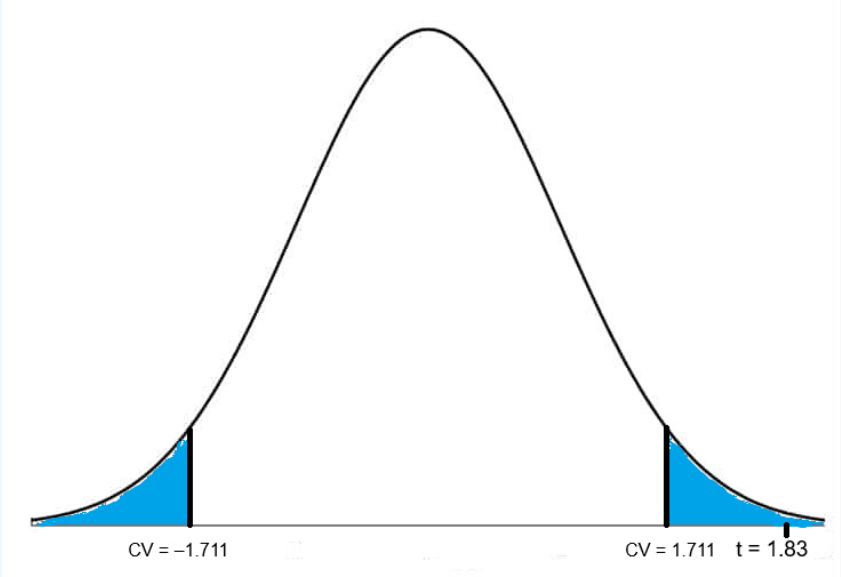


Figure 8.4.4: The t-value = 1.83 Falls in the Critical Region.

5. **Summarize the Results:** There is enough evidence to support the analyst's claim that the average income differs from $68,000.

Example 8.4.3 Traditional t-test

With California's housing market known for its sky-high prices, a housing advocacy group claims that the average monthly rent statewide is $2,300. However, a skeptical tenant rights organization believes that rent may be climbing even higher than reported. To test this belief, the organization surveyed a random sample of 20 California renters. The survey results show that the average monthly rent paid by the sample is $2,480, with a sample standard deviation of $300. At $\alpha = 0.01$, test the organization's claim that the average rent is greater than $2,300.

Solution

1. **State the Hypotheses and Identify the Claim:** The statement "the rent is greater than (>) $2,300" must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

   H_0: μ = 2,300

   H_1: μ > 2,300 (claim)

   This is a right-tailed test with the claim in the alternative hypothesis.

2. **Find the Critical Values:** Use the t-distribution table to find the critical value with the degrees of freedom and $\alpha = 0.01$. .

   The degrees of freedom is d.f. = 20 - 1 = 19. There are 19 data values.

   The critical value is 2.539.

3. **Compute the Test Statistic:** We are using the t-test because we are performing a test about a population mean. We must use the t-test (instead of the z-test) because the population standard deviation σ is unknown. Also, we must compute the sample mean and sample standard deviation using the data.

   Using a calculator to compute $\bar{X}$ = 2,480 and s = 300.

   $$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{2,480 - 62,300}{\left(\frac{300}{\sqrt{20}}\right)} = 2.68$$

4. **Make the Decision:** Since the test statistic t =2.68 is in the critical region, we should reject H_0.

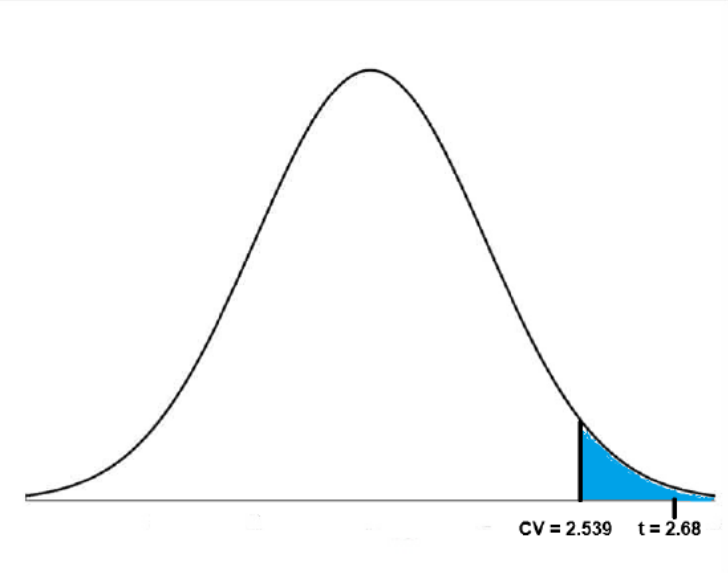


Figure 8.4.5: The t-value = 2.68 Falls in the Critical Region.

5. **Summarize the Results:** There is enough evidence to support the organization's claim that the average monthly rent in California is greater than $2,300.

P-Value Method

Example 8.4.5 t-Test: P-value Method

The label on a particular brand of cream of mushroom soup states that (on average) there is 870 mg of sodium per serving. A nutritionist would like to test if the average is more than the stated value. To test this, 13 servings of this soup were randomly selected, and the amount of sodium was measured. The sample mean was found to be 882.4 mg, and the sample standard deviation was 24.3 mg. Assume that the amount of sodium per serving is normally distributed. Test this claim using the traditional method of hypothesis testing. Use the α = 0.05 level of significance.

Solution

1. **State the Hypotheses and Identify the Claim:** The statement “the average is more (>) than 870” must be in the alternative hypothesis. Therefore, the null and alternative hypotheses are:

   H_0: μ = 870

   H_1: μ > 870 (claim)

   This is a right-tailed test with the claim in the alternative hypothesis.

2. **Compute the Test Statistic:** The test statistic is computed as follows, $t = \dfrac{\bar{x} - \mu_0}{\left(\dfrac{S}{\sqrt{n}}\right)} = \dfrac{882.4 - 870}{\left(\dfrac{24.3}{\sqrt{13}}\right)} = 1.8399.$

3. **Find the P-Value:**

We can use technology to get the test statistic and p-value.

TI-84: If you have raw data, enter the data into a list before you go to the test menu. Press the **[STAT]** key, arrow over to the **[TESTS]** menu, arrow down to the **[2:T-Test]** option, and press the **[ENTER]** key. Arrow over to the [Stats] menu and press the **[ENTER]** key. Then type in the hypothesized mean (μ), sample or population standard deviation, the sample mean, sample size, arrow over to the ≠, <, > sign that is the same as the problem's alternative hypothesis statement then press the **[ENTER]** key, arrow down to **[Calculate]** and press the **[ENTER]** key. The calculator returns the t-test statistic and p-value.

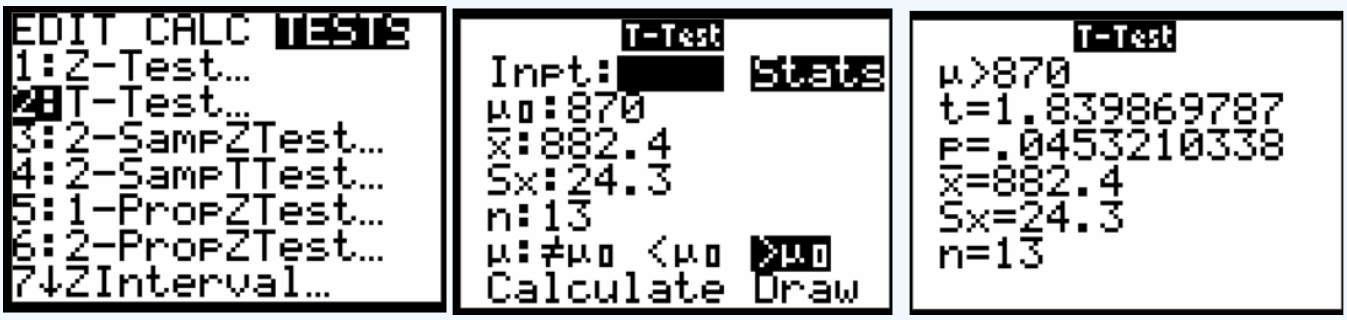


Figure 8.4.6: Steps to Enter Data into TI-84+ for t-test and Output.

Alternatively (If you have raw data in list one) Arrow over to the **[Data]** menu and press the **[ENTER]** key. Then type in the hypothesized mean (μ), L_1, leave Freq:1 alone, arrow over to the ≠, <, > sign that is the same in the problem's alternative hypothesis statement then press the **[ENTER]** key, arrow down to **[Calculate]** and press the **[ENTER]** key. The calculator returns the t-test statistic and the p-value.

4. **Make the Decision:** The rejection rule: reject the null hypothesis if the p-value ≤ α. Decision: Since the p-value = 0.0453 is less than α = 0.05, we Reject H_0. This agrees with the decision from the traditional method. (These two methods should always agree!)
5. **Summarize the Results:** The summary remains the same as in the previous method. There is sufficient evidence to support the claim that the average amount of sodium per serving of cream of mushroom soup exceeds the stated 870 mg amount.

Authors

"8.4: t-Test for a Mean" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.3: Hypothesis Test for One Mean" by Rachel Webb is licensed under CC BY-SA 4.0

"9.E: Hypothesis Testing with One Sample (Exercises)" by OpenStax is licensed under CC BY 4.0

This page titled 8.4: t-Test for a Mean is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

- 8.3: Hypothesis Test for One Mean by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

8.5: z-Test for a Proportion

Learning Objectives

- Understand when to use a Z-test for proportions in hypothesis testing.
- Recognize that a z-test for proportions is appropriate with a sufficiently large sample size to assume normality.
- Learn to compare the sample proportion to the claimed population proportion.
- Apply the critical value method by comparing the test statistic to a z-value cutoff.
- Use the comparison to determine whether to reject the null hypothesis.

Hypothesis testing for a proportion is a statistical method used to determine whether a population proportion differs significantly from a claimed or known value. It involves setting up a null hypothesis, which typically states that the population proportion equals a specific value, and an alternative hypothesis that suggests a difference. Using sample data, a test statistic (usually a z-score) is calculated and compared to a critical value or used to find a p-value. This process helps assess whether the observed sample proportion provides enough evidence to reject the null hypothesis in favor of the alternative.

When you read a question, it is essential that you correctly identify the parameter of interest. The parameter determines which model to use. Make sure you can distinguish between a question regarding a population mean and a question regarding a population proportion.

The normal distribution can be used for a z-test for a proportion because, under certain conditions, the sampling distribution of the sample proportion approximates a normal distribution. This is based on the Central Limit Theorem, which states that as the sample size becomes large, the distribution of the sample proportion becomes approximately normal, even if the population distribution is not.

The z-test is a statistical test for a population proportion. It can be used when $np \geq 5$ and $nq \geq 5$.

Definition: z-Test for Proportion

The formula for the *test statistic* is:

$$Z = \frac{\hat{p} - p}{\sqrt{\left(\frac{p \cdot q}{n}\right)}} \tag{8.5.1}$$

where

n is the sample size

$\hat{p} = \frac{x}{n}$ is the sample proportion (sometimes already given as a %) and

p is the hypothesized population proportion,

$$q = 1 - p.$$

Use the phrases in Figure 8-24 to help with setting up the hypotheses.

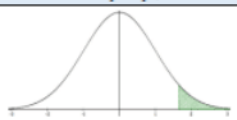
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: p = p_0$ $H_1: p \neq p_0$	$H_0: p = p_0$ $H_1: p > p_0$	$H_0: p = p_0$ $H_1: p < p_0$
		
Claim is in the Null Hypothesis		
=	≤	≥
Is equal to	Is less than or equal to	Is greater than or equal to
Is exactly the same as	Is at most	Is at least
Has not changed from	Is not more than	Is not less than
Is the same as	Within	Is more than or equal to
Claim is in the Alternative Hypothesis		
≠	>	<
Is not	More than	Less than
Is not equal to	Greater than	Below
Is different from	Above	Lower than
Has changed from	Higher than	Shorter than
Is not the same as	Longer than	Smaller than
	Bigger than	Decreased
	Increased	Reduced

Figure 8.5.1: Types of Tests Based on Hypothesis and Keywords.

Note that the t-distribution is not used with proportions. We will use a standard normal z distribution for testing a proportion since this test uses the normal approximation to the binomial distribution (never use the t-distribution).

If you are doing a left-tailed z-test, the critical value will be negative. If you are performing a right-tailed z-test, the critical value will be positive. If you were performing a two-tailed z-test, then your critical values would be ±critical value. The p-value will always be a positive number between 0 and 1. The most important step in any method you use is setting up your null and alternative hypotheses. The critical values and p-value can be found using a standard normal distribution in the same way that we did for the one-sample z-test.

Example 8.5.1 Traditional (Critical Value) Method

It has been found that 85.6% of all enrolled college and university students in the United States are undergraduates. A random sample of 500 enrolled college students in a particular state revealed that 420 of them were undergraduates. Is there sufficient evidence to conclude that the proportion differs from the national percentage? Use α = 0.05. Show that all three methods of hypothesis testing yield the same results.

Solution

At this point, you should be more comfortable with the steps of a hypothesis test and not have to number each step, but know what each step means.

1. **State the Hypothesis and Identify the Claim:** The keywords in this example, “proportion” and “differs,” lead to the following hypotheses and claim.

   H_0: p = 0.856

   H_1: p ≠ 0.856 (claim)

2. **Find the Critical Value:** Look up the critical value in the table.

Use α = 0.05 and look up the critical values in the table to get ± 1.96.

3. **Compute the Test Statistic:** Before computing the test statistic, compute the sample proportion $\hat{p} = \frac{420}{500} = 0.84$ and q = 1 – 0.856 = 0.144.

$$z = \frac{\hat{p} - p}{\sqrt{\left(\frac{p \cdot q}{n}\right)}} = \frac{0.84 - 0.856}{\sqrt{\left(\frac{0.856 \cdot 0.144}{500}\right)}} = -1.02. \tag{8.5.2}$$

4. **Make the Decision:** Draw and label the curve with the critical values and make the decision to reject or not reject the null hypothesis (H_0).

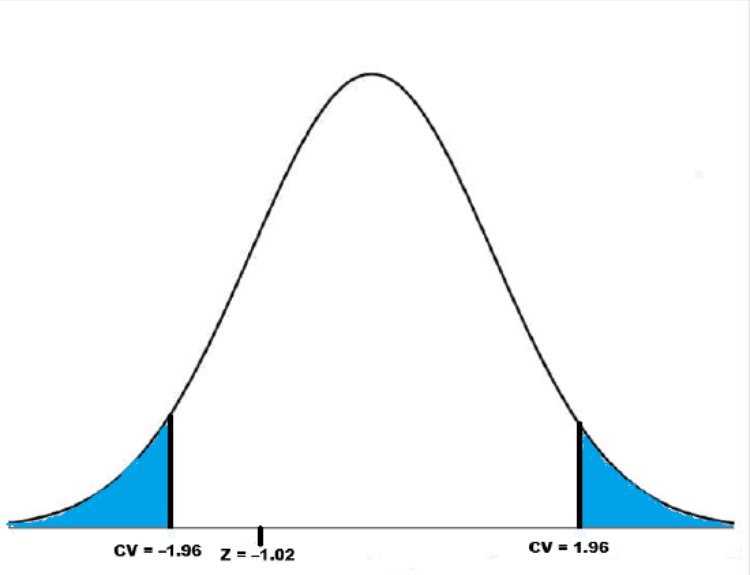


Figure 8.5.2: Standard Normal Distribution with Critical Values Plus and Minus 1.96 and Test Value of –1.02.

State the decision. Since the test statistic is not in the shaded rejection area, do not reject H_0.

5. **Summarize the Results:** There is not enough evidence to conclude that the proportion of undergraduates in college for this state differs from the national average of 85.6%.

Example 8.5.2 Traditional (Critical Value) Method

A California Community College claims that at least 30% of its students major in STEM fields. A student researcher believes that the proportion is greater than. To test this the researcher's, the researcher surveys a random sample of 120 students and finds that 45 of them are majoring in STEM.

At $\alpha = 0.10$, test the claim that the proportion of STEM majors at the college is greater than 30%.

Solution

1. **State the Hypothesis and Identify the Claim:** The keywords in this example, “proportion” and “differs,” ead to the following hypotheses and claim.

   H_0: p = 0.3

   H_1: p > 0.3 (claim)

2. **Find the Critical Value:** Look up the critical value in the table.

Use α = 0.10 and look up the critical values in the table to get 1.28.

3. **Compute the Test Statistic:**

The sample proportions are $\hat{p} = \frac{45}{120} = 0.375$ and q = 1 – p = 1 – 0.3 = 0.7

$$z = \frac{\hat{p} - p}{\sqrt{\left(\frac{p \cdot q}{n}\right)}} = \frac{0.375 - 0.3}{\sqrt{\left(\frac{0.3 \cdot 0.7}{120}\right)}} = 1.79. \tag{8.5.3}$$

4. **Make the Decision:** Draw and label the curve with the critical value and make the decision to reject or not reject the null hypothesis (H_0).

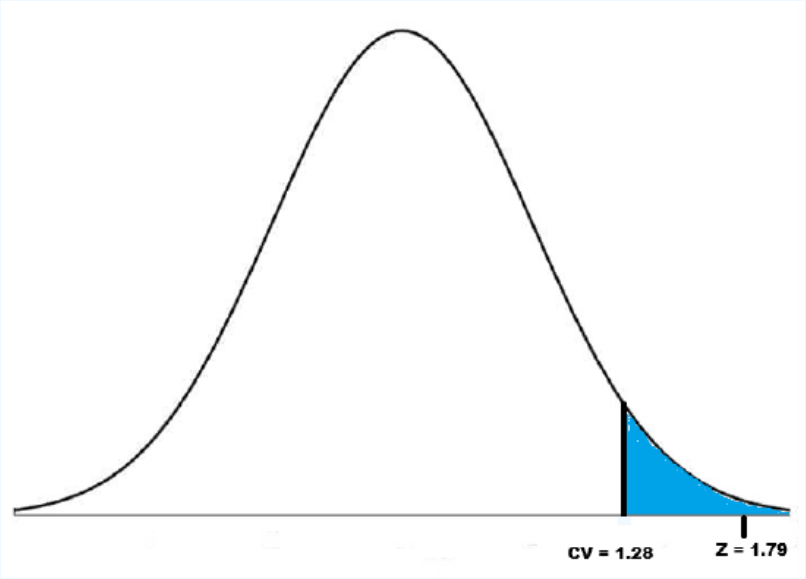


Figure 8.5.3: Standard Normal Distribution with Critical Value 1.28 and Test Point 1.79.

Since the test statistic falls in the critical region, reject H_0.

5. **Summarize the Results:** There is enough evidence to support the researcher's claim that the proportion is greater than 30%.

Example 8.5.3: P-Value Method

It has been found that 85.6% of all enrolled college and university students in the United States are undergraduates. A random sample of 500 enrolled college students in a particular state revealed that 420 of them were undergraduates. Is there sufficient evidence to conclude that the proportion differs from the national percentage? Use α = 0.05. Show that all three methods of hypothesis testing yield the same results.

Solution

1. **State the Hypothesis and Identify the Claim:** The keywords in this example, “proportion” and “differs,” lead to the following hypotheses and claim.

   H_0: p = 0.856

   H_1: p ≠ 0.856 (claim)

2. **Compute the Test Statistic:**

$$Z = \frac{\hat{p} - p}{\sqrt{\left(\frac{p \cdot q}{n}\right)}} = \frac{0.84 - 0.856}{\sqrt{\left(\frac{0.856 \cdot 0.144}{500}\right)}} = -1.019$$

3. **Compute the P-Value:**

To find the p-value, we need to find the area to the left of z = –1.02 and the right of z = 1.02, and then add the areas together. First, find the area to the left of z = –1.02 using the Standard Normal Distribution Table is 0.1539. Then, double this area to get the p-value = 0.3078.

We can also find the p-value using the TI-84+ Graphing Calculator, although the final answer may be slightly different due to the rounding of the z-value. Use the steps below to compute the p-value. In this case, the p-value = 0.3082, which differs slightly from the value found in the table.

TI-84: Press the **[STAT]** key, arrow over to the **[TESTS]** menu, arrow down to the option **[5:1-PropZTest]** and press the **[ENTER]** key. Type in the hypothesized proportion (p), x, sample size, arrow over to the $\neq$, <, > sign that is the same in the problem's alternative hypothesis statement then press the **[ENTER]** key, arrow down to **[Calculate]** and press the **[ENTER]** key.

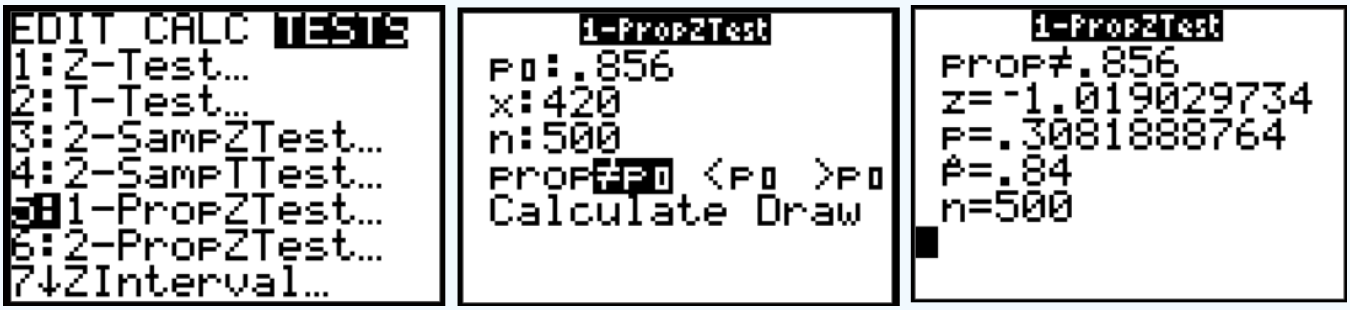


Figure 8.5.4: Output of Normal Area Function using TI-84+ Calculator.

4. **Make the Decision:** Since the p-value > α, the decision is not to reject H_0.

In this case, p-value = 0.3078 (table) or 0.3082 (calculator) > $\alpha = 0.05$ - reject H_0.

4. **Summarize the Results:** There is not enough evidence to conclude that the proportion of undergraduates in college for this state differs from the national average of 85.6%.

General Decision Rule for P-Value Method

- Reject H_0 when the p-value ≤ α.
- Do not reject H_0 when the p-value > α.

Authors

"8.5: z-Test for a Proportion" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"8.4: Hypothesis Test for One Proportion" by Rachel Webb is licensed under CC BY-SA 4.0

This page titled 8.5: z-Test for a Proportion is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 8.4: Hypothesis Test for One Proportion by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

8.6: Formulas for Chapter 8

Chapter 8 Formulas

Test Statistic for t-Test for the Mean

The t-test test statistic for the mean is used when testing a claim about a population mean, and the population standard deviation is unknown. It is typically applied when the sample size is small (less than 30), although it can be used for larger samples as well. The test is based on the t-distribution and uses the sample standard deviation in place of the population standard deviation. The test statistic measures how many estimated standard errors the sample mean is from the hypothesized population mean.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean
- μ is the population mean stated in the null hypothesis
- s is the sample standard deviation (known)
- n is the sample size
- $\frac{s}{\sqrt{n}}$ is the estimated standard error of the mean (based on the sample)

Test Statistic for z-Test for the Mean

The test point for the z-test for the mean is used when testing a claim about a population mean, and the population standard deviation is known. It is appropriate when the sample size is large (typically n ≥ 30) or the population is normally distributed. The test statistic measures how many standard errors the sample mean is from the hypothesized population mean.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{x}$ is the sample mean
- μ is the population mean stated in the null hypothesis
- σ is the population standard deviation (known)
- n is the sample size
- $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean (measures variability in sample means)

Test Statistic for z-Test for the Proportion

The test point for the z-test for a proportion is used to test a claim about a population proportion, typically when the sample size is large enough for the normal approximation to apply (usually when both np and nq are at least 5). This test determines whether the observed sample proportion differs significantly from a hypothesized population proportion. The test statistic measures how many standard errors the sample proportion is from the claimed population proportion.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Where:

- $\hat{p}$ is the sample proportion
- p is the population proportion stated in the null hypothesis
- q is the complement of the population proportion, calculated as $q = 1 - p$
- n is the sample size
- $\sqrt{\frac{pq}{n}}$ is the standard error of the sample proportion

Authors

"8.6: Formulas for Chapter 8" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 8.6: Formulas for Chapter 8 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

8.7: Chapter 8 Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Alternative Hypothesis (H_1): The statement that there is an effect or a difference; it is the logical complement of H_0.

Confidence Level ($1 - \alpha$): The probability that the confidence interval contains the true population parameter.

Critical Value Method: A traditional approach where the test statistic is compared to a cutoff value from the z or t distribution.

Hypothesis Testing: A method for making decisions about population parameters using sample data.

Null Hypothesis (H_0): The assumption that there is no effect or no difference. It is assumed true until evidence suggests otherwise.

One-Sample t-Test for the Mean: Used when the population standard deviation is unknown and the sample is drawn from a normally distributed population.

One-Sample z-Test for a Proportion: Used to test a claim about a population proportion when the sample size is large enough to satisfy normal approximation conditions.

One-Sample z-Test for the Mean: Used when the population standard deviation is known and the sample size is large or the population is normally distributed.

Parameter: A numerical measure that describes a characteristic of a population (e.g., population mean or proportion).

p-Value Method: A modern approach that calculates the probability of the observed sample result under H_0. If the p-value is less than α, H_0 is rejected.

Significance Level (α): The probability of making a Type I Error, or rejecting H_0 when it is true. Common values are 0.05 or 0.01.

Statistic: A numerical measure that describes a characteristic of a sample, used to estimate a population parameter.

Statistical Hypothesis: A claim or conjecture about a population parameter that can be tested with sample data.

Test Statistic: A standardized value (z or t) used to decide whether to reject H_0. It compares the sample result to the expected result under H_0.

Type I Error (α): Rejecting H_0 when H_0 is true.

Type II Error (β): Failing to reject H_0 when H_0 is false.

Key Symbols

H_0 - Null hypothesis.

H_1 - Alternative hypothesis.

α - Significance level (probability of a Type I error).

β - Probability of a Type II error.

$\bar{x}$ - Sample mean.

μ - Population mean.

p - Population proportion.

$\hat{p}$ - Sample proportion.

n - Sample size.

σ - Population standard deviation.

s - Sample standard deviation.

z - Test statistic for z-tests (when population standard deviation is known).

t - Test statistic for t-tests (when population standard deviation is unknown).

$z_{\alpha/2}$ - Critical z-value for two-tailed tests.

$t_{\alpha/2}$ - Critical t-value for two-tailed tests.
p-value - Probability of observing the test result under the assumption that H_0 is true.

SE - Standard error (standard deviation of the sampling distribution).

Authors

"8.7: Chapter 8 Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 8.7: Chapter 8 Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

9: HYPOTHESIS TESTING FOR TWO SAMPLES

9.1: z-Test for the Difference Between Two Means

9.2: t-Test for the Difference Between Two Means

9.3: z-Test for the Difference Between Two Proportions

9.4: Confidence Intervals for the Difference Between Two Means and Proportions

9.5: Formulas for Chapter 9

9.6: Chapter 9 - Key Terms and Symbols

This page titled 9: Hypothesis Testing for Two Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

9.1: z-Test for the Difference Between Two Means

Learning Objectives

- Understand how to perform hypothesis testing for the difference between two means using z-values.
- Recognize that this method is appropriate when both population standard deviations are known and samples are large, or populations are normally distributed.
- Determine whether there is a significant difference between the two population means.
- Understand that the method assumes independent samples.
- Use the standard normal distribution (z-distribution) to compare the test statistics and make decisions.

There are many situations where two means will be compared. For example, assume the average age of students who attend community colleges in California is to be compared to the average age of students who attend community colleges in Florida. The two populations must be independent of each other. This means that there is no data value common to both populations. Random samples of each population can be collected.

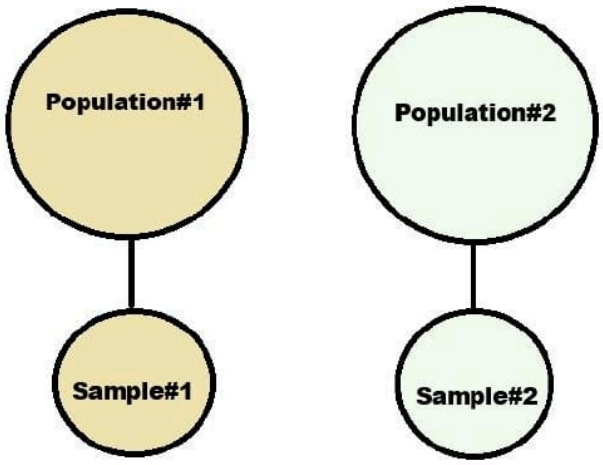


Figure 9.1.1: Visualization of the Difference Between Two Independent Populations and Samples

Moreover, μ_1 represents the mean of population#1 and μ_2 represents the mean of population#2. Moreover, $\bar{X}_1$ is the mean of sample#1 and $\bar{X}_2$ is the mean of sample#2. Using the central limit theorem, multiple samples involving the difference of the two sample means ($\bar{X}_1$-$\bar{X}_2$) can be determined. This will result in a normally shaped distribution in which the mean of the distribution of the differences of the sample means equals zero. To apply the z-test, the following conditions must be true.

Conditions for z-test

1. Both samples must be random.
2. The population standard deviations must be known.
3. Both samples must have sample sizes greater than or equal to 30, or both populations must be normal.

z-test Process for Two Means

1. State the hypothesis, identify the claim, and determine the null and alternative hypotheses. The three possibilities are presented below. One of these statements will be the claim. Draw the standard normal distribution and shade in the proper

tail area. The total tail area equals the given α.

Table 9.1.1: Generalized Hypothesis Statements

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$

2. Look up the critical values in the table. These values separate the critical regions from the non-critical regions.
3. Compute the test point using the formula below.

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Note: Since the test is for H_0, the difference $\mu_1 - \mu_2 = 0$.

4. Make the decision to reject or not reject the null hypothesis.
5. Summarize the results.

Examples

Example 9.1.1 Testing the Differences in Ages of Two Population Means

An educational researcher claims that there is a difference in the average ages of community college students who attend California and Arizona. She collects random samples from both populations and lists the results in the table below. Test the researcher's claim using α = 0.05. The data is measured in years.

Table 9.1.2: Difference in Ages at Two Community Colleges

California Community Colleges	Arizona Community Colleges
$\bar{X}_1 = 28$	$\bar{X}_2 = 25$
$\sigma_1 = 4.5$	$\sigma_2 = 3.2$
$n_1 = 44$	$n_2 = 40$

Solution

1. **State the Hypothesis and Identify the Claim:** The word "difference" translates to "$\neq$," which leads to the following hypotheses and claim.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ Claim

2. **Find the Critical Value:** Draw a standard normal distribution and look up the critical values in the table. They are ± 1.96.
3. **Compute the Test Statistic:** Use the given sample data and the test point formula to compute the value. Recall that $\mu_1 - \mu_2 = 0$.

$$z = \frac{(28 - 25) - 0}{\sqrt{\frac{4.5^2}{45} + \frac{3.2^2}{40}}} = 3.57$$

4. **Make the Decision:** The result of the z-test is to "reject H_0" as the test point falls in the critical region.

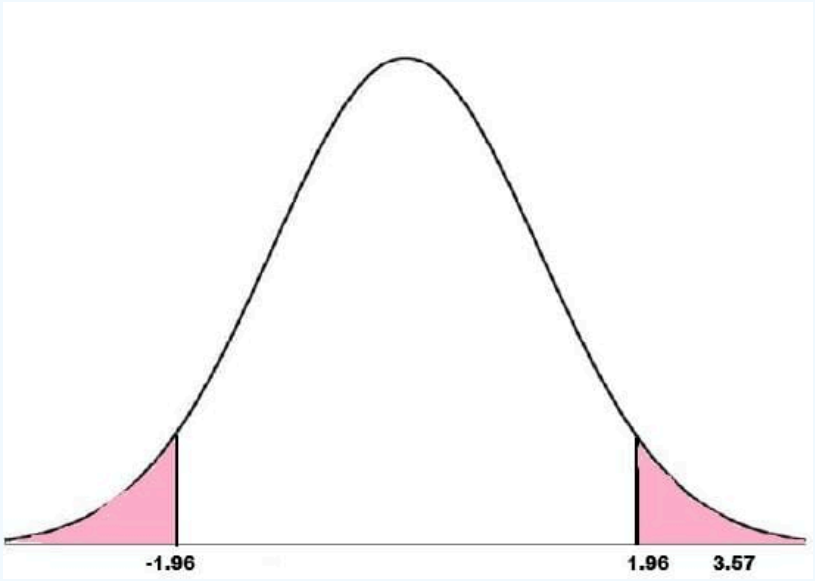


Figure 9.1.2: Standard Normal Distribution with Critical Values Plus and Minus 1.96 and Test Value 3.57.

5. **Summarize the Results:** Write out the summary statement. This summary statement is to "support the claim that there is a difference in ages."

Example 9.1.2

A dietitian compares two types of diets (diet A and diet B) to determine whether one leads to more weight loss on average. He claims that diet A's average weight loss is greater than diet B's average weight loss. The data (collected in pounds) is summarized in the table below. Test the claim using α = 0.05.

Table 9.1.3: Difference in Diets

Diet A	Diet B
$\bar{X}_1 = 15$	$\bar{X}_2 = 13.3$
$\sigma_1 = 2.3$	$\sigma_2 = 2.7$
$n_1 = 35$	$n_2 = 48$

Solution

1. **State the Hypothesis and Identify the Claim:** The word "greater than" translates to ">," which lead to the following hypotheses and claim.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$ Claim

2. **Find the Critical Value:** Draw a standard normal distribution, shade in the right tail, and look up the critical values in the table. It is 1.65.
3. **Compute the Test Statistic:** Use the given sample data and the test point formula, compute the value. Recall that $\mu_1 - \mu_2 = 0$.

$$z = \frac{(15 - 13.3) - 0}{\sqrt{\frac{2.3^2}{35} + \frac{2.7^2}{48}}} = 3.09$$

4. **Make the Decision:** The result of the z-test is to "reject H_0" as the test point falls in the critical region.

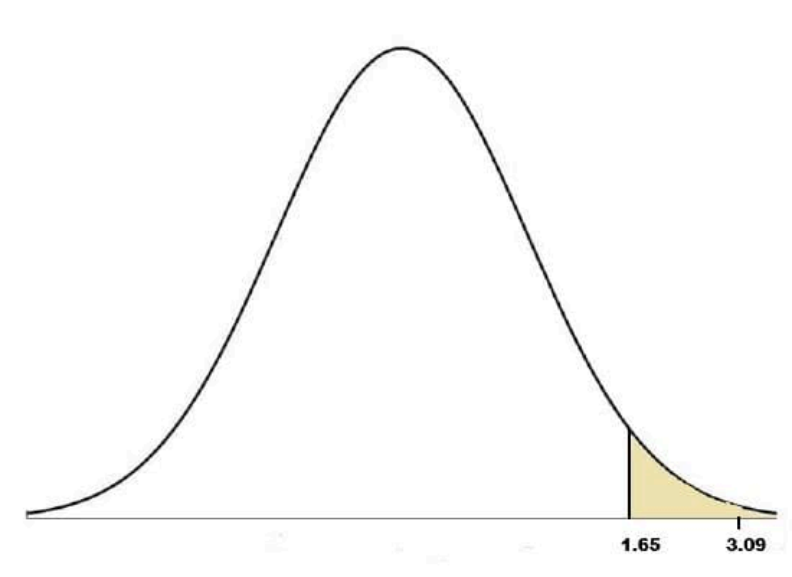


Figure 9.1.3: Standard Normal Distribution with Critical Value 1.65 and Test Value 3.09.

5. **Summarize the Results:** Write out the summary statement. This summary statement is to "support the claim that the average weight loss for diet A is greater than the average weight loss for diet B."

Technology such as the TI-84+ calculator can be used to compute the test point for a z-test for two means. This process will be used in the next example.

Example 9.1.3

A researcher is testing the claim that there is a difference in the average cost spent by a family of four who attend the National Basketball Association (NBA) and National Hockey League (NHL) games. Data is collected and summarized in the table below. The data is units of U.S. dollars. Test the claim using α = 0.01.

Difference in Costs for Sporting Events

NBA Game Costs for Family of Four	NHL Game Costs for Family of Four
$\bar{X}_1 = 468$	$\bar{X}_2 = 448$
$\sigma_1 = 45.69$	$\sigma_2 = 33.98$
$n_1 = 30$	$n_2 = 40$

Table 9.1.4: Difference in Costs at Sporting Events

Solution

1. **State the Hypothesis and Identify the Claim:** The word "difference" translates to "$\neq$," which leads to the following hypotheses and claim.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ Claim

2. **Find the Critical Value:** Draw a standard normal distribution, shade in two tails, and look up the critical values in the table. They are ± 2.58.

3. **Compute the Test Statistic:** Use a TI-84+ calculator to compute the test statistic value. Follow the steps below.

a) Press the **[STAT]** button, use the right arrow to select **[TESTS],** and then select **[3: 2-SampZTest]**.

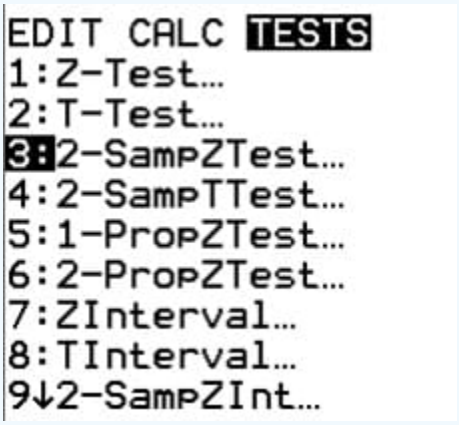


Figure 9.1.4: TI-84+ the Selection of a Two-Sample Z-test Function

b) Select **[STATS]**, and then enter the data from the table into the calculator. The input should look as it does in the image below. Finally, select **[Calculate]** and press **[Enter]**.

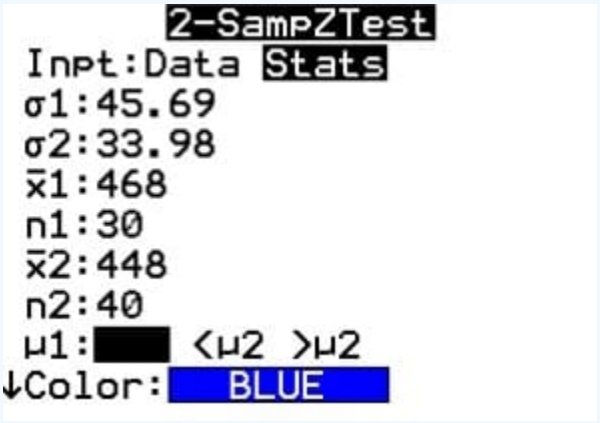


Figure 9.1.5: TI-84+ Input of Data for a Two Sample Test Function

c) The output on the calculator screen is presented in the image below. Note the test point is 2.02 (rounded to two place values) and the p-value is 0.0438 (rounded to four decimal place values).

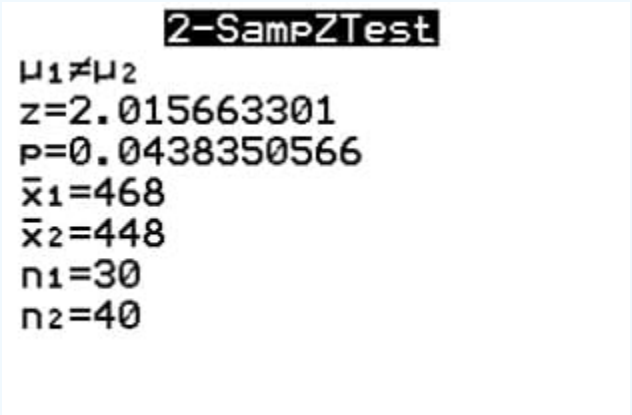


Figure 9.1.6: TI-84+ Output of Data for a Two Sample Test Function

4. **Make the Decision:** The result of the z-test is "do not reject H_0" as the test point does not fall in the critical region. Also, the p-value = 0.0438 > $\alpha = 0.01$

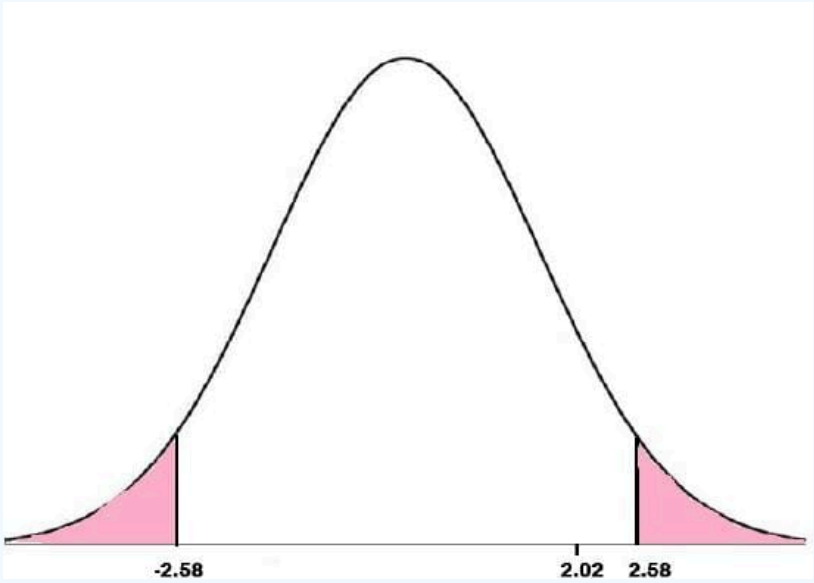


Figure 9.1.7: Standard Normal Distribution with Critical Values Plus and Minus 2.58 and Test Value 2.02

5. **Summarize the Results:** Write out the summary statement. This summary statement is written as "do not support the claim that there is an average difference in costs at NBA and NHL games for a family of four."

Author

"9.1: z-Test for the Difference Between Two Means" by Alfie Swan is licensed under CC BY 4.0

This page titled 9.1: z-Test for the Difference Between Two Means is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

9.2: t-Test for the Difference Between Two Means

Learning Objectives

- Understand how to perform hypothesis testing for the difference between two means using t-values.
- Recognize that this method is appropriate when population standard deviations are unknown and sample sizes are small.
- Apply the test to compare two independent groups.
- Use the t-distribution to account for variability in sample standard deviations.
- Make decisions based on the comparison of the test statistic to critical t-values.

A t-test for the difference between two means is used to determine whether there is a statistically significant difference between the means of two independent groups. The test compares the sample means while accounting for variability within each group. The null hypothesis assumes no difference between the means, while the alternative hypothesis suggests a difference exists. The test statistic is calculated using the sample means, standard deviations, and sample sizes and is then compared to a critical value from the t-distribution.

Hypothesis Test for Independent t-Test (2-Sample t-Test)

1. State the claim and determine the null and alternative hypotheses. Also, the number of tails for the test will be determined, and the tails will be shaded in the t-distribution.

Types of Tests

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
$H_0 : \mu_1 = \mu_2$ $H_0 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_0 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_0 : \mu_1 < \mu_2$

Table 9.2.1: Types of Test for Hypotheses Testing

2. Look up the critical value in the table using degrees of freedom. The degrees of freedom is the minimum value between $n_1 - 1$ and $n_2 - 1$
3. If needed, compute the sample statistics - sample means and standard deviations.

Sample Statistic:
Calculate $\overline{x}_1, \overline{x}_2, s_1, s_2, n_1, n_2$

4. Find the sample statistic, test statistic, and if needed the p-value

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Note: Since the test is for H_0, the difference $\mu_1 - \mu_2 = 0$.

5. Make a decision and summarize the results.

Example 9.2.1 t-Test Using Traditional (Critical Value) Method

A researcher claims that there is a difference in the average weight of an adult rhinoceros and an adult hippopotamus. The researcher collected data to test the claim and summarized the results in the table below. The data is measured in kilograms.Test the claim using using $\alpha = 0.01$

Data Set

Rhinoceros	Hippopotamus
$\bar{X}_1 = 2300$	$\bar{X}_2 = 1820$
$s_1 = 225$	$s_2 = 180.5$
$n_1 = 10$	$n_2 = 11$

Table 9.2.2: Summarized Data for the Weights of Adult Rhinoceroses and Hippopotamuses.

Solution

1. State the claim and determine the null and alternative hypotheses. The word "difference" translates to $\neq$.

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 \neq \mu_2$ Claim

The test has two tails.

2. Determine the critical values.

Calculate $n_1 - 1 = 10 - 1$ = 9 and $n_2 - 1 = 11 - 1$ = 10. The degrees of freedom is 9 as it is the minimum of the two values. Looking up the values in the table using d.f. = 9, two tails, and $\alpha = 0.01$, the critical values are $\pm$3.250.

3. Compute the test point.

$$t = \frac{(2300 - 1820) - (0)}{\sqrt{\frac{225^2}{10} + \frac{180.5^2}{11}}} = 5.36$$

4. Make a decision. For this problem, the decision is to "reject H_0" because the t-test point falls in the critical region

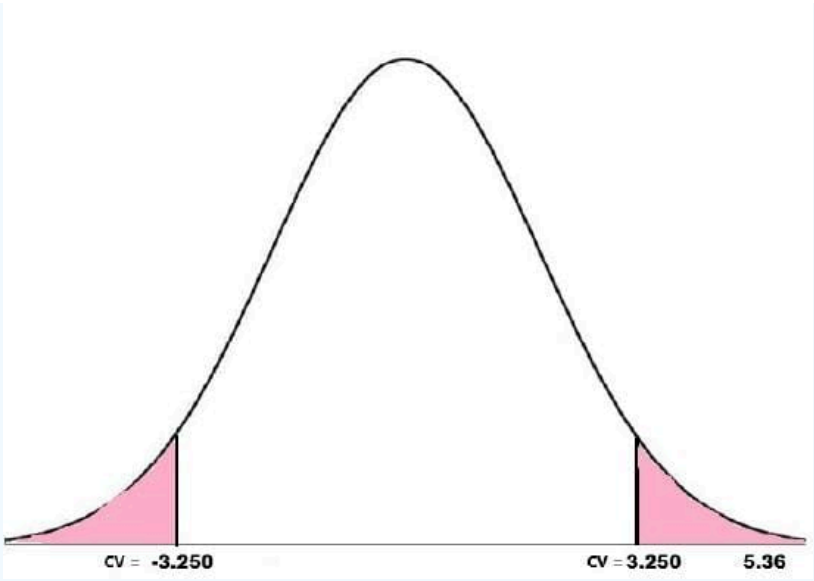


Figure 9.2.1: The Test Point (5.36) Falls into the Critical Region - Reject H_0.

5. The summary statement is that "there is enough evidence to support that there is a difference in the weights of the two animals."

Example 9.2.2 t-Test Using Traditional (Critical Value) Method

A math instructor wishes to test a claim that the average score on exams taken by night students is greater than the average score on exams taken in the morning at the school where the instructor works. The instructor collects two random samples of exam scores from both populations. The data is listed in the table below Test the claim using $\alpha = 0.05$.

Data Set

Evening Student Scores	Morning Student Scores
98	49
80	78
88	92
85	60
86	87
88	61
95	70
86	71
96	95
75	85

Table 9.2.3: Data for the Evening and Morning Student Scores.

Solution

1. State the claim and determine the null and alternative hypotheses. The word "greater than" translates to ">."

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 > \mu_2$ Claim

The test has one tail.

2. Determine the critical value.

Calculate $n_1 - 1 = 10 - 1$ = 9 and $n_2 - 1 = 11 - 1$ = 10. The degrees of freedom are 9 as 9 is the minimum of the two values. Looking up the values in the table using d.f. = 9, one tail, and $\alpha = 0.05$, the critical value is 1.833.

3. Compute the test point. First, use the calculator to compute the sample means and sample standard deviations. Round to two place values. The results are listed below.

Data Set

Evening Student Scores	Morning Student Scores
$\bar{X}_1 = 87.7$	$\bar{X}_2 = 74.8$
$s_1 = 7.17$	$s_2 = 15.2$
$n_1 = 10$	$n_2 = 10$

Table 9.2.4: Summarized Data for the Evening and Morning Student Scores.

$$t = \frac{(87.7 - 74.8) - (0)}{\sqrt{\frac{7.17^2}{10} + \frac{15.2^2}{10}}} = 2.43$$

4. Make a decision. For this problem, the decision is to "reject H_0" because the t-test point falls in the critical region.

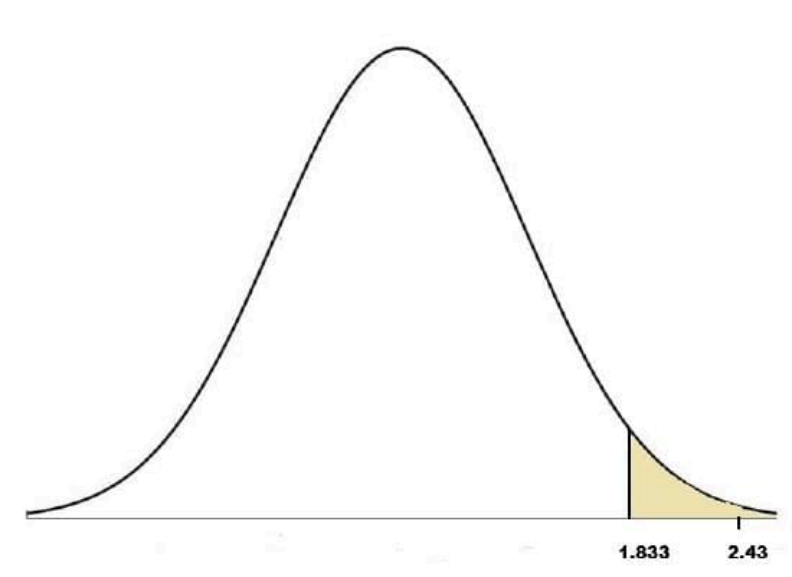


Figure 9.2.2: The Test Point (2. 43) Falls into the Critical Region - Reject H_0.

5. The summary statement is that "there is enough evidence to support the claim that the average of the exam scores of the evening students is greater than the average of the exam scores of the morning students."

Example 9.2.3 t-Test Using Traditional (Critical Value) Method

At Willow Creek High School, students are gearing up for the annual “Battle of the Genders” sports competition. The coach believes that female athletes are putting in fewer hours of practice than the male athletes this season. To test this claim, the coach collects a random sample of practice times (in hours per week) from each group.

Female and Male Athletes' Summarized Data

Female Athletes	Male Athletes
$\bar{X}_1 = 6.2$	$\bar{X}_2 = 7.4$
$s_1 = 1.1$	$s_2 = 1.3$
$n_1 = 12$	$n_2 = 13$

Table 9.2.5: Summarized Data for the Female and Male Athletes

Using a significance level of $\alpha = 0.005$, test the claim that female athletes practice fewer hours than male athletes.

Solution

1. State the claim and determine the null and alternative hypotheses. The word "greater than" translates to ">."

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 < \mu_2$ Claim

2. Determine the critical value.

Calculate $n_1 - 1 = 12 - 1$ = 11 and $n_2 - 1 = 13 - 1$ = 12. The degrees of freedom are 11, as 11 is the minimum of the two values. Looking up the values in the table using d.f. = 11, one tail, and $\alpha = 0.005$, the critical value is –3.106.

3. Compute the test point.

$$t = \frac{(6.2 - 7.4) - (0)}{\sqrt{\frac{1.1^2}{12} + \frac{1.3^2}{13}}} = -2.50$$

4. Make a decision. For this problem, the decision is to "not reject H_0" because the t-test point does not fall into the critical region.

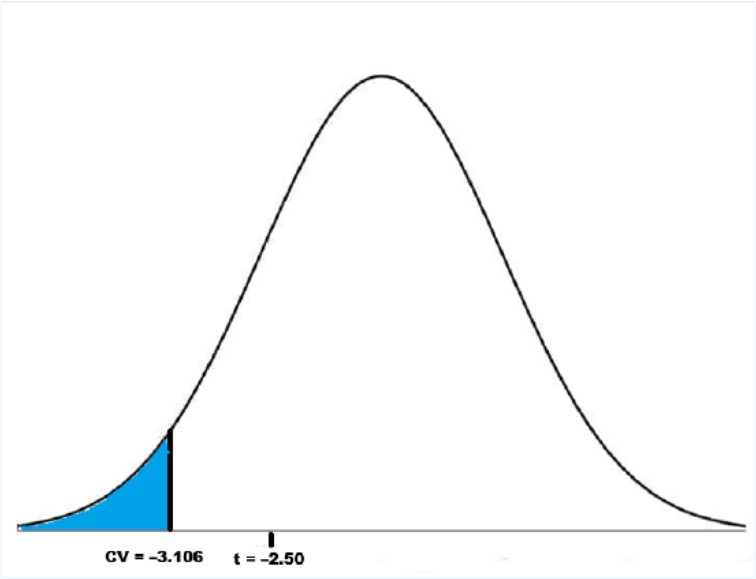


Figure 9.2.3: The Test Point (t = –2.50) Falls into the Non - Critical Region - Do Not Reject H_0.

The TI-84+ calculator can also be used to calculate the test point for a t-test with two independent samples.

Example 9.2.4 t-Test Using Traditional (Critical Value) Method Using a Graphing Calculator

The average prices for new 15-lb turkeys sold at commercial markets in California and Texas will be compared. The researcher claims that the average price of a 15-lb Thanksgiving turkey in California is greater than the average price of a 15-lb Thanksgiving turkey in Texas. The data is measured in dollars and presented in the table below. Test the researcher's claim using $\alpha = 0.01$.

Data Set

California	Texas
54	34
44	30
26	18
43	18
22	29
47	41
44	35
43	20
41	39
47	21
25	28

Table 9.2.6: Data for the Prices of 15-lb Turkeys in Texas and California.

Solution

1. State the claim and determine the null and alternative hypotheses. The word "greater than" translates to ">."

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 > \mu_2$ Claim

The test has one tail.

2. Determine the critical value.

Calculate $n_1 - 1 = 11 - 1$ = 10 and $n_2 - 1 = 11 - 1$ = 10. The degrees of freedom are 10 as 10 is the minimum of the two values. Looking up the values in the table using d.f. = 10, one tail, and $\alpha = 0.01$, the critical value is 2.764.

3. Using the calculator to compute the test point. Enter the data into the TI-84+ calculator by pressing **[STAT]** and then select **[1:EDIT].** Enter the data in column 1 into List 1 (L1) and the data in column 2 into List 2 (L2). The results should look like the image below.

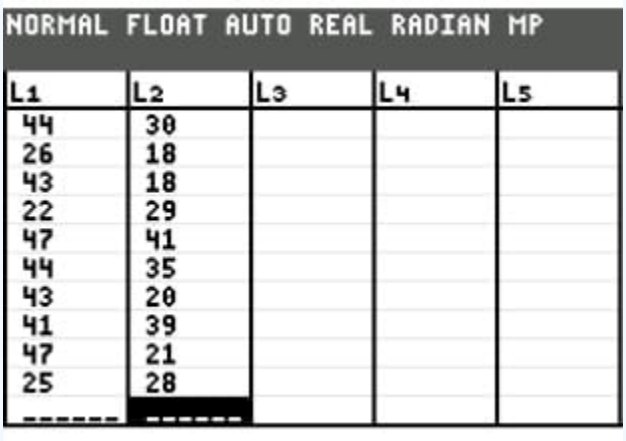


Figure 9.2.4: Enter the Data into the TI-84+ Calculator

Press **[STAT]** and use the right arrow to select **[TESTS]** and then the down arrow to select **[4:2-SampTTest]**, and then hit the [ENTER] button. Use the left arrow to select "DATA" and then fill in the information provided in the image below.

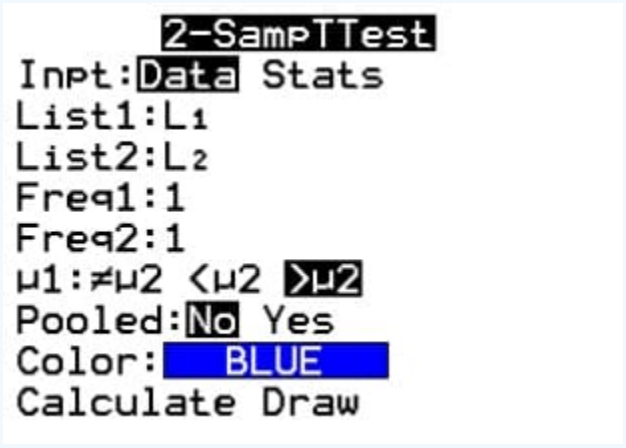


Figure 9.2.5: Check to Make Sure the Correct Lists and Other Key Items are Selected.

After selecting **[Calculate]** and hitting **[ENTER]**, the output in the image below should appear on the calculator. The test point is t = 2.78 (rounded to two place values) and the p-value is 0.0060 (Rounded to four decimal places).

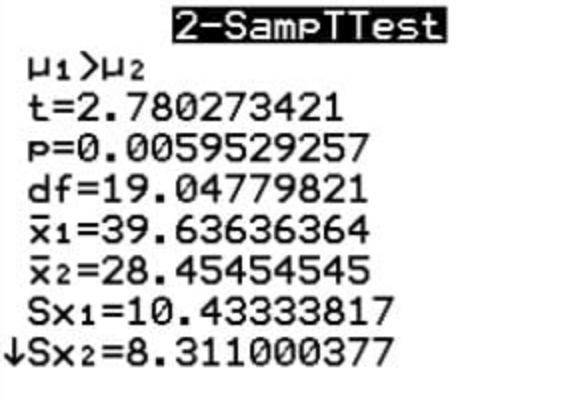


Figure 9.2.6: The Output of the Test Point and Other Data

4. Make a decision. For this problem, the decision is to "reject H_0" because the t-test point falls in the critical region.

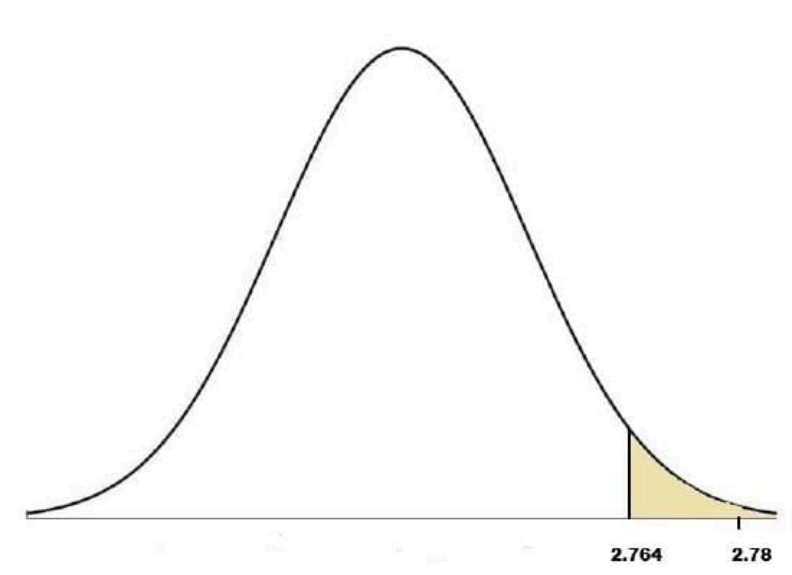


Figure 9.2.7: The Test Point (2.78) Falls into the Critical Region - Reject H_0

5. The summary statement is that "there is enough evidence to support the claim that the average price of 15-lb turkeys in California is greater than the average price of 15-lb turkeys in Texas.

Author

"9.2: t-Test for the Difference Between Two Means" by Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"9.3: Independent Samples for Two Means" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 9.2: t-Test for the Difference Between Two Means is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 9.3: Independent Samples for Two Means by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

9.3: z-Test for the Difference Between Two Proportions

Learning Objectives

- Understand how to perform hypothesis testing for the difference between two proportions using z-values.
- Recognize that this method is appropriate for comparing proportions from two independent groups with large sample sizes.
- Determine whether the difference between the two population proportions is statistically significant.
- Apply the normal approximation for large samples.
- Use the standard normal distribution (z-distribution) to make decisions based on the test statistic.

This section will examine how to analyze a difference in the proportions for two independent samples. As with all other hypothesis tests and confidence intervals, the testing process is the same, although the formulas and assumptions differ. In this section, we compare the differences between 2 population proportions labeled p_1 and p_2. The two samples used must be random and independent. Also, to ensure normality, the following inequalities must be valid.

- $n_1 \cdot p_1 \geq 5$
- $n_1 \cdot (1 - p_1) \geq 5$
- $n_2 \cdot p_2 \geq 5$
- $n_2 \cdot (1 - p_2) \geq 5$

There are three types of hypothesis tests for comparing the difference in 2 population proportions $p_1 - p_2$, see Figure 9-7.

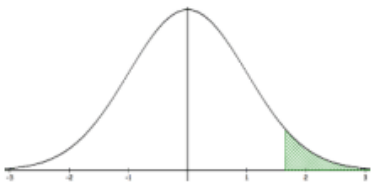
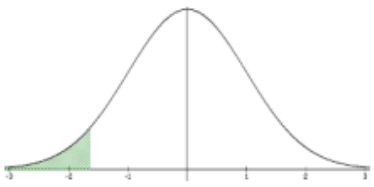
Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$	$H_0: p_1 = p_2$ $H_1: p_1 > p_2$	$H_0: p_1 = p_2$ $H_1: p_1 < p_2$
		
$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 \neq 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 > 0$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 < 0$

Figure 9.3.1: Types of Tests Used in Hypothesis Testing for the Difference Between Two Proportions

Hypothesis Test for the Difference of Two Means

1. State the claim using the hypotheses presented in Figure 9-7.
2. Draw the probability distribution and shade in the tail area using the proper number of tails. Use Figure 9-1 as a reference.
3. Look up the critical values using the table or other technology.
4. To compute the **test statistic** (or test point) for a hypothesis test involving two proportions, you can use the formula listed in the formula box below or technology.
5. Decide on the null hypothesis (H_0) and summarize the results.

Definition: Test Statistic Formula for Difference of 2 Proportions

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\bar{p} \cdot \bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}}$$

Where,

- $\hat{p}_1 =$ sample proportion for sample 1, calculated as $\hat{p}_1 = \frac{X_1}{n_1}$
- $\hat{p}_2 =$ sample proportion for sample 2, calculated as $\hat{p}_2 = \frac{X_2}{n_2}$
- p_1 = the proportion of population 1.
- p_2 = the proportion of population 2.
- Note that for our purposes, $p_1 - p_2 = 0$.
- $\bar{p}$ = pooled proportion and is computed as $\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$
- X_1 = number of successes in sample 1.
- X_2 = number of successes in sample 2.
- n_1 = sample size of population 1.
- n_2 = sample size of population 2.
- $\bar{q} = 1 - \bar{p}$

Examples

Example 9.3.1 Traditional (Critical Value) Method

At a university study lounge, a rumor spreads that social science majors spend less time in group study sessions than STEM majors. To investigate, a curious researcher surveys students across campus.

Out of 120 surveyed social science majors, 54 report attending a weekly study group.
Out of 100 surveyed STEM majors, 55 report the same.

Using $\alpha = 0.05$, test the claim that the proportion of social science majors attending weekly study groups is less than STEM majors.

Solution

1. The claim and hypotheses are written as follows.

$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$ Claim

This is a one-tailed test.

2. The critical value is found using the table and $\alpha = 0.05$. It is -1.65.
3. Compute the test point.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{54}{120} = 0.45$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{55}{100} = 0.55$$

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(54 + 55)}{(120 + 100)} = 0.4955$$

$\hat{q} = 1 - \hat{p} = 1 - 0.4955 = 0.5045$

The test statistic is, $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\bar{p} \cdot \bar{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\right)}} = \dfrac{(0.45 - 0.55)}{\sqrt{\left(0.4955 \cdot 0.5045\left(\dfrac{1}{120} + \dfrac{1}{100}\right)\right)}} = -1.48.$

4. The result is to " not reject H_0" as the test point does not fall in the critical region.

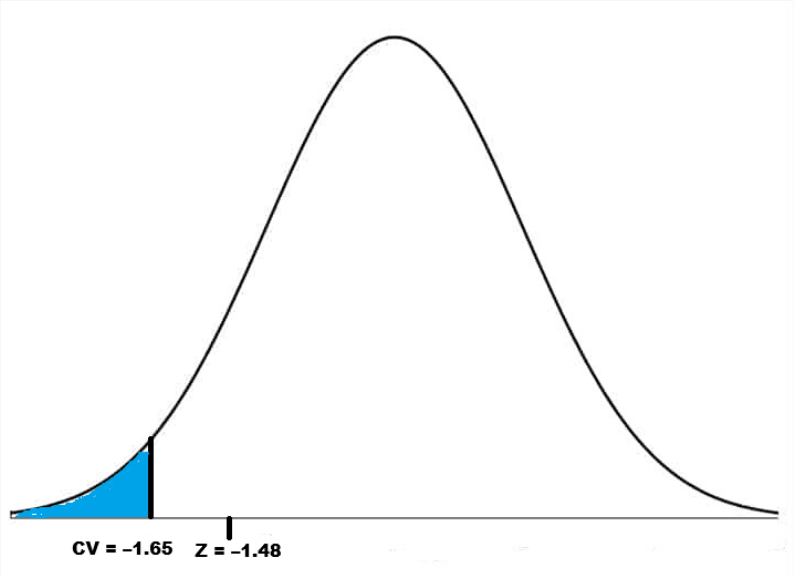


Figure 9.3.2: The Test Point (-1.48) Falls into the Non-Critical Region - Do Not Reject H_0

5. The summary "there is not enough evidence to support the claim that social science majors spend less time in study groups than STEM majors."

Example 9.3.2 Traditional (Critical Value) Method

A marketing researcher is testing the claim of whether there is a significant difference between the proportion of burger lovers and pizza lovers who attend a local community college. To test the claim, the researcher uses $\alpha = 0.10$. Two random and independent samples were collected, and the results are presented in the table below.

Data For Burger Lovers and Pizza Lovers

Burger Lovers	Pizza Lovers
$X_1 = 90$	$X_2 = 100$
$n_1 = 150$	$n_2 = 200$

Table 9.3.1: Data for Burger Lovers and Pizza Lovers

Solution

1. Since the word "difference" translates to "not equal", the claim and hypotheses are written as follows.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$ Claim

This is a two-tailed test.

2. The critical values are found using the table and the level of significance (0.01). They are ± 1.65.
3. Compute the test point.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{90}{150} = 0.6$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{100}{200} = 0.5$$

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(90 + 100)}{(150 + 200)} = 0.543$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.543 = 0.457$$

The test statistic is, $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\bar{p} \cdot \bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \dfrac{(0.6 - 0.5)}{\sqrt{\left(0.543 \cdot 0.457\left(\frac{1}{150} + \frac{1}{200}\right)\right)}} = 1.86.$

4. The result is to "reject H_0" as the test point falls in the critical region.

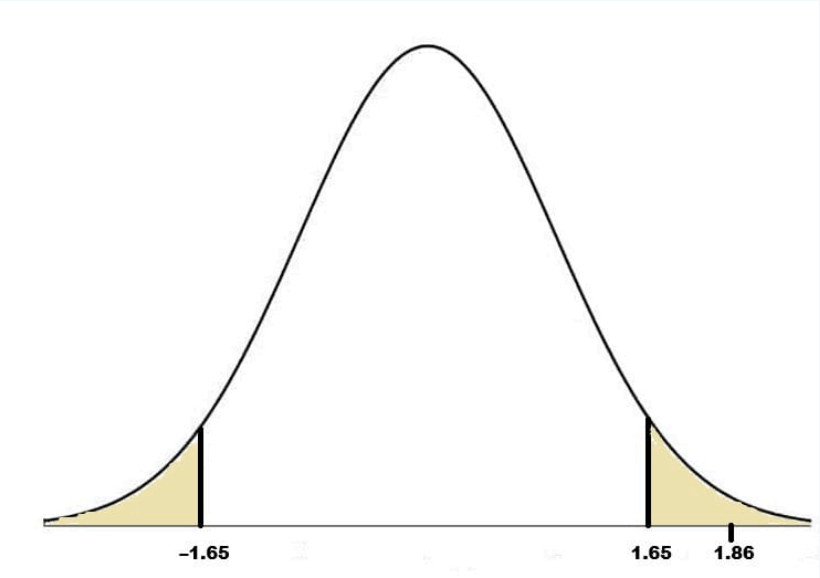


Figure 9.3.3: The Test Point (1.86) Falls into the Critical Region - Reject H_0

5. The summary is "there is enough evidence to support the claim."

Example 9.3.3 Traditional (Critical Value) Method Using Graphing Calculator

A vice principal wants to see if there is a difference between the number of students who are late to class for the first class of the day compared to the students' class right after lunch. To test their claim to see if there is a difference in the proportion of late students between first and after-lunch classes, the vice-principal randomly selects 200 students from the first class and records if they are late, then randomly selects 200 students from their class after lunch and records. The number of students who were late to the first class was 13, and the number of students who were late to class after lunch was 16. At $\alpha = 0.05$, can a difference be concluded?

Solution

1. Since the word "difference" translates to "not equal", the claim and hypotheses are written as follows.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$ Claim

This is a two-tailed test.

2. The critical values are found using the table and the level of significance (0.05). They are ± 1.96.
3. Compute the test point.

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{13}{200} = 0.065$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{16}{200} = 0.08$$

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)} = \frac{(13 + 16)}{(200 + 200)} = 0.0725$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.0725 = 0.9275$$

The test statistic is, $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\bar{p} \cdot \bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} = \dfrac{(0.065 - 0.08)}{\sqrt{\left(0.0725 \cdot 0.9275\left(\frac{1}{200} + \frac{1}{200}\right)\right)}} = -0.5784 = -0.58.$

TI-84: Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option [6:2-PropZTest] and press the [ENTER] key. Type in the x_1, n_1, x_2, and n_2 arrow over to the $\neq$, <, > sign that is the same in the problem's alternative hypothesis statement, then press the [ENTER] key, arrow down to [Calculate] and press the [ENTER] key. The calculator returns the z-test statistic and the p-value.

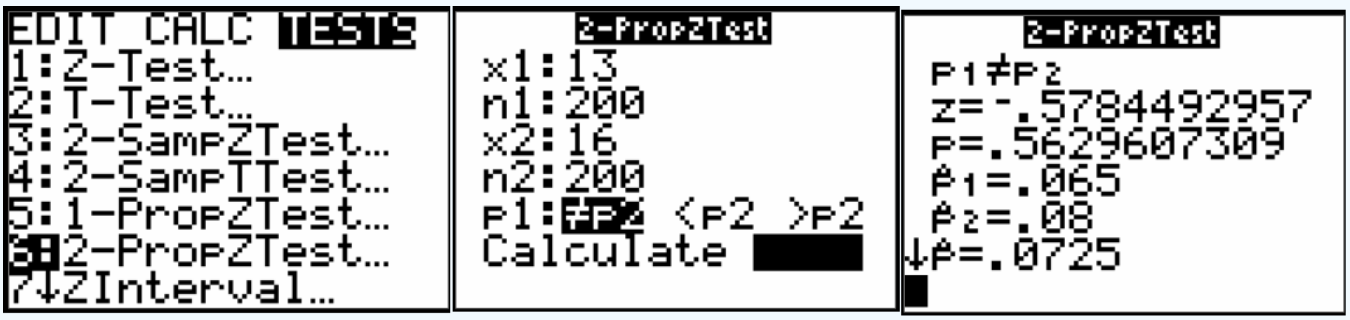


Figure 9.3.4: Test selection, Input, and Output for the TI-84+

4. The result is to "not reject H_0 as the test point falls in the non-critical region.

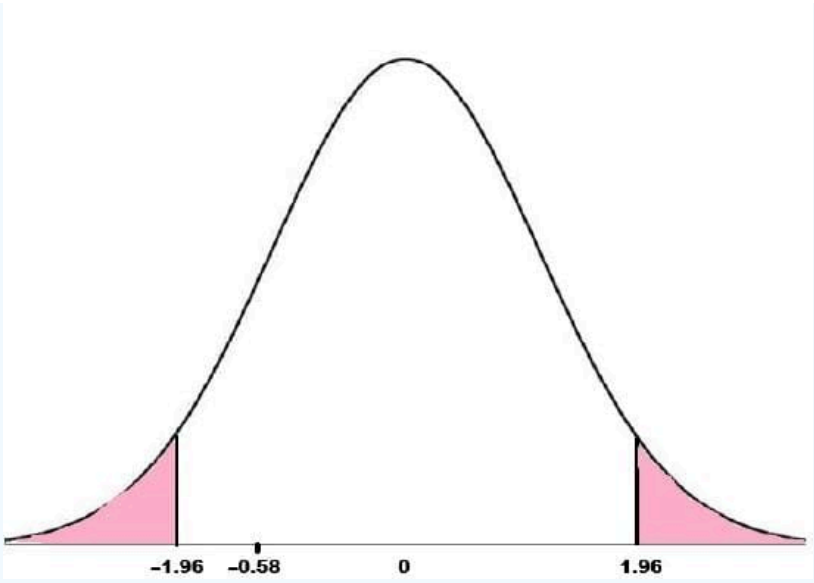


Figure 9.3.5: The Test Point (-0.58) Falls into the Non-Critical Region - Do Not Reject H_0

5. The summary is "there is not enough evidence to support the claim."

Authors

"9.3: z-Test for the Difference Between Two Proportions" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"9.1: Two Proportions" by Kathryn Kozak is licensed under CC BY-SA 4.0

"9.3: Two Proportion Z-Test and Confidence Interval" by Rachel Webb is licensed CC BY-SA 4.0

This page titled 9.3: z-Test for the Difference Between Two Proportions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 9.3: Two Proportion Z-Test and Confidence Interval by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

9.4: Confidence Intervals for the Difference Between Two Means and Proportions

Learning Objectives

- Understand how to construct confidence intervals for the difference between two means or two proportions.
- Use confidence intervals to estimate the range where the true difference between two populations likely lies.
- Apply this method when comparing two independent groups.
- Assess whether a meaningful difference exists without performing a formal hypothesis test.
- Interpret the confidence interval to evaluate the significance of the difference between groups.

In statistical analysis, **confidence intervals** are crucial in estimating unknown population parameters based on sample data. A confidence interval provides a range of plausible values within which the true parameter value is expected to lie, given a specified confidence level, such as 90%, 95%, or 99%. These intervals not only account for variability due to sampling but also reflect the uncertainty inherent in any estimation process.

This section presents two examples of confidence intervals for estimating the difference between two population parameters. The first type illustrates a confidence interval for the difference between two population **means** based on a normally distributed sample and t-distribution. The second type demonstrates a confidence interval for the difference of two **proportions**. The two types of confidence intervals highlight the methodology, assumptions, and interpretation of the resulting confidence intervals, providing insights into their application in real-world problems. Moreover, it is assumed that the two samples are collected randomly from two populations and are independent samples.

By constructing these confidence intervals, we gain a robust framework for statistical inference that allows researchers and decision-makers to make data-driven conclusions with a quantifiable level of confidence.

Confidence Interval for the Difference of Two Means using Z-values

This formula is used when both population standard deviations σ_1 and σ_2 are given or both sample sizes n_1 and n_1 are greater than or equal to 30.

Definition: Confidence Interval Formula for the the Difference of 2 means using Z-values

$$(\bar{X}_1 - \bar{X}_2) - Z_C \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) - Z_C \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where,

- $\bar{X}_1$ = sample mean of the first sample.
- $\bar{X}_2$ = sample mean of the second sample.
- σ_1 = population standard deviation of the first population.
- σ_2 = population standard deviation of the second population.
- n_1 = sample size of the first sample.
- n_2 = sample size of the second sample.
- Z_C = critical value from the standard normal distribution for a given confidence level.

Example 9.4.1

An educational researcher is investigating whether there is a significant difference in the academic performance of male and female students at a university. To measure academic performance, the researcher compares the grade point averages (GPAs) of female and male students. The data collected is provided in the table below. Also, the researcher uses a confidence level of 95%. Compute the confidence interval for the difference between the two means. Round all calculations to three decimal places.

Females	Males
$\bar{X}_1$ =3.15	$\bar{X}_2$ =2.78
σ_1 =0.52	σ_2 =0.48
n_1 =120	n_2 =100

Solution

Step 1) Look up the critical value in a table or using a calculator. The value is Z_C = 1.96.

Step 2) Plug information into the formula and calculate using the order of operations.

$$(\bar{X}_1 - \bar{X}_2) - Z_C \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) - Z_C \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(3.15 - 2.78) - 1.96 \cdot \sqrt{\frac{0.52^2}{120} + \frac{0.48^2}{100}} < \mu_1 - \mu_2 < (3.15 - 2.78) + 1.96 \cdot \sqrt{\frac{0.52^2}{120} + \frac{0.48^2}{100}}$$

$$0.37 - 0.132 < \mu_1 - \mu_2 < 0.37 + 0.132$$

$$0.238 < \mu_1 - \mu_2 < 0.502$$

Since the difference does not contain zero, the average GPA for women is significantly higher than the average GPA for men.

The TI-84+ can be used to compute the confidence interval as well.

Step 1) Click on the "stat" button and use the right arrow to select "TESTS." Select option 9 using the down arrow or typing in 9. This option has the title "2-SampZint."

Step 2) Under "2-SampZint," select "Stats." Type in the given data and select "Calculate."

Step 3) The output is at the top of the screen. Round the two answers to three decimal places. They are written in parentheses and are separated by a comma. After rounding the confidence interval is (0.237, 0.502).

The image below captures all three steps.

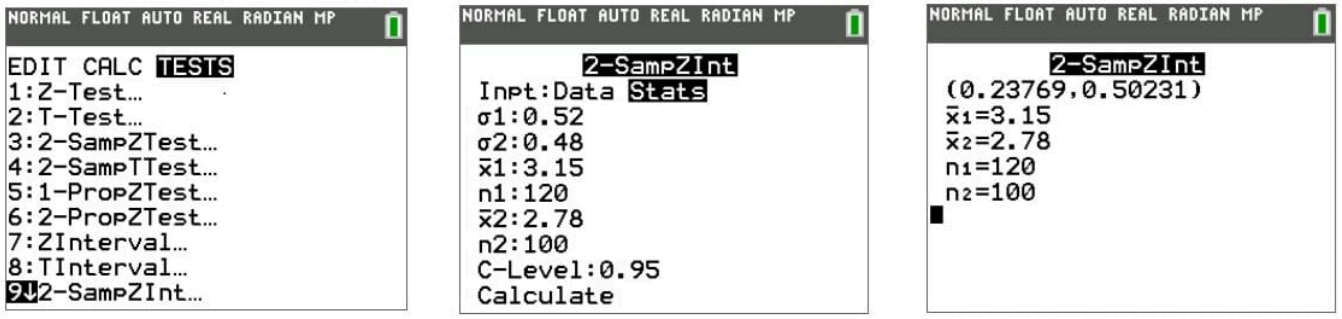


Figure 9.4.1: Copy and Paste Caption here. (Copyright; author via source)

Confidence Interval for the Difference of Two Means using t-values

This formula is used when both sample standard deviations s_1 and s_2 and both sample sizes n_1 and n_1 are less than 30.

Definition: Confidence Interval for the Difference of Two Means using t-values

$$(\bar{X}_1 - \bar{X}_2) - t_C \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) - t_C \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where,

- $\bar{X}_1$ = sample mean of the first sample.
- $\bar{X}_2$ = sample mean of the second sample.
- s_1 = sample standard deviation of the first sample.
- s_2 = sample standard deviation of the second sample.
- n_1 = sample size of the first sample.
- n_2 = sample size of the second sample.
- t_C = critical value from the t-distribution for a given confidence level and minimum degrees of freedom.
- The degrees of freedom is the minimum value between $n_1 - 1$ and $n_2 - 1$.

Example 9.4.2

A STEM education researcher is interested in understanding the differences in weekly study habits between math majors and chemistry majors at a local university. She randomly selects two independent groups of students to participate in her study and collects the data presented in the table below. The units are in hours per week. Use a confidence level of 99% to construct a confidence interval for the difference between the two means. Round the calculation results to three decimal places.

Math Majors	Chemistry Majors
$\bar{X}_1$ =10.5	$\bar{X}_2$ =12.9
s_1 =3.6	s_2 =2.68
n_1 =10	n_2 =13

Solution

Step 1) This confidence interval requires t-values since the sample standard deviations are given and both sample sizes are less than 30.

Step 2) Compute the degrees of freedom. It is the minimum value between $n_1 - 1 = 10 - 1 = 9$ and $n_2 - 1 = 13 - 1 = 12$. Thus, the degrees of freedom is 9.

Step 3) Look up the t_C value in the table using the confidence level and degrees of freedom. It is $t_C = 3.25$.

Step 4) Plug information into the formula and calculate using the order of operations.

$$(\bar{X}_1 - \bar{X}_2) - t_C \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) - t_C \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(10.5 - 12.9) - 3.25 \cdot \sqrt{\frac{3.6^2}{10} + \frac{2.68^2}{10}} < \mu_1 - \mu_2 < (10.5 - 12.9) + 3.25 \cdot \sqrt{\frac{3.6^2}{10} + \frac{2.68^2}{10}}$$

$$-2.4 - 4.613 < \mu_1 - \mu_2 < -2.4 + 4.613$$

$$-7.2013 < \mu_1 - \mu_2 < 2.213$$

Since the interval does include zero, this suggests there is no evidence of a difference in study hours between the two groups

Confidence Interval for the Difference Between Two Proportions

The formula is presented in the box below and assumes normality if the following inequalities are valid: $n_1 \cdot p_1 \geq 5$ and $n_2 \cdot p_2 \geq 5$.

Definition: Confidence Interval for the Difference Between Two Proportions

$$(\hat{p}_1 - \hat{p}_2) - z_C \cdot \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}\backslash($$

Where,

- $\hat{p}_1$ = sample proportion for sample 1, calculated as $\hat{p}_1 = \frac{X_1}{n_1}$
- $\hat{p}_2$ = sample proportion for sample 2, calculated as $\hat{p}_2 = \frac{X_2}{n_2}$
- $\hat{q}_1$ = complement of sample proportion for sample 1, calculated as $\hat{q}_1 = 1 - \hat{p}_1$
- $\hat{q}_2$ = complement of sample proportion for sample 2, calculated as $\hat{q}_2 = 1 - \hat{p}_2$
- p_1 = the proportion of population 1.
- p_2 = the proportion of population 2.
- X_1 = number of successes in sample 1.
- X_2 = number of successes in sample 2.
- n_1 = sample size of population 1.
- n_2 = sample size of population 2.
- Z_C = critical value from the standard normal distribution for a given confidence level.

Example 9.4.3

A community college is interested in comparing the proportions of students who take classes during the **winter intercession** versus the **summer intercession**. A survey of students yielded the results listed in the table below. Construct a confidence interval for the difference of the 2 proportions using a confidence level of 98%. Round all calculations to three decimal places.

Winter Intercession	Summer Intercession
$X_1 = 98$	$X_2 = 178$

$n_1 = 300$	$n_2 = 400$

Solution

Step 1) Compute the sample proportions and their complements.

$\hat{p}_1 = \frac{X_1}{n_1} = \frac{98}{300} = 0.327$

$\hat{q}_1 = 1 - 0.327 = 0.673$

$\hat{p}_2 = \frac{X_2}{n_2} = \frac{165}{400} = 0.445$

$\hat{q}_2 = 1 - 0.413 = 0.555$

Step 2) Look up the critical value in the table or a graphing calculator. The value is $Z_C = 2.33$

Step 3) Plug information into the formula and calculate using the order of operations.

$(\hat{p}_1 - \hat{p}_2) - z_C \cdot \sqrt{\frac{\hat{p}_1 \cdot \hat{q}_1}{n_1} + \frac{\hat{p}_2 \cdot \hat{q}_2}{n_2}}$\(

$(0.327 - 0.445) - 2.33 \cdot \sqrt{\frac{0.327 \cdot 0.673}{300} + \frac{0.445 \cdot 0.555}{400}}$\(

$-0.118 - 0.086$ \(

-0.204 \(

Since the interval includes 0, there is no statistically significant evidence of a difference in the proportions of students taking classes during these two intercessions at the 98% confidence level.

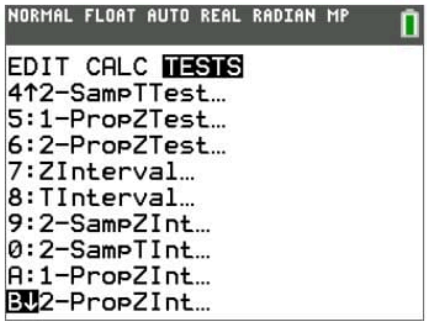
The TI-84+ can be used to compute the confidence interval as well.

Step 1) Click on the "stat" button and use the right arrow to select "TESTS." Use the down arrow to select option B. This option has the title "2-PopZint."

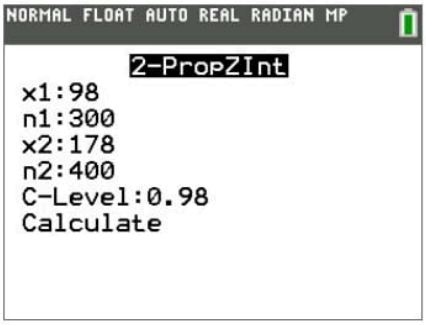
Step 2) Under "2-PropZint," select "Stats." Type in the given data and select "Calculate."

Step 3) The output is at the top of the screen. Round the two answers to three decimal places. They are written in parentheses and are separated by a comma. After rounding the confidence interval is (–0.204, –0.033). The right endpoint slightly differs from the answer computed using the formula. The reason is that the calculator did not round intermediate steps as done when using the formula.

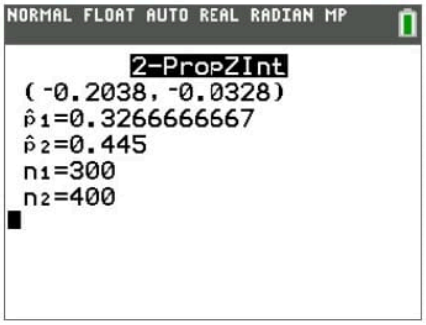
The image below captures all three steps.



Select TESTS and option B.



Enter the data.



The results are outputted at the top of the screen.

Figure 9.4.1: Copy and Paste Caption here. (Copyright; author via source)

Author

"9.4: Confidence Intervals for the Difference Between Two Means and Proportions" by Alfie Swan is licensed under CC BY 4.0

This page titled 9.4: Confidence Intervals for the Difference Between Two Means and Proportions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

9.5: Formulas for Chapter 9

Formulas for Chapter 9

Test Statistic for the z-Test for the Difference of Two Means

The test for the difference between two means is used to determine whether there is a statistically significant difference between the means of two independent populations. It compares the observed difference in sample means to the hypothesized difference in population means. In most cases, the null hypothesis assumes that the population means are equal, meaning $\mu_1 - \mu_2 = 0$, but the formula allows for testing any hypothesized value.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ are the sample means from the two independent samples
- μ_1, μ_2 are the population means being compared
- σ_1^2, σ_2^2 are the population variances of the two groups
- n_1, n_2 are the sample sizes of the two groups
- z is the test statistic used to evaluate the difference between the means

Test Statistic for the t-Test for the Difference of Two Means

The t-test for the difference between two means is used when comparing the means of two independent samples, and the population standard deviations are unknown. It helps determine whether the observed difference in sample means reflects a real difference in the populations or is due to sampling variability. The null hypothesis typically assumes that the population means are equal, so $\mu_1 - \mu_2 = 0$. However, the formula is written to accommodate any hypothesized difference between the population means.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ are the sample means from the two independent samples
- μ_1, μ_2 are the population means being compared
- s_1^2, s_2^2 are the sample variances of the two groups
- n_1, n_2 are the sample sizes of the two groups
- t is the t-test statistic used to assess the difference between means

Test Statistic for the z-Test for the Difference of Two Proportions

The z-test for the difference of proportions is used to determine whether there is a statistically significant difference between the proportions of two independent groups. It is commonly applied in survey or experimental contexts when comparing two categories, such as success rates or approval levels, across different populations. The test assumes that the sample sizes are large enough for the normal approximation to be valid.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where:

- $\hat{p}_1 = \frac{x_1}{n_1}$ is the sample proportion from the first group
- $\hat{p}_2 = \frac{x_2}{n_2}$ is the sample proportion from the second group
- p_1, p_2 are the population proportions being compared (commonly $p_1 - p_2 = 0$ under the null hypothesis)
- $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ is the pooled sample proportion
- $\bar{q} = 1 - \bar{p}$ is the complement of the pooled sample proportion
- n_1, n_2 are the sample sizes of the two groups
- z is the test statistic used to assess whether the observed difference is statistically significant

Authors

"9.5: Formulas for Chapter 9" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 9.5: Formulas for Chapter 9 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

9.6: Chapter 9 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Confidence interval: a range of values, derived from sample data, that is likely to contain the true value of an unknown population parameter, expressed with a specified level of confidence (such as 90%, 95%, or 99%).

Confidence interval for the difference of two means: estimates the true difference between two population means by using the difference of sample means, the standard error of the difference, and a critical value from the z or t distribution.

Confidence interval for the difference of two proportions: estimates the difference between two population proportions using the sample proportions, the standard error of their difference, and a critical z value.

Hypothesis testing: a statistical procedure used to evaluate claims or questions about population parameters using sample data. It requires setting up a null hypothesis (usually a statement of no effect or no difference) and an alternative hypothesis (a statement indicating the presence of an effect or difference), then using sample evidence to decide whether to reject the null hypothesis.

The null hypothesis (H_0): the default assumption that there is no effect or no difference between groups or variables.

The alternative hypothesis (H_1): the statement that there is a true effect or difference in the population.

Hypothesis testing for the difference between two means: evaluates whether there is a statistically significant difference between the means of two independent groups, based on sample data. This test is used when comparing, for example, average scores, measurements, or outcomes.

Z-test for the difference of two means: a hypothesis test that is used when the population standard deviations are known or the sample sizes are large. It uses the standard normal distribution to assess statistical significance.

T-test for the difference of two means: a hypothesis test that is used when population standard deviations are unknown and sample sizes are small. It relies on the t-distribution and takes into account the variability in both samples.

Hypothesis testing for the difference between two proportions: evaluates whether the proportion of a certain outcome in one population is significantly different from that in another. It uses the sample proportions and a pooled estimate to calculate the test statistic.

Test statistic: a standardized value computed from sample data, used to decide whether to reject the null hypothesis. For differences in means or proportions, this often takes the form of a z or t value.

P-value: the probability, assuming the null hypothesis is true, of obtaining a result as extreme as or more extreme than the observed sample result. A small p-value indicates strong evidence against the null hypothesis.

Level of significance (α): is the threshold for rejecting the null hypothesis, often set at 0.05. If the p-value is less than alpha, the null hypothesis is rejected.

Standard error: the estimated standard deviation of a sampling distribution. It measures how much a sample statistic is expected to vary from the true population parameter.

Key Symbols

α - Significance level. The probability of rejecting the null hypothesis when it is true (Type I error).

CL - Confidence level. The probability that a confidence interval contains the true population parameter.

H_0 - Null hypothesis. The default assumption that there is no effect or no difference between groups or variables.

H_1 - Alternative hypothesis. The claim that there is a true effect or difference in the population.

$\hat{p}_1$ - Sample proportion from the first group. The observed proportion of successes in the sample from the first population.

$\hat{p}_2$ - Sample proportion from the second group. The observed proportion of successes in the sample from the second population.

μ_1 - Population mean of the first group. The true average value in the first population.

μ_2 - Population mean of the second group. The true average value in the second population.

n_1 - Sample size of the first group. The number of observations in the first sample.

n_2 - Sample size of the second group. The number of observations in the second sample.

p_1 - Population proportion of the first group. The true proportion of successes in the first population.

p_2 - Population proportion of the second group. The true proportion of successes in the second population.

$\bar{p}$ - The pooled sample proportion is the combined proportion of successes from both samples, used in hypothesis testing for the difference between two proportions.

p-value - The probability of observing the sample result, or more extreme, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against H_0.

$\bar{q}$ -The pooled complement of the sample proportions is the complement of the pooled proportion, representing the proportion of failures.

t - T-test statistic. Used in hypothesis tests when the population standard deviation is unknown and the sample size is small.

$t_{\alpha/2}$ - Critical t-value. The value from the t-distribution corresponding to the desired confidence level and degrees of freedom.

Test statistic - A standardized value (such as z or t) calculated from sample data and used to determine whether to reject the null hypothesis.

z - Z-test statistic. Used when the population standard deviation is known or the sample size is large.

$z_{\alpha/2}$ - Critical z-value. The z-score corresponding to the desired confidence level (e.g., 1.96 for 95\%).

$\bar{x}_1$ - Sample mean from the first group. The average of the sample data from the first group.

$\bar{x}_2$ - Sample mean from the second group. The average of the sample data from the second group.

Authors

"9.6: Chapter 9 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 9.6: Chapter 9 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10: CORRELATION AND LINEAR REGRESSION

10.1: Bivariate Data and Scatter Plots
10.2: Correlation Coefficient
10.3: Hypothesis Test for Correlation Using r
10.4: Hypothesis Test for a Correlation Using t-Test
10.5: Linear Regression
10.6: Coefficient of Determination and the Standard Error of the Estimate
10.7: Formulas for Chapter 10
10.8: Chapter 10 - Key Terms and Symbols

This page titled 10: Correlation and Linear Regression is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10.1: Bivariate Data and Scatter Plots

Learning Objectives

- Define bivariate data and recognize that it involves pairs of related values.
- Interpret scatter plots to identify patterns, trends, or relationships between variables.
- Use scatter plots to detect correlations, clusters, or outliers in a dataset.

Bivariate data involves two variables that are analyzed together to determine their relationship. One common way to visualize bivariate data is through a **scatterplot**, a graph displaying data points on a coordinate plane. Each point represents a pair of values, plotting one variable on the x-axis and the other on the y-axis. Scatterplots help identify patterns, trends, and correlations between the variables, such as positive, negative, or no correlation, making them useful for statistical analysis and predictions.

Definition: Bivariate Data

Bivariate data involves two variables, typically measured simultaneously for each observation or data point. Each variable is paired, and each pair represents a single observation in the dataset. The pairs are represented using the variables (x, y) where x is called the independent variable and y is called the dependent variable.

Bivariate data is used to analyze and understand the relationship between two variables. Below are a few key reasons for using bivariate data.

1. **Exploring Relationships**:
   Bivariate data allows you to determine if and how two variables are related. For example, you might explore the relationship between study hours and test scores.

2. **Identifying Correlation**:
   It helps quantify the strength and direction of the relationship between two variables. Correlation coefficients (e.g., Pearson's r) are often used.

3. **Predicting Outcomes**:
   By analyzing bivariate data, you can predict the value of one variable based on the value of another. For example, using height to predict weight.

Example 10.1.1

An educational researcher is interested in studying the relationship between the amount of sleep (measured in hours) a student gets the night before an exam and their performance on the exam (measured out of 100 points). The collected data will consist of ordered pairs, where the first value represents the hours of sleep and the second value represents the exam score. One data point collected during the study was (6, 87). Use this pair to answer the following questions.

1. Determine the independent variable value in this context and explain what it represents.
2. Determine the dependent variable value in this context and explain what it represents.

Solution

1. The independent variable value is 6, representing the number of hours the student slept before the exam.
2. The dependent variable value is 87, representing the number of points out of 100 that a student earned on the exam.

Definition of Independent and Dependent Variables

The independent or explanatory variable (commonly represented as x) explains or causes changes in the response variable. The independent variable can be manipulated or changed by the researcher.

The dependent variable or response variable (commonly represented by y) reflects the outcome of a study. This variable is measured or observed by the researcher. For example, if we are studying how the amount of time spent studying impacts exam scores, study time serves as the predictor variable, while the exam score is the response variable.

In experimental data, it is usually straightforward to identify the independent and dependent variables. However, this distinction can be more challenging in observational data. The dependent variable is the one you aim to study or gain insights about.

When examining the relationship between the unemployment rate and the economic growth rate, it may not be immediately clear which variable should be x and which should be y. The choice depends on whether we aim to predict the unemployment rate or the economic growth rate. It is important to avoid drawing cause-and-effect conclusions from observational data. A strong relationship between these two variables does not necessarily imply that one directly causes changes in the other. Other factors, such as retirements or pandemics, could influence both rates simultaneously.

Scatter Plot

A common method for examining the relationship between independent and dependent variables is to plot all the data points on a coordinate system. This type of graph, known as a scatter plot, helps visualize the relationship (if any) between x and y.

Definition: Scatter Plot

A **scatter plot** is a type of graph used to display and analyze the relationship between two numerical variables. Each data point in a scatter plot represents an ordered pair (x, y) and is plotted on a two-dimensional coordinate system. The independent variable x is typically placed on the horizontal axis, while the dependent variable y is on the vertical axis. Scatter plots are commonly used to identify patterns, trends, correlations, or potential outliers in the data.

Example 10.1.2

An educational researcher is interested in studying the relationship between the amount of sleep (measured in hours) a student gets the night before an exam and their performance on the exam (measured out of 100 points). The collected data will consist of ordered pairs, where the first value represents the hours of sleep and the second value represents the exam score. A list of the collected data points is provided below. Construct a graph and a scatter plot and describe the relationship (if any) between them.

Bivariate Data

x: Hours of Sleep on the Night Before Exam	y: Points (out of 100) Earned on the Exam
8	75
6	86
3	72
4	65
2	68
0	50
12	90
7	98
10	84

Table 10.1.1: The Data Consists of Hours of Sleep the Night Before an Exam and Exam Score

Solution

The scatter plot is plotted on the graph represented in the image below.

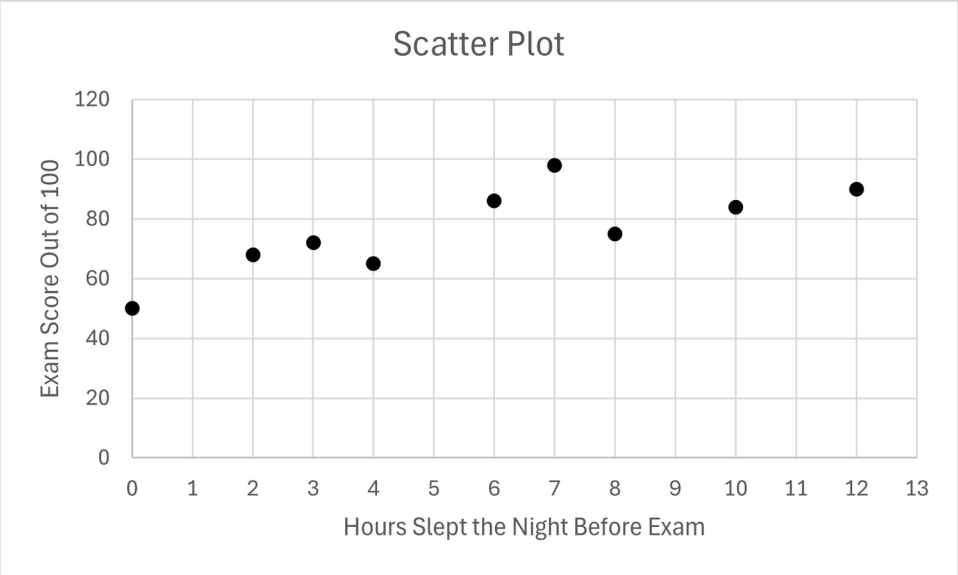


Figure 10.1.1: The Scatterplot for the Hours of Sleep and Corresponding Exam Scores. The relationship between the variables seems to be somewhat linear, as the y-values tend to increase along with the x-values. Further testing is needed to confirm whether this relationship is truly linear.

Author

"10.1: Bivariate Data and Scatter Plots" by Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"12.1: Correlation" by Rachel Webb is licensed under CC BY-SA 4.0

This page titled 10.1: Bivariate Data and Scatter Plots is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 12.1: Correlation by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

10.2: Correlation Coefficient

Learning Objectives

- Define the correlation coefficient (r) and explain its purpose.
- Describe how the correlation coefficient measures the strength and direction of a linear relationship between two variables.
- Interpret the value of r on a scale from -1 to 1.
- Identify that values of r near -1 or 1 indicate stronger linear relationships.

Correlation in terms of two variables measures how much they move together. If one variable increases when the other does, they have a **positive correlation**. If one increases while the other decreases, they have a **negative correlation**. If they don't affect each other, there's **no correlation**. The correlation coefficient will be computed to measure the strength and direction of the correlation between two independent variables x and y.

Definition: Term

The **correlation coefficient**, denoted as r, measures the strength and direction of the linear relationship between two variables. Its purpose is to provide a numerical value that quantifies how closely the variables x and y are related. The range of r is from -1 to 1.

Correlation and Scatter Plots

When the correlation coefficient r is near -1, it indicates a strong negative linear relationship. As the x-values on the horizontal axis increase, the y-values on the vertical axis will decrease. The closer r is to -1, the stronger the linear relationship. The shape of the scatter plot will appear to be linear. An example of a scatter plot with r close to -1 is presented below.

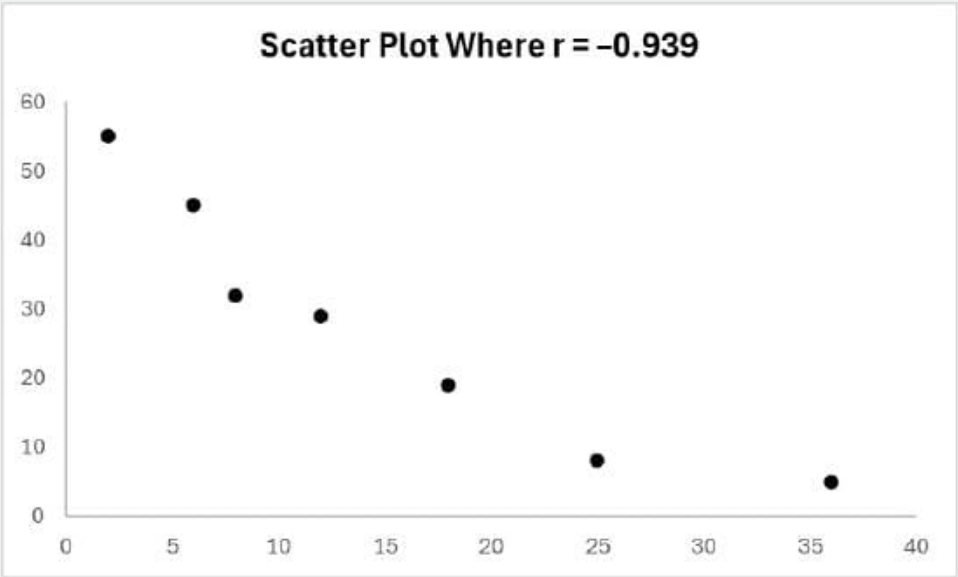


Figure 10.2.1: Scatter Plot Where There is a Strong Negative Linear Relationship

When the correlation coefficient r is near 0, it indicates there isn't a linear relationship between x and y. The scatter plot will have an amorphous shape when r is close to 0 as presented in the image below.

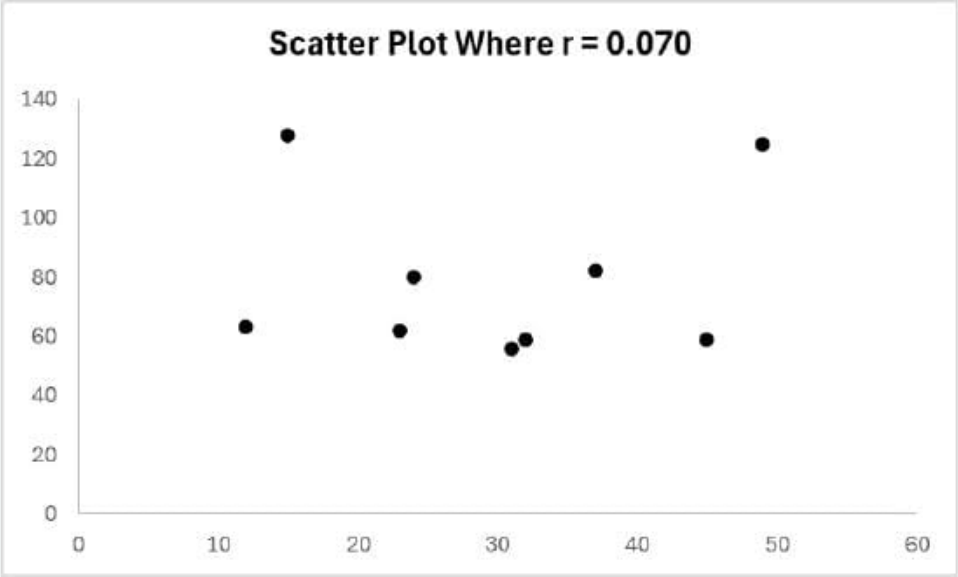


Figure $10.2.2$: Scatter Plot Where There is No Linear Relationship

When the correlation coefficient r is near 1, it indicates a strong positive linear relationship. As the x-values on the horizontal axis increase, the y-values on the vertical axis will also increase. The closer r is to 1 the stronger the linear relationship. The shape of the scatter plot will appear to be linear. An example of a scatter plot with r close to 1 is presented below.

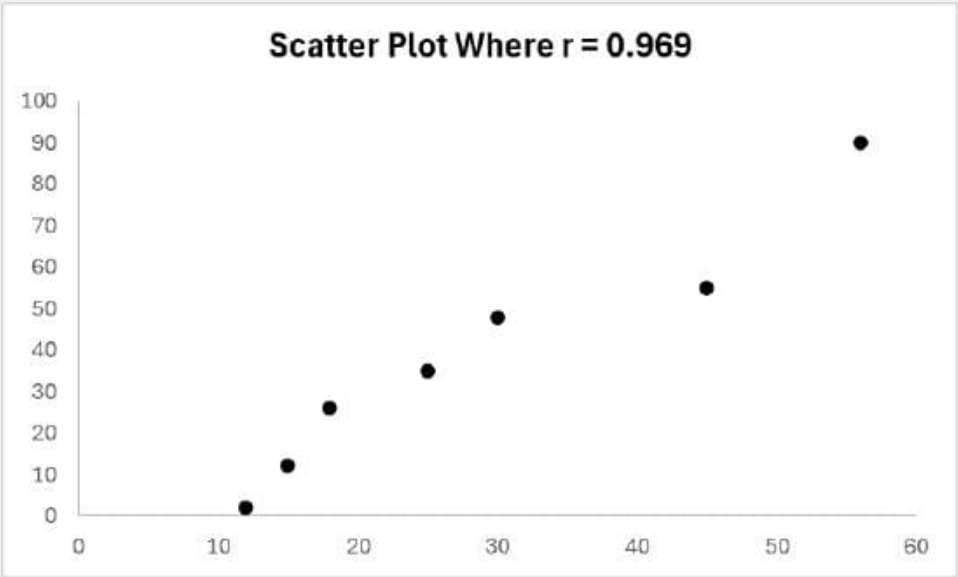


Figure $10.2.3$: Scatter Plot Where There is a Strong Positive Linear Relationship

Correlation Coefficient Formula

Definition: Formula to Compute Correlation Coefficient

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Where,

- n = total number of pairs.
- $\sum x$ = sum of all x values.
- $\sum y$ = sum of all y values.
- $\sum x^2$ = sum of all the squares of the x values.
- $\sum y^2$ = sum of all the squares of the y values.
- $\sum xy$ = sum of all the products of the corresponding x and y values.

Example 10.2.1

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points and time is measured in hours per week. Compute the correlation coefficient for the data set below. Use the formula and round the final answer to three decimal places.

Bivariate Data

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.2.1: Hours Studied Per Week and Midterm Exam Score

Solution

1. Add three columns to the table for xy, x^2, and y^2. Fill in the columns by performing the required computations.

Bivariate Data With Expanded Columns

x	y	xy	x^2	y^2
10	51	510	100	2601
10	53	530	100	2809
12	64	768	144	4096
13	68	884	169	4624

x	y	xy	x^2	y^2
14	71	994	196	5041
15	79	1185	225	6241
16	84	1344	256	7056
20	92	1840	400	8464

Table 10.2.2: Added Columns Needed for Computation of the Correlation Coefficient.

2. Find the sum of each column.

- n = 8
- $\sum x$ = 110
- $\sum y$ = 562
- $\sum x^2$ = 1590
- $\sum y^2$ = 40932
- $\sum xy$ = 8055

3. Plug the information into the formula and compute using the order of operations.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$r = \frac{8(8055) - (110)(562)}{\sqrt{[8(1590) - (110)^2][8(40932) - (562)^2]}}$$

$$r = \frac{64440 - 61820}{\sqrt{[620][11612]}}$$

$$r = \frac{2620}{\sqrt{719944}}$$

$$r = \frac{2620}{2683.17722}$$

$$r \approx 0.976$$

Correlation Coefficient and Technology

Using technology such as a scientific calculator, graphing calculator, Microsoft Excel, or other computational tools is a more efficient way to calculate the correlation coefficient. These methods minimize the risk of errors during calculations and save time by eliminating the need for extensive intermediate computations. Below is an example of how to compute the correlation coefficient using the TI-84+ calculator. All future calculations will be done using the TI-84+ calculator.

Example 10.2.2

Using the data from example 1, compute the correlation coefficient r using a TI-84+ calculator.

Solution

To turn on **Stat Wizard** on a **TI-84+**, follow these steps:

- **Press the [MODE]** on your calculator.
- **Scroll down** using the arrow keys until you find **[STAT WIZARDS]**.
- **Highlight [ON]** using the right arrow key.

- **Press [ENTER]** to confirm your selection.

1. Press the **[STAT]** button, make sure that **[EDIT]** and **[1:Edit]** are selected, then press **[ENTER]**.

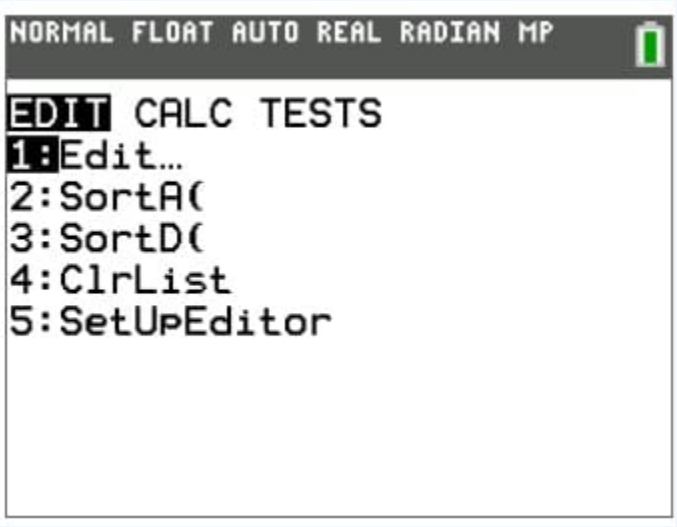


Figure 10.2.4: Selecting Edit Function in TI-84+

2. Enter the x-values in List 1 **[L_1]** and the y-values in List 2 **[L_2]**.

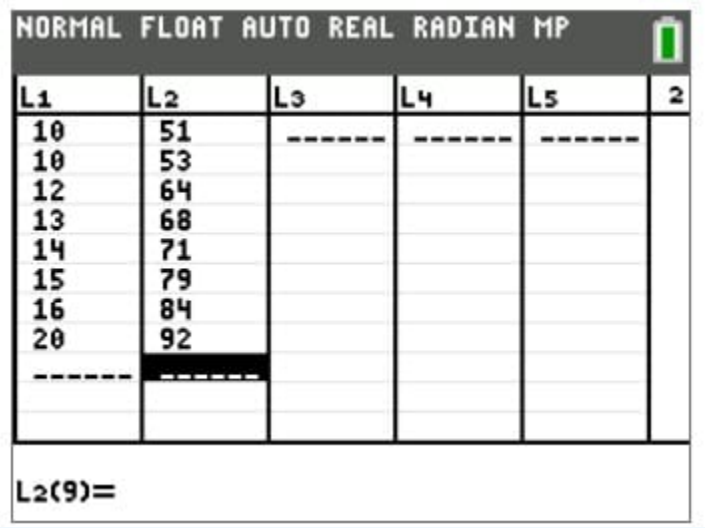


Figure 10.2.5: Enter Data into List One and List Two in TI-84+

3. Press the **[STAT]** button again, use the right arrow to select **[CALC]**, use the down arrow to select **[8: LinReg(a+bx)]**, and then press **[ENTER]**.



Figure 10.2.6: Selection of Linear Regression Function in TI-84+

4. Make sure that Xlist has L_1 and the Ylist has L_2. Use the down arrow to select **[Calculate]** and press [ENTER].

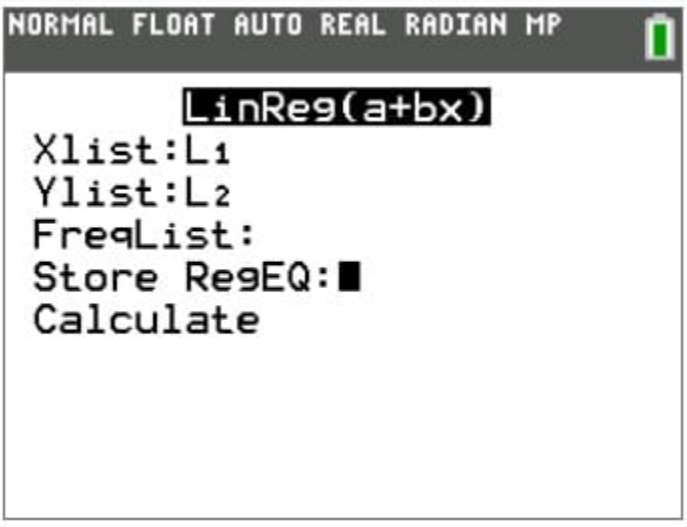


Figure 10.2.7: Check Screen to Ensure the Proper Lists are Selected

5. On the output page, r will be on the last line. After rounding to three places values it is $r = 0.976$.

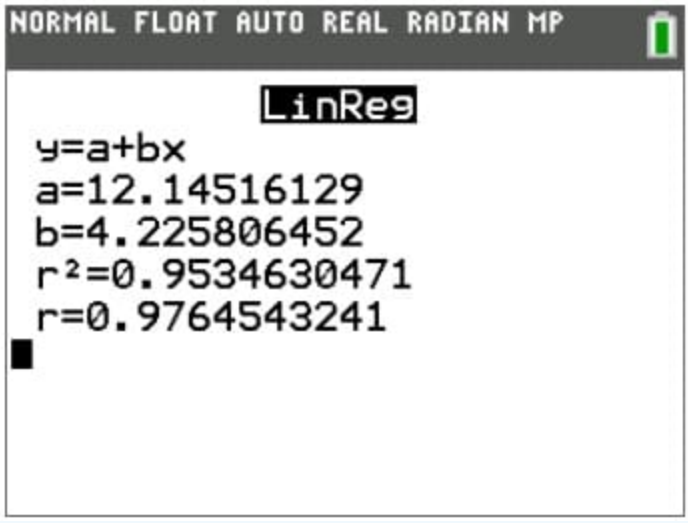


Figure 10.2.8: Output of Correlation Coefficient r = 0.976

Author

"10.2: Correlation Coefficient" by Alfie Swan is licensed under CC BY 4.0

This page titled 10.2: Correlation Coefficient is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10.3: Hypothesis Test for Correlation Using r

Learning Objectives

- Explain the purpose of hypothesis testing for correlation using the correlation coefficient r.
- Determine whether a significant linear relationship exists between two variables in a population through hypothesis testing.
- Compare the sample correlation coefficient to critical values based on sample size.
- Evaluate whether the observed correlation is likely due to chance or reflects a true association in the population.

The correlation coefficient r measures the strength and direction of a linear relationship between two variables. A hypothesis test for r is conducted to determine whether an observed correlation is statistically significant. This test evaluates whether the observed relationship in the sample reflects a true relationship in the population or if it could have occurred due to random chance. The population correlation coefficient ρ, spelled as rho and pronounced as row, will be used to write out the claim for the test. The process for the test is outlined below.

Hypothesis Test for Correlation

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim (There is no linear relationship between the variables)

$H_1 : \rho \neq 0$ (There is a linear relationship between the variables)

2. Look up the critical values using the Pearson's Correlation Matrix (PMC) table below using the given α. The degrees of freedom are $n - 2$. The PMC table can also be accessed in the book by clicking PMC Table.

Degrees of Freedom (n – 2)	α = 0.005	α = 0.01	α = 0.05	α = 0.10
1	0.999	0.999	0.997	0.988
2	0.995	0.990	0.950	0.900
3	0.974	0.959	0.878	0.805
4	0.942	0.917	0.811	0.729
5	0.906	0.875	0.754	0.669
6	0.870	0.834	0.707	0.621
7	0.836	0.798	0.666	0.582
8	0.805	0.765	0.632	0.549
9	0.776	0.735	0.602	0.521
10	0.750	0.708	0.576	0.497
11	0.726	0.684	0.553	0.476
12	0.703	0.661	0.532	0.458
13	0.683	0.641	0.514	0.441
14	0.664	0.623	0.497	0.426
15	0.647	0.606	0.482	0.412
16	0.631	0.590	0.468	0.400
17	0.616	0.575	0.456	0.389
18	0.602	0.561	0.444	0.378
19	0.589	0.549	0.433	0.369
20	0.576	0.537	0.423	0.360
21	0.565	0.526	0.413	0.352
22	0.554	0.515	0.404	0.344
23	0.543	0.505	0.396	0.337
24	0.534	0.496	0.388	0.330
25	0.524	0.487	0.381	0.323
26	0.515	0.479	0.374	0.317
27	0.507	0.471	0.367	0.311
28	0.499	0.463	0.361	0.306

Figure 10.3.1: Correlation Coefficient Table

3. Represent the range of correlation coefficient values on a spectrum of r values. The minimum value is -1 and the maximum value is 1.



Figure 10.3.2: The Scale for the Correlation Values

4. Compute r (correlation coefficient) using the formula, a calculator, or another technology.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

5. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

- If r falls in the left critical region (shaded area to the left of the left critical value), then reject H_0 and there is a negative linear relationship.
- If r falls in the non-critical region (between the two critical values), then do not reject H_0, and there is no linear relationship.
- If r falls in the right critical region (shaded area to the right of the right critical value), then reject H_0 and there is a positive linear relationship.

Example of a Test for Correlation

Example 10.3.1

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points, and time is measured in hours per week. Test for correlation using an $\alpha = 0.05$. Please click on the link PMC table to access the table in the book.

Bivariate Data

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.3.1: Hours Studied Per Week and Midterm Exam Scores

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the table below, using the given $\alpha = 0.05$.

Since n represents the number of pairs, which is 8, then $d.f. = 8 - 2 = 6$. Using the table above, the critical values are ± 0.707 .

3. Represent the range of correlation coefficient values on a spectrum of r values.

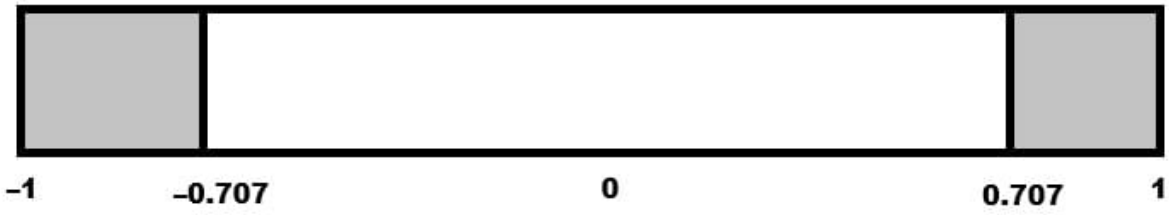


Figure 10.3.3: The Scale of the Correlation Coefficient Values With Critical Values ± 0.707

4. Compute r (correlation coefficient) using the formula, a calculator, or another technology.

This value was computed in 10.2. It is $r = 0.976$.

5. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

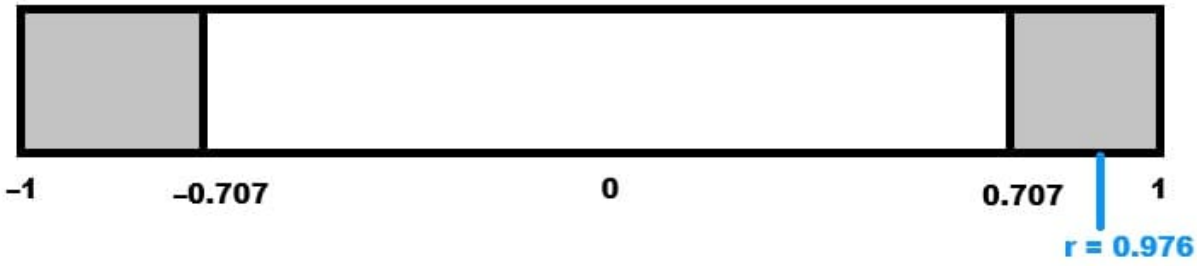


Figure 10.3.4: The Correlation Coefficient Value of r = 0.976 Falls into the Right Critical Region

Since r falls in the right critical region, reject H_0 and there is a positive linear relationship.

Example 10.3.2

A **health researcher** at the **Health Department** at a large university is conducting a study to explore the relationship between physical activity and health outcomes among college students aged **18–25 years old**. The researcher is specifically interested in determining whether there is a correlation between the number of hours students work out per week and the number of days they spend being ill in a year. The researcher collected data provided in the table below. Test for correlation using an $\alpha = 0.01$. Please click on the link PMC table to access the table in the book.

Bivariate Data

X: Hours Worked Out per Week	Y: Days Spent Ill in a Year
0	14
2	10
4	8
5	6

X: Hours Worked Out per Week	Y: Days Spent Ill in a Year
7	5
10	3
12	2

Table 10.3.2: Hours Worked Out Per Week and Days Spent Ill in a Year

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the table below using the given $\alpha = 0.01$.

Since n represents the number of pairs, which is 8, then $d.f. = 7 - 2 = 5$. Using the table above, the critical values are ± 0.875 .

3. Represent the range of correlation coefficient values on a spectrum of r values.

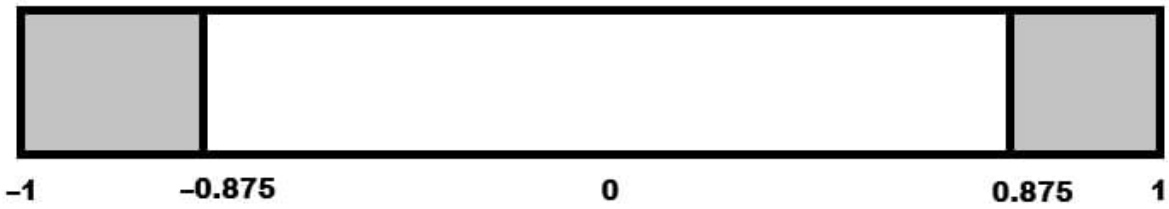


Figure 10.3.5: The Scale of the Correlation Coefficient Values With Critical Values ± 0.875

4. Compute r (correlation coefficient) using the formula, a calculator, or another technology.

A Ti-84+ was used to compute the value. It is $r = -0.964$.

5. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

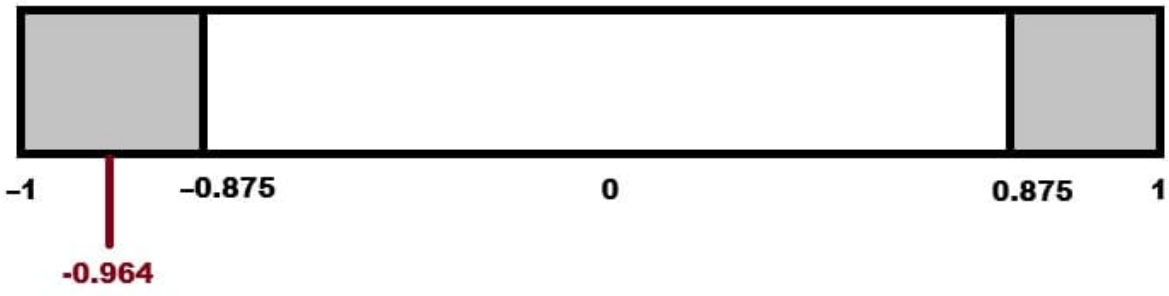


Figure 10.3.6: The Correlation Coefficient Value of r = -0.964 Falls into the Left Critical Region

Since r falls in the left critical region, reject H_0, and there is a negative linear relationship.

Example 10.3.3

A researcher is exploring if there is any correlation between the amount of money students spend on lunch and their GPA in a college setting. Hypothetically, we are testing if students who spend more money on lunch tend to have higher or lower GPAs.

The researcher collected 10 pairs of data representing the amount of money students spend on lunch and their corresponding GPA. The data is presented below. Test for correlation using $\alpha = 0.05$. Please click on the link PMC table to access the table in the book.

Bivariate Data

Amount Spent on Lunch ($)	GPA
$ 10.00	1.95
$ 7.50	3.20
$ 4.00	3.60
$ 8.45	2.80
$ 6.95	3.40
$ 9.00	2.70
$ 8.90	2.56
$ 12.50	3.30
$ 19.80	3.00
$ 6.90	3.49

Table 10.3.3: Amount Spent on Lunch and GPA

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the table below, using the given $\alpha = 0.05$.

Since n represents the number of pairs, which is 10, then $d.f. = 10 - 2 = 8$. Using the table above, the critical values are ± 0.632.

3. Represent the range of correlation coefficient values on a spectrum of r values.

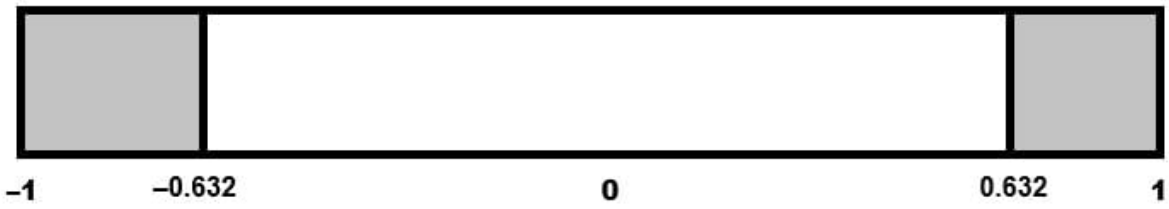


Figure $10.3.7$: The Scale of the Correlation Coefficient Values With Critical Values ± 0.632

4. Compute r (correlation coefficient) using the formula, a calculator, or another technology.

A Ti-84+ was used to compute the value. It is $r = -0.256$.

5. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

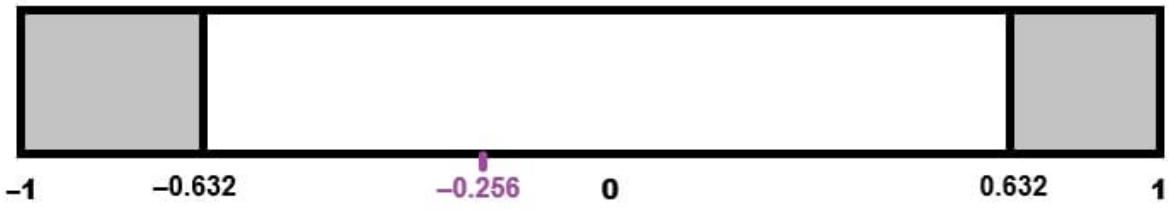


Figure $10.3.8$: The Correlation Coefficient Value of r = -0.256 Falls into the Non-Critical Region

Since r falls in the non-critical region, do not reject H_0, and there is no linear relationship.

Author

"10.3: Hypothesis Test for Correlation Using r" by Alfie Swan is licensed under CC BY 4.0

This page titled 10.3: Hypothesis Test for Correlation Using r is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10.4: Hypothesis Test for a Correlation Using t-Test

Learning Objectives

- Describe the purpose of hypothesis testing for correlation using the t-value method.
- Convert a sample correlation coefficient r to a corresponding t-value.
- Identify the critical t-value using the appropriate degrees of freedom.
- Determine whether the observed correlation provides evidence of a real linear relationship in the population.

A t-test is another hypothesis test used to determine if there is a statistically significant relationship between the independent and dependent variables. The population correlation coefficient ρ (this is the Greek letter rho, which sounds like "row" and is not a p) is the correlation among all possible pairs of data values (x, y) taken from a population.

We will only be using the two-tailed test for a population correlation coefficient ρ. The hypotheses are:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

The null hypothesis of a two-tailed test states that there is no correlation (there is no linear relation) between x and y. The alternative hypothesis states that there is a significant correlation (there is a linear relation) between x and y.

The t-test is a statistical test for the correlation coefficient. It can be used when x and y are linearly related, the variables are random variables, and when the population of the variable y is normally distributed.

Definition: t-Test Formula for Correlation

The formula for the t-test statistic is $t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)}$.

Use the t-distribution with degrees of freedom equal to $df = n - 2$.

Where,

- r = correlation coefficient
- n = sample size (number of pairs)

Note the $df = n - 2$ since we have two variables, x and y.

Examples

Example 10.4.1

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points, and time is measured in hours per week. Test for correlation using an $\alpha = 0.05$.

Click on this link for the t-distribution table to locate the critical values.

Bivariate Data

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.4.1: Hours Studied Per Week and Midterm Exam Score

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the t-table and the given \(\alpha\ = 0.05).

Since n represents the number of pairs, which is 8, then $d.f. = 8 - 2 = 6$. Using the table for t-values and two tails, the critical values are ± 2.447.

3. Compute the test point using the formula, a calculator, or another technology.

The value of $r = 0.976$ can be calculated using a calculator or a formula. Also, n = 8. Plug this information into the formula.

$$t = r\sqrt{\left(\frac{n-2}{1-r^2}\right)} = 0.9765\sqrt{\left(\frac{8-2}{1-0.9765^2}\right)} = 11.09$$

4. Determine if the null hypothesis (H_0) is rejected or not and write out a summary statement.

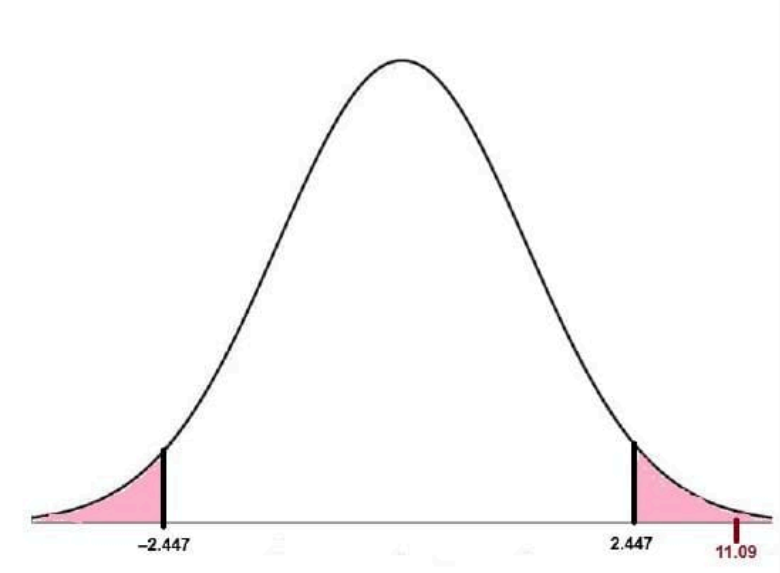


Figure 10.4.1: The Test Point t = 11.09 Falls Into the Right Critical Region

Since the test point is in the critical region, reject (H_0) and there is a linear relationship.

A TI-84+ can also be used to compute the test point more efficiently. The next example will demonstrate the process for this type of computation.

Example 10.4.2

A **health researcher** at the **Health Department** at a large university is conducting a study to explore the relationship between physical activity and health outcomes among college students aged **18–25 years old**. The researcher is specifically interested in determining whether there is a correlation between the number of hours students work out per week and the number of days they spend being ill in a year. The researcher collected data provided in the table below. Test for correlation using an $\alpha = 0.01$.

Click on this link for the t-distribution table to locate the critical values.

Bivariate Data

X: Hours Worked Out per Week	Y: Days Spent Ill in a Year
0	14
2	10
4	8
5	6
7	5
10	3
12	2

Table 10.4.2: Hours Worked Out Per Week and Days Spent Ill in a Year

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the t-table and the given \(\alpha\ = 0.05).

Since n represents the number of pairs, which is 7, then $d.f. = 7 - 2 = 5$. Using the table for t-values and two tails, the critical values are ± 4.032.

3. Compute the test point using the formula, a calculator, or another technology.

In this example, the value will be computed using the calculator.

4. Press **[STAT]** and make sure [**EDIT]** and **[1:Edit]** are highlighted, then press **[ENTER]**. Enter the x-values into L_1 and they-values into L_2.

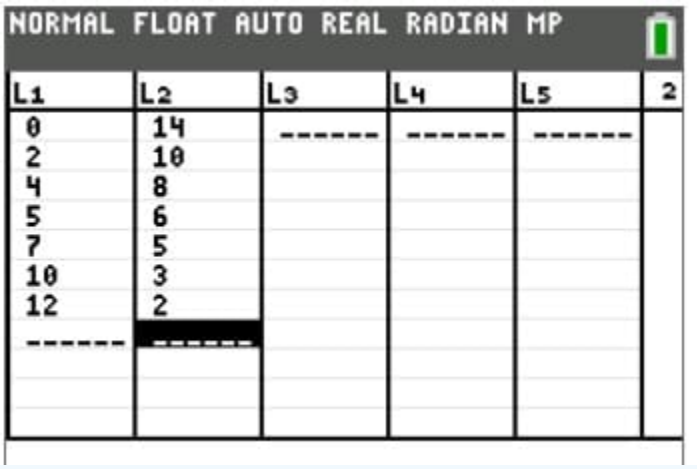


Figure 10.4.2: Enter Data into List One and List Two

5. Press the **[STAT]** key, arrow over to the **[TESTS]** menu, arrow down to the option F **[LinRegTTest]**, and press the **[ENTER]** key. The default is Xlist: L_1, Ylist: L_1, Freq:1, β and $\rho :\neq 0$. Arrow down to Calculate and press the [ENTER] key.

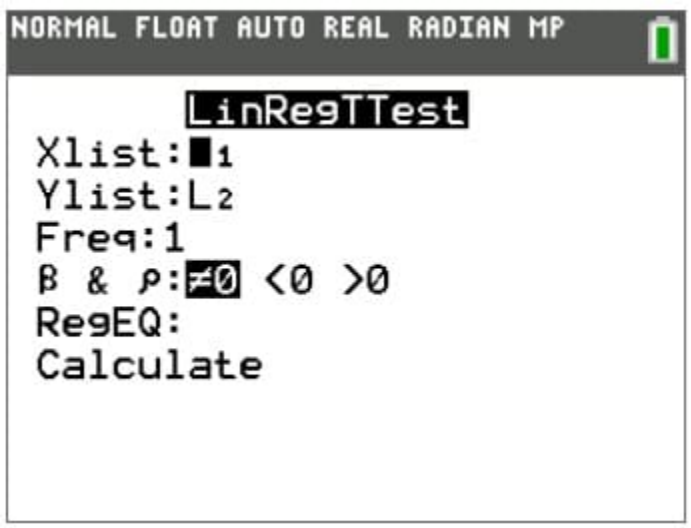


Figure 10.4.3: Enter Information Into Linear Regression Test Screen

6. The output screen on the calculator returns the t-test statistic and the p-value.

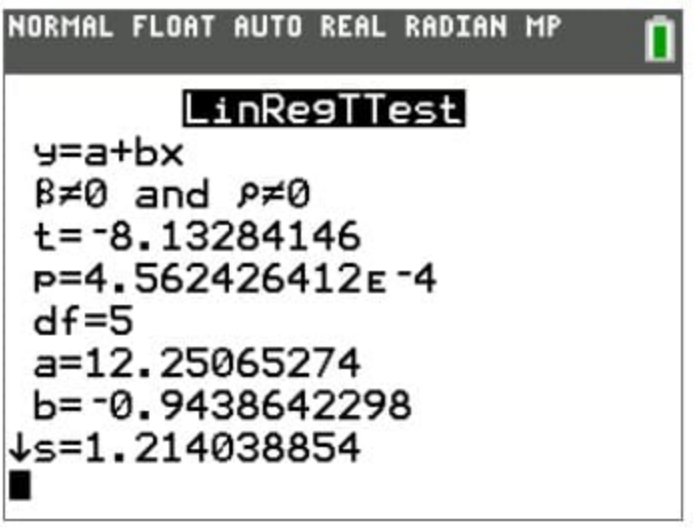


Figure 10.4.4: Output Linear Regression for T-test

The test statistic is $t = -8.13$.

7. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

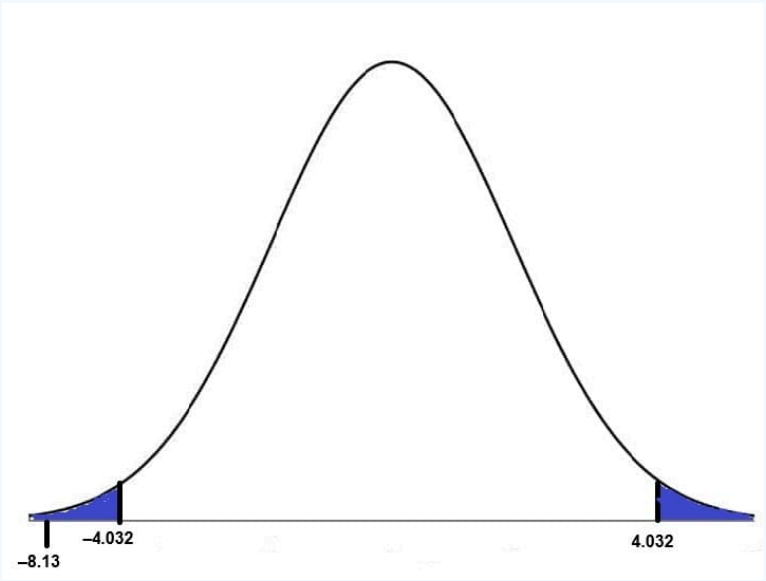


Figure 10.4.5: The Test Point t = -8.13 Falls Into the Right Critical Region

Since the test point is in the critical region, reject (H_0) and there is a linear relationship.

Example 10.4.3

A researcher is exploring if there is any correlation between the amount of money students spend on lunch and their GPA in a college setting. Hypothetically, we are testing if students who spend more money on lunch tend to have higher or lower GPAs.

The researcher collected 10 pairs of data representing the amount of money students spend on lunch and their corresponding GPA. The data is presented below. Test for correlation using $\alpha = 0.05$.

Click on this link for the t-distribution table to locate the critical values.

Bivariate Data

Amount Spent on Lunch ($)	GPA
$ 10.00	1.95
$ 7.50	3.20
$ 4.00	3.60
$ 8.45	2.80
$ 6.95	3.40
$ 9.00	2.70
$ 8.90	2.56
$ 12.50	3.30
$ 19.80	3.00
$ 6.90	3.49

Table 10.4.3: Amount Spent on Lunch and GPA

Solution

1. Write out the claim and the null and alternative hypotheses.

$H_0 : \rho = 0$ Claim

$H_1 : \rho \neq 0$

2. Look up the critical values using the t-table and the given \(\alpha\ = 0.05).

Since n represents the number of pairs, which is 10, then $d.f. = 10 - 2 = 8$. Using the table for t-values and two tails, the critical values are ± 2.036.

3. Compute the test point using the formula, a calculator, or another technology.

In this example, the value will be computed using the calculator. The output of the calculator is presented in the image below.

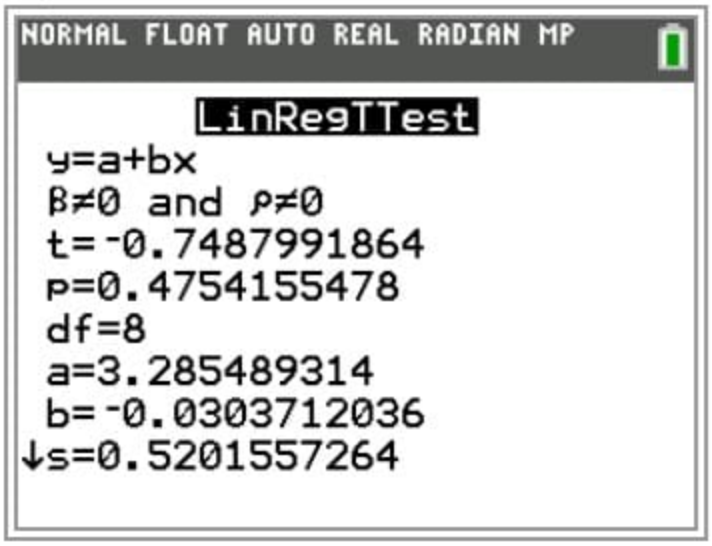


Figure 10.4.6: Output Linear Regression for T-test

The test statistic is $t = -0.75$.

4. Determine if the null hypothesis (H_0) is rejected or not, and write out a summary statement.

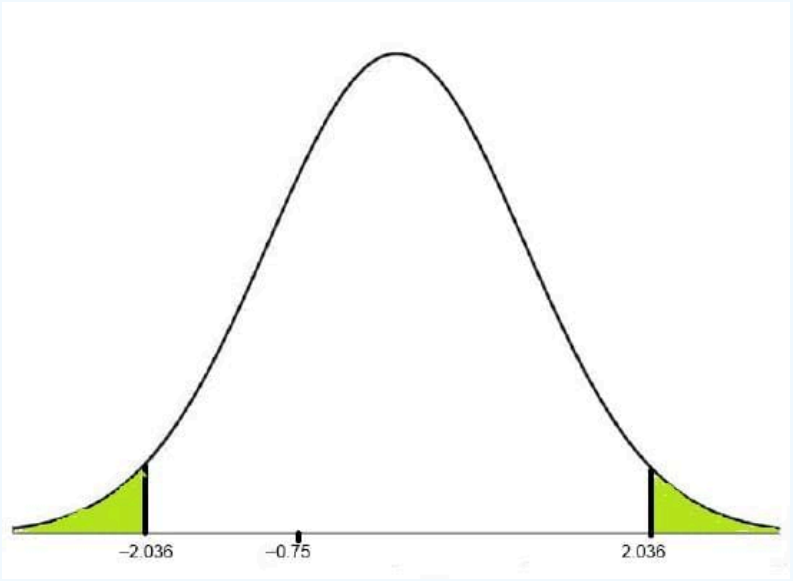


Figure 10.4.7: The Test Point t = -0.75 Falls in Non-Critical Region

Since the test point is in the non-critical region, do not reject (H_0), and there is no linear relationship.

Authors

"10.4: Hypothesis Test for a Correlation Using t-Test" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"12.1.2: Hypothesis Test for a Correlation" by Rachel Webb is licensed under CC BY-SA 4.0

This page titled 10.4: Hypothesis Test for a Correlation Using t-Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

- 12.1.2: Hypothesis Test for a Correlation by Rachel Webb is licensed CC BY-SA 4.0.

10.5: Linear Regression

Learning Objectives

- Explain the purpose of linear regression in modeling the relationship between two variables.
- Identify the components of a linear regression equation: the y-intercept (a) and the slope (b).
- Describe how the y-intercept and slope define the position and steepness of the regression line.
- Use a linear regression model to make predictions based on the trend between variables.

A **line of regression** is a straight line that best represents the relationship between two correlated variables in a dataset. It is used in **regression analysis** to predict a dependent variable's value based on the independent variable's value. The line is determined using the **least squares method**, which minimizes the sum of the squared differences between the observed data points and the predicted values from the line. The line of regression is the line that **best fits** those points. In other words, the line **minimizes the distance** between itself and the data points. Below is a graph of the line of regression with the corresponding equation for the line.

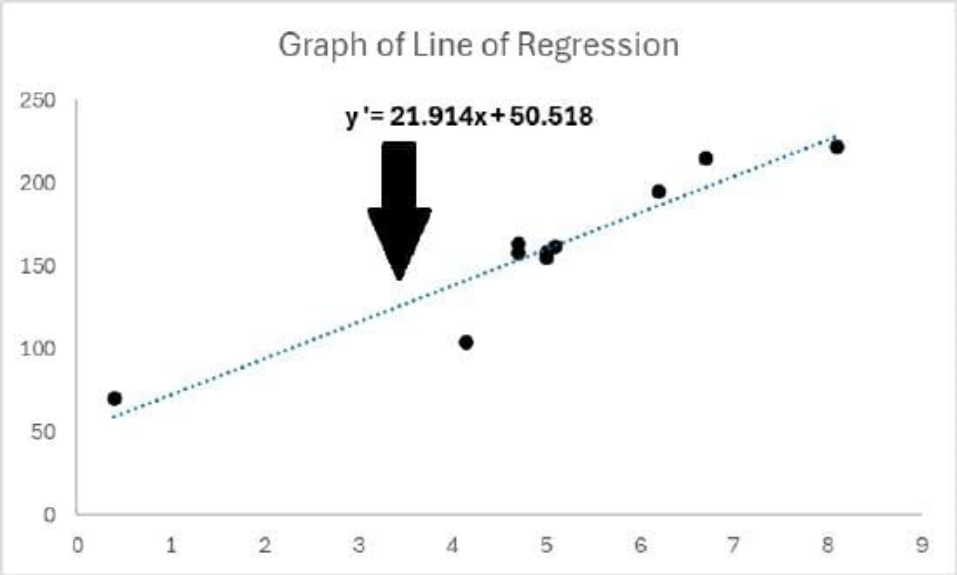


Figure 10.5.1: Graph of Line of Regression and Scatterplot

Definition: Equation for the Line of Regression

$y' = a + bx$

Where,

- y' is the estimated value of y at x. Note that it does not represent the derivative of y as in a calculus class.
- The intercept of the line is $a = \dfrac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$
- The slope of the line is $b = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$

Examples

Example 10.5.1

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points, and time is measured in hours per week. It was shown that the data values are positively correlated in 10.3. Find the line of regression and estimate the exam score if a student studies for 8 hours per week. Round the results to three decimal places.

Data Values

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.5.1 Hours Studied Per Week and Midterm Scores out of 100 Points.

Solution

In this example, the formulas will be used.

Step 1) Add three columns to the table for xy, x^2, and y^2. Fill in the columns by performing the required computations.

Data Values

x	y	xy	x^2	y^2
10	51	510	100	2601
10	53	530	100	2809
12	64	768	144	4096
13	68	884	169	4624
14	71	994	196	5041
15	79	1185	225	6241
16	84	1344	256	7056
20	92	1840	400	8464

Table 10.5.2 Key Columns Needed to Compute the Regression Coefficients a and b

Step 2) Find the sum of each column.

- $n = 8$
- $\sum x = 110$
- $\sum y = 562$

- $\sum x^2$ = 1590
- $\sum y^2$ = 40932
- $\sum xy$ = 8055

Step 3) Plug the information into the formulas for the regression coefficients and compute them using the order of operations.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(562)(1590) - (110)(8055)}{8(1590) - (110)^2} = \frac{893,580 - 886,050}{12,720 - 12,100} = \frac{7,530}{620} = 12.145$$

$$b = \frac{8(8055) - (110)(562)}{8(1590) - (110)^2} = \frac{64,440 - 61,820}{12,720 - 12,100} = \frac{2,620}{620} = 4.226$$

Step 4) Write out the equation for the line of regression using the computed regression coefficients.

$$y' = 12.145 + 4.226x$$

Step 5) Use the line of regression to compute y' (estimated Midterm Exam score) when x = 8 (hours per week). Plug x = 8 into the equation and compute the estimated value.

$$y' = 12.145 + 4.226x = 12.145 + 4.226(8) = 12.145 + 33.808 = 45.953$$

Thus, if a student studies 8 hours per week, the estimated Midterm Exam score is around 46 points.

It is more efficient to work out a linear regression problem using technology. The example above will be reworked using the TI-84+ calculator in the next example.

Example 10.5.2

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points and time is measured in hours per week. It was shown that the data values are positively correlated in 10.3. Find the line of regression and estimate the exam score if a student studies for 8 hours per week. Use the TI-84 + calculator. Round the results to three decimal places.

Data Values

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.5.3 Hours Studied Per Week and Midterm Scores out of 100 Points.

Solution

Step 1) Press the [STAT] button, make sure that [Edit and 1:EDIT] are selected, then press [ENTER].



Figure 10.5.2: Edit Function to Enter Data in TI-84+

Step 2) Enter the x-values in List 1 [L_1] and the y-values in List 2 [L_2].

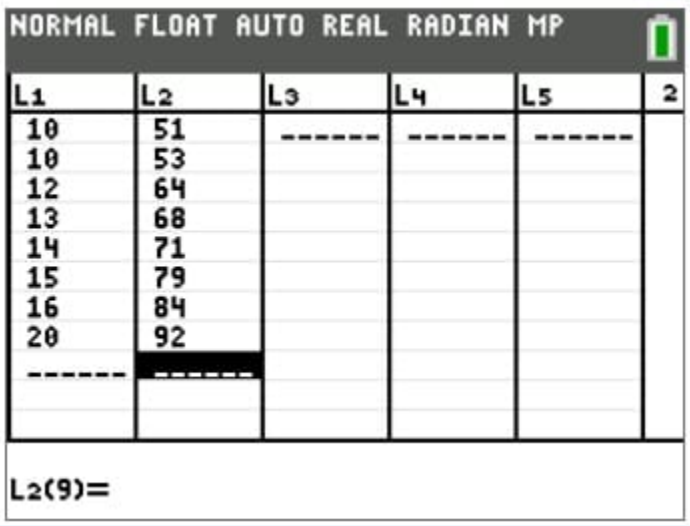


Figure 10.5.3: Lists (L_1 and L_2) Columns Where X and Y Values are Entered

Step 3) Press the [STAT] button again, use the right arrow to select [CALC], use the down arrow to select [8: LinReg(a+bx)], and then press [ENTER].



Figure 10.5.4: Linear Regression Function to Compute Regression Coefficients

Step 4) Make sure that X-list has L_1 and the Y-list has L_2. Use the down arrow to select [Calculate] and press [ENTER].

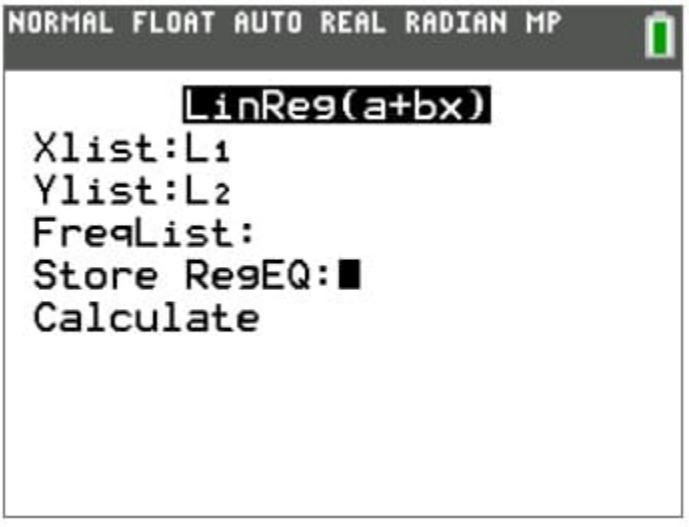


Figure 10.5.5: Check Screen to Ensure Proper Lists are Selected

Step 5) On the output page, a and b will be on the first two lines. After rounding to three places, they are $a = 12.145$ and $b = 4.226$.

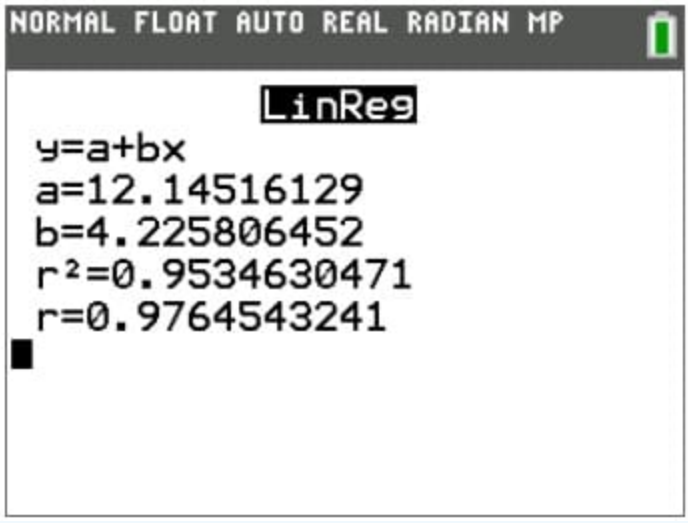


Figure 10.5.6: Output of Regression Coefficients (a = 12.145 and b = 4.226)

Step 6) Use the line of regression to compute y' (estimated Midterm Exam score) when x = 8 (hours per week). Plug x = 8 into the equation and compute the estimated value.

$y' = 12.145 + 4.226x = 12.145 + 4.226(8) = 12.145 + 33.808 = 45.953$

Thus, if a student studies 8 hours per week, the estimated Midterm Exam score is around 46 points.

Example 10.5.3

A **health researcher** at the **Health Department** at a large university is conducting a study to explore the relationship between physical activity and health outcomes among college students aged **18–25 years old**. The researcher is specifically interested in determining whether there is a correlation between the number of hours students work out per week and the number of days they spend being ill in a year. The researcher collected data provided in the table below. It was shown that the data values are positively correlated in 10.3. Find the line of regression and estimate the exam score if a student studies for 9 hours per week. Use the TI-84 + calculator. Round the results to three decimal places.

Data Values

X: Hours Worked Out per Week	Y: Days Spent Ill in a Year
0	14
2	10
4	8
5	6
7	5
10	3
12	2

Table 10.5.4 Hours Worked Out Per Week and Days Ill Per Year.

Solution

In this example, the TI-84+ calculator will be used to compute the regression coefficients. Follow the steps in Example 2 to compute the regression coefficients. The output is provided in the image below.

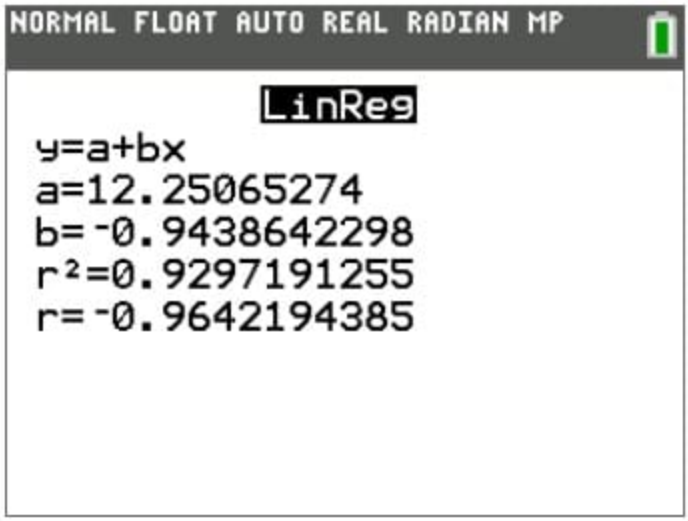


Figure 10.5.7: Output of Regression Coefficients (a = 12.251 and b = –0.964)

Therefore, the line of regression is $y' = 12.251 - 0.944x$.

The estimated value is $y' = 12.251 - 0.944(9) = 12.251 - 8.496 = 3.755$.

Thus, according to the linear model, if a person works out 9 hours per week, they will be ill for around 4 days during the year.

Example 10.5.4

A researcher is exploring if there is any correlation between the amount of money students spend on lunch and their GPA in a college setting. Hypothetically, we are testing if students who spend more money on lunch tend to have higher or lower GPAs.

The researcher collected 10 pairs of data representing the amount of money students spend on lunch and their corresponding GPA. The researcher collected data provided in the table below. It was shown that the data values are not correlated in 10.3. Find the line of regression and estimate the exam score if a student studies for $13.00 on lunch. Use the TI-84 + calculator. Round the results to three decimal places.

Data Values

Amount Spent on Lunch ($)	GPA
$ 10.00	1.95
$ 7.50	3.20
$ 4.00	3.60
$ 8.45	2.80
$ 6.95	3.40
$ 9.00	2.70
$ 8.90	2.56
$ 12.50	3.30
$ 19.80	3.00

Amount Spent on Lunch ($)	GPA
$ 6.90	3.49

Table 10.5.5 Amount Spent on Lunch in Dollars and Grade Point Average (GPA).

Solution

Since there is no correlation, the line of regression is not computed as it is not valid.

Authors

"10.5: Linear Regression" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"10.1: Regression" by Kathryn Kozak is licensed under CC BY-SA 4.0

This page titled 10.5: Linear Regression is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 10.1: Regression by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

10.6: Coefficient of Determination and the Standard Error of the Estimate

Learning Objectives

- Define the coefficient of determination and explain its role in evaluating regression models.
- Interpret the coefficient of determination as a measure of how well the regression line explains variation in the dependent variable.
- Recognize that values closer to 1 indicate a better model fit.
- Define the standard error of the estimate and describe its purpose in measuring prediction accuracy.
- Interpret smaller standard error values as indicators of more accurate predictions and a better model fit.

Two key metrics, the coefficient of determination and the standard error of the estimate, are computed to assess the quality of the line of regression. The **coefficient of determination**, often denoted as r^2 and typically written as a percentage, is a statistical measure that helps assess the quality of a regression model. It provides insights into how well the independent variables explain the variance in the dependent variable. The **standard error of the estimate**, denoted as SE, measures how much the actual data points in a regression model differ from the predicted values. It simply tells you how far off your predictions are, on average.

Coefficient of Determination

The formula below is used to calculate the coefficient of determination; however, it can also be conveniently computed using technology.

Definition: Coefficient of Determination Formula

It is found by squaring r, the correlation coefficient, to get r^2.

The table below presents a scale of key values that assist in interpreting the quality of the line of regression based on the coefficient of determination.

Coefficient of Determination Scale

Value Range	Interpretation
$r^2 = 0\%$	The model explains none of the variance in the dependent variable.
$0\% < r^2 < 25\%$	The model explains very little variance; the independent variables are weakly related to the dependent variable.
$25\% < r^2 < 50\%$	The model explains a moderate amount of variance, but there is significant unexplained variation.
$50\% < r^2 < 75\%$	The model explains a substantial amount of variance; the predictors are relatively strong.
$75\% < r^2 < 100\%$	The model explains most of the variance; it has a high level of explanatory power.
$r^2 = 100\%$	The model explains all of the variance perfectly (typically unrealistic in real-world data).

Table 10.6.1: Scale for Different Coefficient of Determination Values and Their Interpretations

Examples of Coefficient of Determination

Example 10.6.1

In examples 1 and 2 from section 10.5, the line of regression is $y' = 12.145 + 4.226x$ and the correlation coefficient is $r = 0.976$. Calculate the coefficient of determination and explain its significance.

Solution

Method 1) Square r and write the result as a percentage rounded to two place values. The coefficient of determination is $r^2 = 0.976^2 = 0.95258 = 95\%$.

Method 2) Using a TI-84+ calculator, follow the steps in example 2 of section 10.5 to enter the data and calculate the line of regression. The output page of the calculator is presented below and displays r^2 on the fourth line. Convert to a percentage and round to two decimal places to get $r^2 = 95\%$.

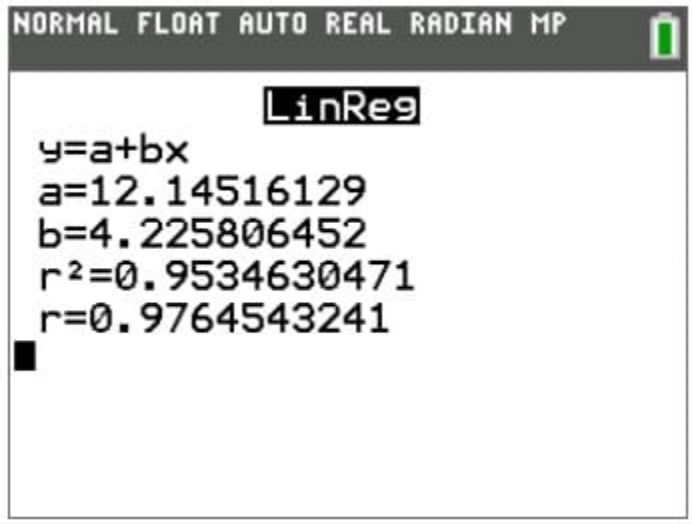


Figure 10.6.1: Output of Coefficient of Determination r^2 = 0.95

Since $r^2 = 95\%$ the line of regression has a high level of explanatory power.

Example 10.6.2

In example 3 from section 10.5, the line of regression is $y' = 12.251 - 0.944x$ and the correlation coefficient is $r = -0.964$. Calculate the coefficient of determination and explain its significance.

Solution

Method 1) Square r and write the result as a percentage rounded to two place values. The coefficient of determination is $r^2 = (-0.964)^2 = 0.92930 = 93\%$.

Method 2) Using a TI-84+ calculator, follow the steps in example 2 of section 10.5 to enter the data and calculate the line of regression. The output page of the calculator is presented below and displays r^2 on the fourth line. Convert to a percentage and round to two decimal places to get $r^2 = 93\%$.

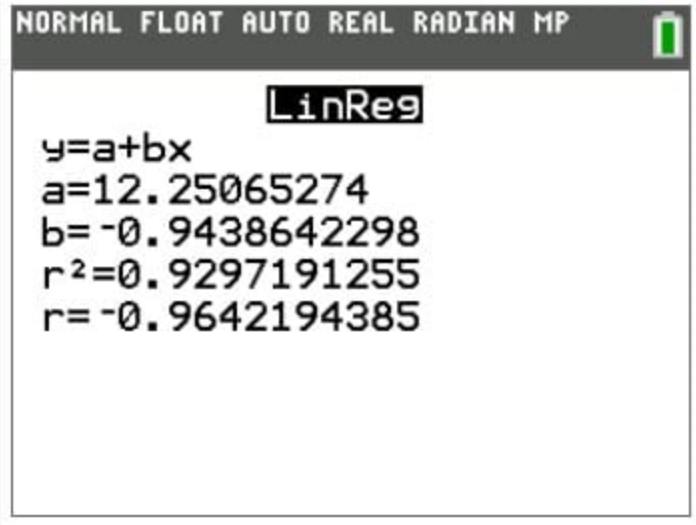


Figure 10.6.2: Output of Coefficient of Determination r^2 = 0.93

Since $r^2 = 95\%$ the line of regression has a high level of explanatory power.

Standard Error of the Estimate

The standard error of the estimate indicates how closely the actual data points align with the regression line. Smaller values of the standard error of the estimate reflect a closer fit between the data points and the regression line. In the ideal case, the standard error of the estimate would be zero, meaning all data points lie exactly on the regression line. The two graphs below illustrate the impact of different standard errors of the estimate, allowing for a comparison of their effects on the regression line.

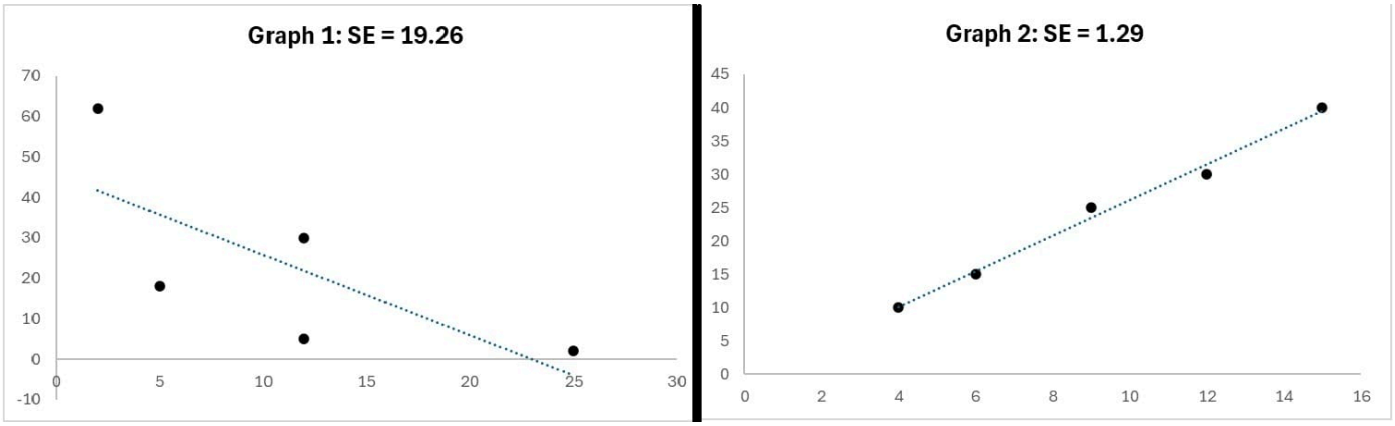


Figure 10.6.3: Graph of Two Scatter Plots Showing Standard Error of the Estimates

- In Graph 1, SE = 19.26, and the actual values are more spread out from the line of regression. This will result in less accurate predictions for the line of regression.
- In Graph 2, SE = 1.29, and the actual values are closer to the line of regression. This will result in more accurate predictions for the line of regression.

Definition: Standard Error of the Estimate Formula

$$SE = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

Where,

- n = number of pairs of data.
- a = regression coefficient of the intercept.
- b = regression coefficient of the slope.
- $\sum y$ = sum of the y-values.
- $\sum y^2$ = sum of the y^2-values.
- $\sum xy$ = sum of the xy-values.

Examples of Standard Error of the Estimate

Example 10.6.3

Professor Martinez is conducting a study to understand the relationship between the number of hours students study per week and their performance on the midterm exam in **Math 400**, an advanced calculus course at the university. She collects data from 8 randomly selected students in her class. The exam is out of 100 points and time is measured in hours per week. It was shown that the data values are correlated in 10.3. Compute the standard error of the estimate.

Bivariate Data

x: Hours Studied Per Week	y: Midterm Exam Score (out of 100 points)
10	51
10	53
12	64
13	68
14	71
15	79
16	84
20	92

Table 10.6.2: Bivariate Data for Hours Studied Per Week and Midterm Exam Score.

Solution

The TI-84+ will be used to compute the sums and regression coefficients.

Step 1) Press the [STAT] button, make sure that [Edit and 1:EDIT] are selected, then press [ENTER].



Figure 10.6.4: Select Edit Function to Enter Data

Step 2) Enter the x-values in List 1 [L_1] and the y-values in List 2 [L_2].

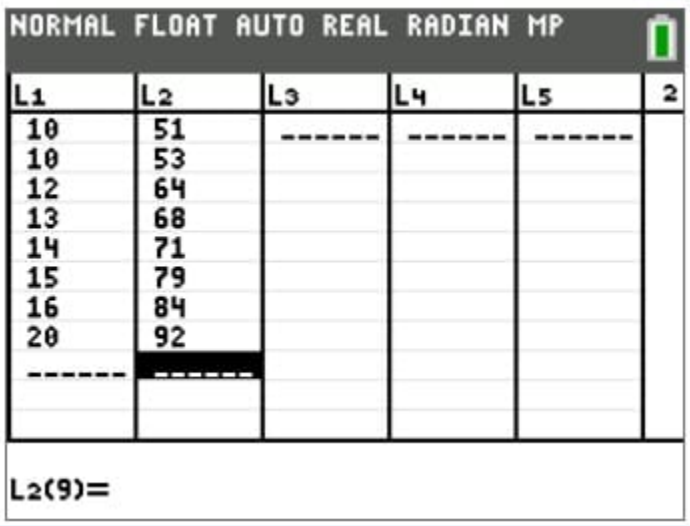


Figure 10.6.5: Enter X-values in List 1 and Y-values in List 2

Step 3) Press the [STAT] button again, use the right arrow to select [CALC], use the down arrow to select [8: LinReg(a+bx)], and then press [ENTER]



Figure 10.6.6: Under Calculate Select Option #8 (Linear Regression Function)

Step 4) Make sure that X-list has L_1 and the Y-list has L_2. Use the down arrow to select [Calculate] and press [ENTER].

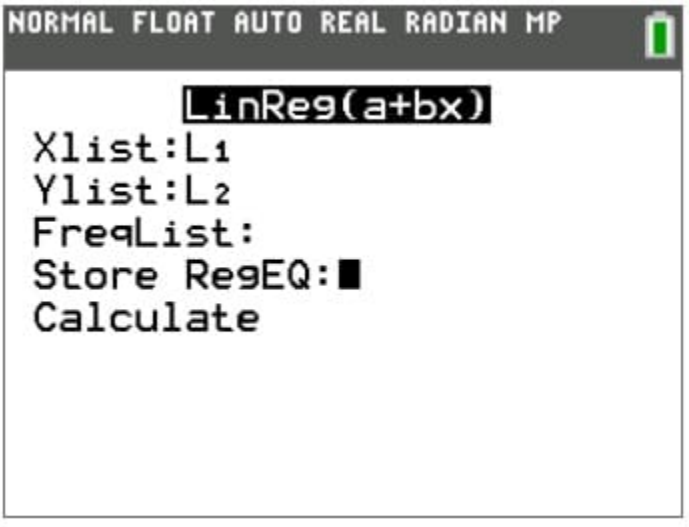


Figure 10.6.7: Check Screen to Make Sure Proper Lists are Selected

Step 4) On the output page, the regression coefficients will be on the first two lines. After rounding to three places values they are $a = 12.145$ and $b = 4.226$.

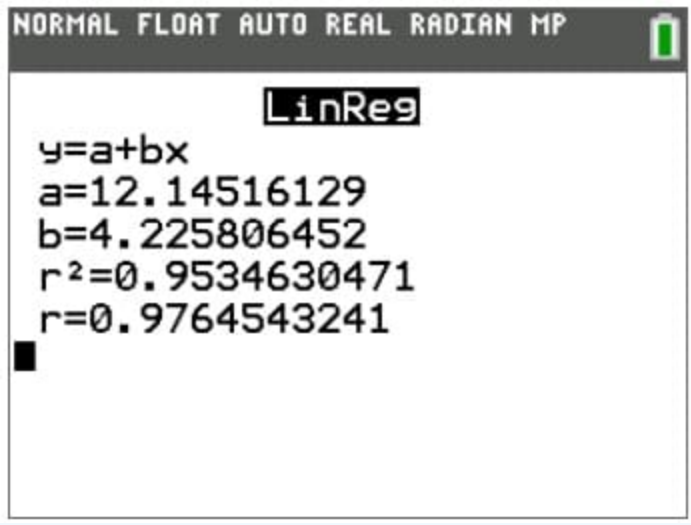


Figure 10.6.8: Output of Regression Coefficients (a = 12.145 and b =4.226)

Step 5) Press the [STAT] button again, use the right arrow to select [CALC], use the down arrow to select [2: 2-VAR STATS], and then press [ENTER]

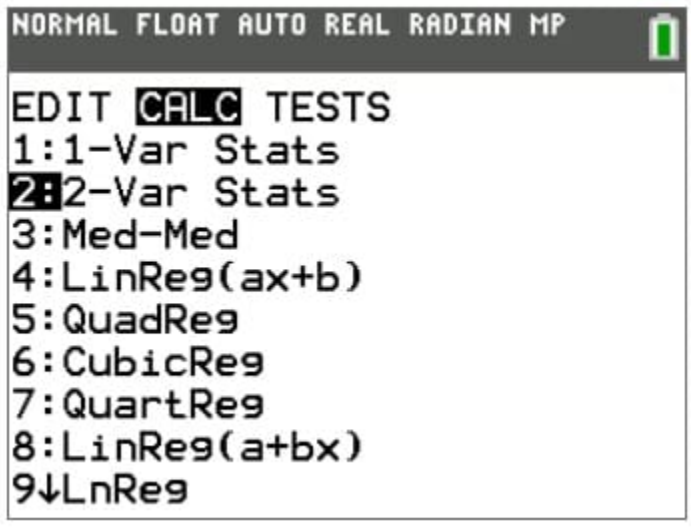


Figure 10.6.9: Select Two Variable Statistics to Compute Sums Needed for Calculations

Step 6) Make sure that X-list has L_1 and the Y-list has L_2. Use the down arrow to select [Calculate] and press [ENTER].



Figure 10.6.10: Check Screen to Make Sure Proper Lists are Selected

Step 7) When the output page appears, scroll down using the down arrow until the following sums appear: $\sum y, \sum y^2, and \sum xy.$

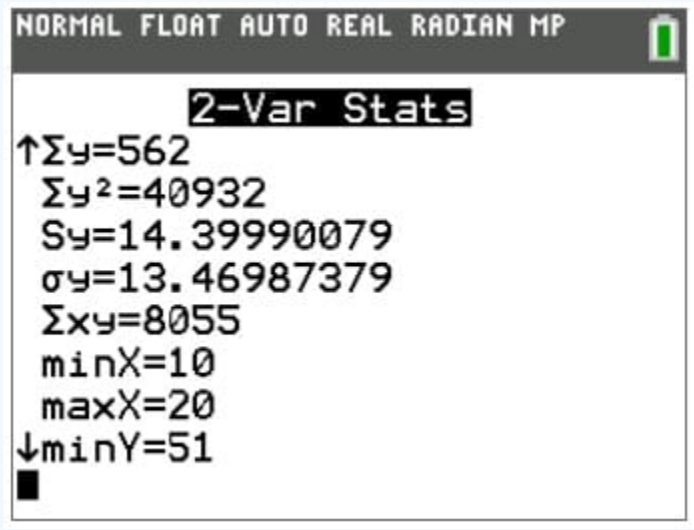


Figure 10.6.11: Output of Sums Needed for Calculations

Step 8) The results of steps 4 and 7 can be plugged into the formula to calculate the standard error of the estimate.

The key values are $n = 8, a = 12.145, b = 4.226, \sum y = 562, \sum y^2 = 40932, and \sum xy = 8055.$

$$SE = \sqrt{\frac{40932 - (12.145)562 - (4.226)8055}{8-2}} = \sqrt{\frac{40932 - 6825.49 - 34040.43}{6}} = \sqrt{\frac{66.08}{6}} = \sqrt{10.01333} = 3.31$$

This is a relatively small value, which means the data values are close to the line of regression and will result in good predictions.

Example 10.6.4

This example will demonstrate how to calculate the standard error of the estimate without using the formula. The TI-84+ calculator has a built-in function that directly calculates the standard error of the estimate. The data from example 3 above will be reused for this second example.

Solution

Step 1) Repeat steps 1 and 2 in example 1 above. The data must be entered into the calculator.

Step 2) Press the [STAT] key, arrow over to the [TESTS] menu, arrow down to the option F [LinRegTTest], and press the [ENTER] key. The default is Xlist: L_1, Ylist: L_1, Freq:1, β, and ρ:≠0. Arrow down to Calculate and press the [ENTER] key.

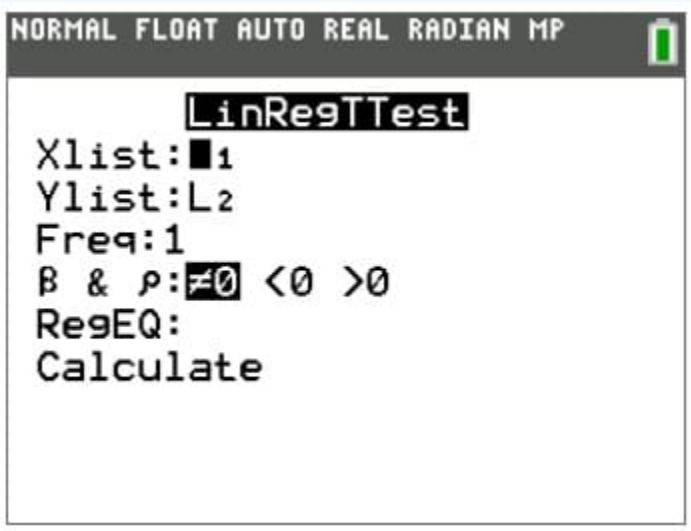


Figure 10.6.12: Make Sure the Proper Lists are Selected After Data Input

Step 3) The output screen on the calculator returns the standard error of the estimate. It is the value s on the last line.

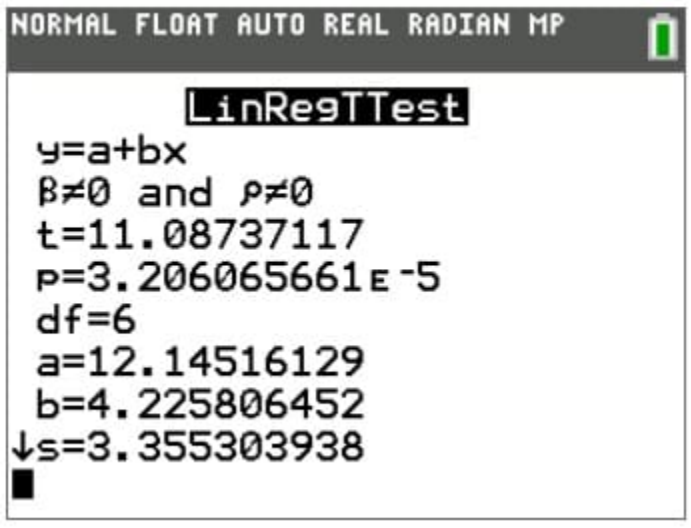


Figure 10.6.13: The Standard Error of the Estimate is s = 3.36

The standard error of the estimate is $SE = 3.36$ (rounded to two places. This result is slightly different from the equation above because this calculated value had no rounding during intermediate steps in the computation. Overall, this is a relatively small value, which means the data values are close to the line of regression and will result in good predictions.

Example 10.6.5

A **health researcher** at the **Health Department** at a large university is conducting a study to explore the relationship between physical activity and health outcomes among college students aged **18–25 years old**. The researcher is specifically interested in determining whether there is a correlation between the number of hours students work out per week and the number of days they spend being ill in a year. The researcher collected data provided in the table below. It was shown that the data values are correlated in 10.3. Compute the standard error of the estimate.

Bivariate Data

X: Hours Worked Out per Week	Y: Days Spent Ill in a Year
0	14
2	10
4	8
5	6
7	5
10	3
12	2

Solution

Following the steps from example 4 above and 2 from 10.4, the output of the t-test is provided below.

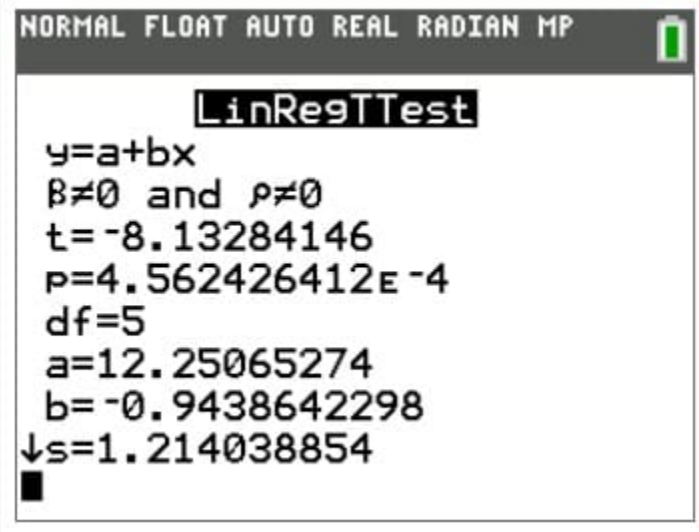


Figure 10.6.14: The Standard Error of the Estimate is s = 1.21

The standard error of the estimate is $SE = 1.21$. Overall, this is a relatively small value, which means the data values are close to the line of regression and will result in good predictions.

Author

"10.6: Coefficient of Determination and the Standard Error of the Estimate" by Alfie Swan is licensed under CC BY 4.0

This page titled 10.6: Coefficient of Determination and the Standard Error of the Estimate is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10.7: Formulas for Chapter 10

Formulas for Chapter 10

Correlation Coefficient

The correlation coefficient, commonly denoted as r, is a statistical measure that describes the strength and direction of a linear relationship between two quantitative variables. Its value ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative linear relationship, and values near 0 suggest little to no linear correlation.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}}$$

Where:

- r is the correlation coefficient
- n is the number of paired data values
- $\sum xy$ is the sum of the products of each x and y value
- $\sum x$ is the sum of all x values
- $\sum y$ is the sum of all y values
- $\sum x^2$ is the sum of the squares of all x values
- $\sum y^2$ is the sum of the squares of all y values

Test Statistic for the t-Test of Correlation

The t-test for the correlation coefficient is used to evaluate whether the observed sample correlation reflects a true linear relationship in the population. It tests the null hypothesis that the population correlation $\rho = 0$, indicating no linear relationship. The formula below shows r multiplied by the square root of the ratio involving the sample size.

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

- t is the test statistic for the correlation
- r is the sample correlation coefficient
- n is the number of paired data points
- $n-2$ is the degrees of freedom for the t-distribution
- $1-r^2$ is the unexplained variation (proportion of variability not explained by the linear relationship)

Line of Regression and Regression Coefficients

The line of regression, also called the least squares regression line, is a linear equation that best fits a set of data by minimizing the sum of the squared differences between the observed values and the predicted values. It is used to describe the relationship between an independent variable $x$and a dependent variable y. The equation of the regression line is displayed below.

$$y' = a + bx$$

Where y' is the predicted value of y for a given x, a is the y-intercept, and b is the slope of the line.

The formulas for calculating the regression coefficients are provided below.

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y \cdot \sum x^2 - \sum x \cdot \sum xy}{n \sum x^2 - (\sum x)^2}$$

Where:

- y' is the predicted value of the dependent variable
- a is the y-intercept of the regression line
- b is the slope of the regression line
- x is the independent variable
- n is the number of data points
- $\sum xy$ is the sum of the products of each x and y value
- $\sum x$ is the sum of the x values
- $\sum y$ is the sum of the y values
- $\sum x^2$ is the sum of the squares of the x values

Coefficient of Determination

The coefficient of determination, denoted as r^2, represents the proportion of the variance in the dependent variable that is predictable from the independent variable. It provides a measure of how well the regression line fits the data. A value of r^2 closer to 1 indicates that a greater proportion of variation in the dependent variable is explained by the model, while a value closer to 0 suggests that the model explains very little of the variation.

Where:

- r^2 is the coefficient of determination
- r is the correlation coefficient
- r^2 represents the proportion of variation in the dependent variable explained by the linear model
- The value of r^2 ranges from 0 to 1

Standard Error of the Estimate

The standard error of the estimate, denoted as S_{est}, measures the typical distance that the actual data points fall from the predicted values on the regression line. It indicates how well the regression model predicts the dependent variable. A smaller standard error means the model provides a better fit to the data.

$$S_{\text{est}} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

Where:

- S_{est} is the standard error of the estimate
- $\sum y^2$ is the sum of the squares of the observed y-values
- $\sum y$ is the sum of the observed y-values
- $\sum xy$ is the sum of the products of the corresponding x and y values
- a is the y-intercept of the regression line
- b is the slope of the regression line
- n is the number of data points

Authors

"10.7: Formulas for Chapter 10" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 10.7: Formulas for Chapter 10 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

10.8: Chapter 10 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Bivariate Data: Data involving two numerical variables that are analyzed together to explore possible relationships.

Coefficient of Determination (r^2): A value between 0 and 1 that indicates how much of the variation in the dependent variable is explained by the independent variable.

Correlation: A statistical measure that describes the strength and direction of the linear relationship between two variables.

Correlation Coefficient (r): A numerical value between –1 and 1 that quantifies the strength and direction of a linear relationship. Values near 1 indicate strong positive correlation, near –1 indicate strong negative correlation, and near 0 indicate no linear correlation.

Identifying Relationships: Bivariate analysis helps determine if two variables are related and the nature of that relationship (positive, negative, or none).

Intercept (a): In a linear regression equation, this is the value of the dependent variable when the independent variable is zero; the point where the regression line crosses the y-axis.

Linear Regression: A method used to model the relationship between a dependent variable and an independent variable by fitting a straight line to the data.

Predicting Outcomes: Regression models use one variable to estimate or predict the value of another, based on observed trends.

Regression Coefficients: The values in the regression equation (slope and intercept) that quantify how the independent variable affects the dependent variable.

Scatterplot: A graph that displays pairs of values as points on a coordinate plane, helping to visualize the relationship between two variables.

Slope (b): In a linear regression equation, the slope represents the rate of change; it tells how much the dependent variable changes for each unit increase in the independent variable.

Standard Error of the Estimate: A measure of the typical distance between the observed data points and the regression line; it reflects the accuracy of predictions.

Key Symbols

a - y-intercept of the regression line (value of y when $x = 0$).

b - slope of the regression line (rate of change in y with respect to x).

$\bar{x}$ - mean of the x-values in the sample.

$\bar{y}$ - mean of the y-values in the sample.

y' - predicted value of the dependent variable from the regression equation.

r - sample correlation coefficient (measures strength and direction of a linear relationship in sample data).

r^2 - coefficient of determination (proportion of the variance in y explained by x).

ρ - population correlation coefficient (measures strength and direction of a linear relationship in the population).

s_{EST} or SE - standard error of the estimate (typical deviation between observed and predicted values).

x - independent variable (predictor or explanatory variable).

y - dependent variable (response or outcome variable).

Authors

"10.8: Chapter 10 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 10.8: Chapter 10 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

11: CHI-SQUARE TEST

11.1: Goodness of Fit Test
11.2: Test for Independence
11.3: Chapter 11 Formulas
11.4: Chapter 11 - Key Terms and Symbols

This page titled 11: Chi-Square Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

11.1: Goodness of Fit Test

Learning Objectives

- Understand how to perform a chi-square goodness of fit test.
- Compare observed frequencies to expected frequencies based on a theoretical distribution.
- Determine whether a sample distribution matches an expected distribution.
- Use the test to assess how well data fits a claimed or known distribution.
- Interpret the results to evaluate the fit between observed data and expected outcomes.

The χ^2 **goodness-of-fit test** can be used to test the distribution of three or more proportions within a single population. The χ^2 goodness of fit test is used to determine whether observed categorical data match a theoretical or expected distribution based on known or claimed proportions.

Definition: χ^2 Goodness-of-Fit Test

The χ^2-test is a statistical test for testing the goodness-of-fit of a variable. It can be used when the data are obtained from a random sample and when the expected frequency (E) from each category is 5 or more.

The formula for the χ^2 -test statistic is:

$$\chi^2 = \sum \frac{(o-E)^2}{E} \tag{11.1.1}$$

Use a right-tailed χ^2 -distribution with $\mathrm{df} = k - 1$ where k = the number of categories.

with

- O = the observed frequency (what was observed in the sample) and
- E = the expected frequency (based on H_0 and the sample size).
- H_0 : The observed proportions match the expected proportions.
- H_1 : The observed proportions do not match the expected proportions.

Example 11.1.1 Goodness-of-Fit Test

An instructor claims that their students' grade distribution is different than the department's grade distribution. The department's grades have the following proportion of students who get A's is 35%, B's is 23%, C's is 25%, D's is 10%, and F's is 7%, in introductory statistics courses. For a sample of 250 introductory statistics students with this instructor, there were 80 A's, 50 B's, 58 C's, 38 D's, and 24 F's. Test the instructor's claim at the 5% level of significance.

Solution

1. State the claim and determine the null and alternative hypotheses.

This is a test for three or more proportions within a single population, so use the goodness-of-fit test. We will always use a right-tailed χ^2 test. The hypotheses for this example would be:

H_0 : The observed proportions match the expected proportions

H_1 : The observed proportions do not match the expected proportions (claim)

Even though there is an inequality in H_1, the goodness-of-fit test is always right-tailed. This is because we are testing to see if there is a large variation between the observed versus the expected values. If the variance between the observed and expected values is large, then there is a difference in the proportions.

2. Determine the critical value using the χ^2 table. The degrees of freedom are k – 1 (where k is the number of categories). Thus, d.f. = 5 – 1 = 4 and $\alpha = 0.05$. The critical value is 9.488.

   There are $k = 5$ categories that we are comparing: A's, B's, C's, D's, and F's.

   The observed counts are the actual number of A's, B's, C's, D's, and F's from the sample.

3. Compute the expected count for each of the five categories. Find the expected counts by multiplying the expected proportion of A's, B's, C's, D's, and F's by the sample size (which is 250). Round the final answer to three decimal places.

   It will be helpful to make a table to organize the work using the table below.

Observed and Expected Data Values

Grade	A's	B's	C's	D's	F's
Observed (O)	80	50	58	38	24
Expected (E)	0.35 · 250 = 87.5	0.23 · 250 = 57.5	0.25 · 250 = 62.5	0.10 · 250 = 25	0.07 · 250 = 17.5
$\frac{(O-E)^2}{E}$	$\frac{(80-87.5)^2}{87.5}$ = 0.6429	$\frac{(50-57.5)^2}{57.5}$ = 0.9738	$\frac{(58-62.5)^2}{62.5}$ = 0.324	$\frac{(38-25)^2}{25}$ = 6.76	$\frac{(24-17.5)^2}{17.5}$ = 2.4143

Table 11.1.1: Calculation of the Chi-Square Test Statistic

- The first row is grades.
- The second row is the observed values.
- The third row is the expected value for each grade.
- In the last row for each cell, subtract the observed value from the expected value. Then, square each difference and divide each square by the expected value.

The test statistic is the sum of this last row: $\chi^2 = \sum \frac{(O-E)^2}{E} = 0.6429 + 0.9783 + 0.324 + 6.76 + 2.4143 = 11.120$

4. Decide whether to reject or not reject the null hypothesis.

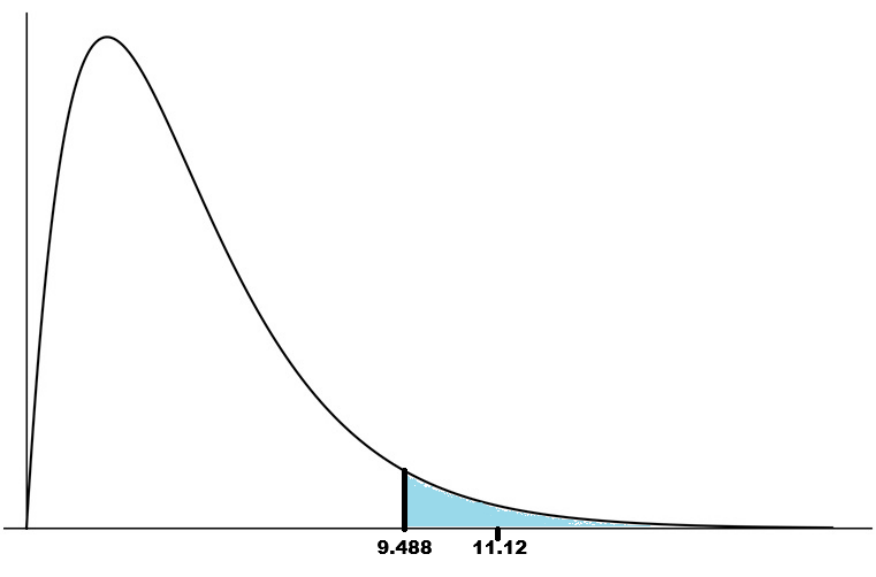


Figure 11.1.1: Reject the Null Hypothesis as the Test Point Falls in the Critical Region

The test statistic of $\chi^2 = 11.12 > 9.4877$ is in the rejection area, so the decision is to reject H_0.

5. Summarize the results.

There is sufficient evidence to support the claim that the proportion of students who get A's, B's, C's, D's and F's in introductory statistics courses for this instructor is different than the department's proportions of 35%, 23%, 25%, 10% and 7% respectively.

TI-84+: The test point can be computed using a built-in function in the calculator.

1. Enter the data by pressing **[STAT]** and then choosing **1:Edit**.
2. In **L1**, enter the **observed values,** and in **L2**, enter the **expected values**

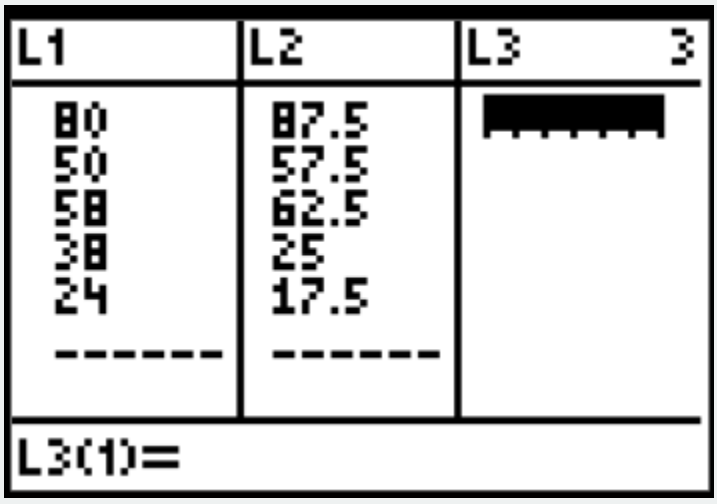


Figure 11.1.2: Enter the Observed Values and Expected Values

3. Press **[STAT]**, then scroll right to **TESTS**. Select **D:χ^2GOF-Test** (scroll down if needed).



Figure 11.1.3: Select the χ^2 Goodness of Fit Test

3. In the prompt, make sure **Observed: L_1, Expected: L_2,** and **df: 4** (this is degrees of freedom). Also, press **[Calculate]**.

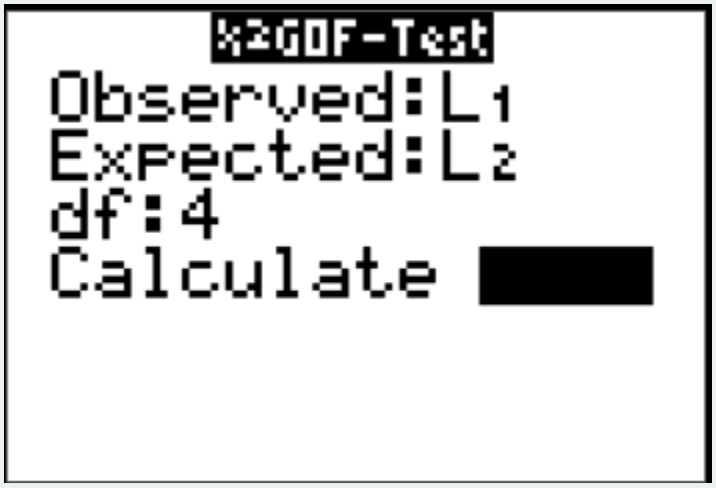


Figure 11.1.4: Make Sure Proper Lists are Selected Along With Degrees of Freedom

4. The test point is the value χ^2 = 11.11940373, which rounds to 11.12.

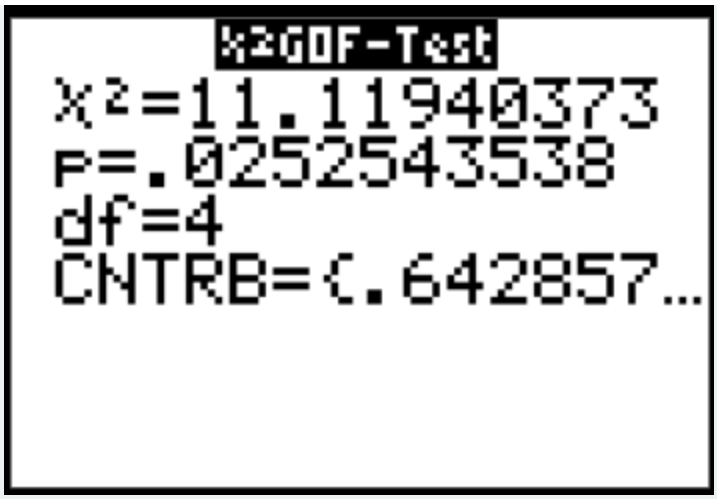


Figure 11.1.5: The Output Displays the χ^2 Test Point

Example 11.1.2

Suppose you have a die that you are curious about whether it is fair or not. If it is fair, then the proportion for each value should be the same. You need to find the observed frequencies, and to accomplish this, you roll the die 500 times and count how often each side comes up. Do the data show that the die is fair? Test at the 10% level.

Table 11.1.1: Observed Frequencies of Die

Die values	1	2	3	4	5	6	Total
Observed Frequency	78	87	87	76	85	87	100

Solution

1. State the null and alternative hypotheses and the level of significance

H_0: The observed frequencies are consistent with the distribution for a fair die (the die is fair)

H_1: The observed frequencies are not consistent with the distribution for a fair die (the die is not fair)

α = 0.10

2. Determine the critical value using the χ^2 table. The degrees of freedom are k – 1 (where k is the number of categories). Thus, d.f. = 6 – 1 = 5 and $\alpha = 0.10$. The critical value is 9.236.

3. Find the test statistic and p-value

First, you need to find the probability of rolling each side of the die. The sample space for rolling a die is {1, 2, 3, 4, 5, 6}. Since you are assuming that the die is fair, then $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

Now you can find the expected frequency for each side of the die. Since all the probabilities are the same, each expected frequency is the same.

$$\text{Expected Frequency} = E = n \cdot p = 500 \cdot \frac{1}{6} \approx 83.33$$

Test Statistic:

It is easier to calculate the test statistic using a table.

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
78	83.33	-5.22	28.4089	0.340920437
87	83.33	3.67	13.4689	0.161633265
87	83.33	3.67	13.4689	0.161633265
76	83.33	-7.33	53.7289	0.644772591
85	83.33	1.67	2.7889	0.033468139
87	83.33	3.67	13.4689	0.161633265
Total		0.02		$\chi^2 \approx 1.504060962$

Table 11.1.2: Calculation of the Chi-Square Test Statistic

The test statistic is rounded to $\chi^2 \approx 1.504$

4. Decide whether to reject or not reject the null hypothesis.

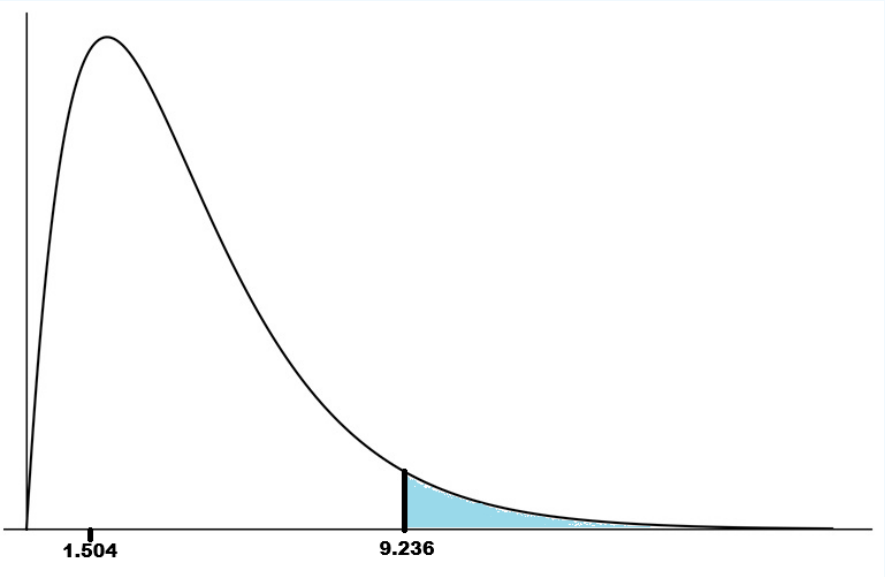


Figure 11.1.6: Do not reject the Null Hypothesis as the Test Point Falls in the Non-Critical Region

The result is do not reject the null hypothesis (H_0).

5. Summarize the results.

There is not enough evidence to show that the die is not consistent with the distribution for a fair die. There is not enough evidence to show that the die is not fair.

Authors

"11.1: Goodness of Fit Test" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"11.2: Chi-Square Goodness of Fit" by Kathryn Kozak is licensed under CC BY-SA 4.0

"10.2: Goodness of Fit Test" by Rachel Webb is licensed under CC BY-SA 4.0

This page titled 11.1: Goodness of Fit Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 10.2: Goodness of Fit Test by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.
- 11.2: Chi-Square Goodness of Fit by Kathryn Kozak is licensed CC BY-SA 4.0. Original source: https://s3-us-west-2.amazonaws.com/oerfiles/statsusingtech2.pdf.

11.2: Test for Independence

Learning Objectives

- Understand how to perform a chi-square test for independence.
- Evaluate whether two categorical variables are related or independent.
- Use a contingency table to compare observed and expected frequencies.
- Apply the test to analyze relationships in survey data or between traits in populations.
- Interpret the results to determine the presence or absence of an association between variables.

Use the **Chi-Square Test for Independence** to determine whether two **categorical (qualitative)** variables are associated. Qualitative data involves characteristics or categories (such as names or labels), and we analyze it by **counting how many individuals** fall into each category.

For example, suppose researchers are investigating whether there is a relationship between **breastfeeding duration** and the likelihood of being diagnosed with **autism spectrum disorder (ASD)**. They would collect data on how long each mother breastfed her child and whether the child was later diagnosed with ASD. This information would be organized into a **contingency table**, showing the frequency of individuals across the combined categories.

The test then evaluates whether the **distribution of counts** in each table cell differs from what we would expect if the two variables were truly **independent**. In this context, **independence** means that breastfeeding does **not affect** whether a child develops ASD.

The goal of the test is to see if there is **evidence of dependence**—that is, whether one variable **affects** the other. This is framed statistically as:

- **Null hypothesis (H_0):** The variables are independent.
- **Alternative hypothesis (H_1):** The variables are dependent.

To test this, we use the **Chi-Square Test for Independence**, which compares observed counts to expected counts under the assumption of independence.

Relation to Independent Probability Rule

For each χ^2 test for independence, data is collected, counted, and then organized into a contingency table. These values are known as the observed frequencies, and the symbol for an observed frequency is O. Total each row and column. The null hypothesis is that the two variables are independent. If two events are independent then $P(B) = P(B|A)$ and we can use the multiplication rule for independent events, to calculate the probability that variable A and B as the $P(A \text{ and } B) = P(A) \cdot P(B)$. Remember, in a hypothesis test, you assume that H_0 is true, and the two variables are assumed to be independent.

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B) \text{ if } A \text{ and } B \text{ are independent} \\ &= \frac{\text{Number of ways A can happen}}{\text{Total number of individuals}} \cdot \frac{\text{Number of ways B can happen}}{\text{Total number of individuals}} \\ &= \frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n} \end{aligned} \tag{11.2.1}$$

To compute the expected frequency for each cell in the contingency table, perform the following steps. Find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies. =

$$\begin{aligned} \text{Expected frequency (cell A and B)} &= E(A \text{ and } B) \\ &= n\left(\frac{\text{Row Total}}{n} \cdot \frac{\text{Column Total}}{n}\right) = \frac{\text{Row Total} \cdot \text{Column Total}}{n} \end{aligned} \tag{11.2.2}$$

If the variables are independent, the expected frequencies and the observed frequencies should be the same.

The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the "total difference" of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true.

How is the test statistic calculated? First, compute the differences between the observed and expected frequencies. Since some of the differences may be positive and others negative, they are squared to ensure all values are positive and to emphasize larger deviations. These squared differences might be large simply because the frequencies themselves are large, so each one is divided by its corresponding expected frequency to scale the values appropriately. Afterward, all of these scaled values are added together. This process measures the overall variance between observed and expected frequencies. The result is compared to a chi-square distribution to determine the critical value or p-value. For this reason, the procedure is known as a chi-square test.

Definition: χ^2 Test for Independence

The χ^2-test is a statistical test for testing the independence between two variables. It can be used when the data are obtained from a random sample, and when the expected value (E) from each cell is 5 or more.

- The test statistic is: $\chi^2 = \sum \frac{(O-E)^2}{E}$
- Use the χ^2 distribution table with degrees of freedom and α.
- The degrees of freedom are df = (the number of rows – 1) (the number of columns – 1), that is, $\mathrm{df} = (R-1)(C-1)$
- The observed frequency (sample results) is denoted as O
- The expected frequency is denoted as E

Steps to Perform the χ^2 Test for Independence

1. Identify the claim and state the hypotheses.
2. Find the critical value using the χ^2 distribution table with df = (the number of rows – 1) (the number of columns – 1) and α.
3. Compute the test statistic using $\chi^2 = \sum \frac{(O-E)^2}{E}$.
4. Decide whether to reject or not reject the null hypothesis (H_0).
5. Summarize the results.

Examples of χ^2 Test for Independence

Example 11.2.1

Is there an association between autism spectrum disorder (ASD) and breastfeeding? To explore this, a researcher surveyed mothers of children with and without ASD about the duration of time they breastfed their children. Does the data provide sufficient evidence to determine whether breastfeeding and ASD status are independent? Conduct the test using an $\alpha = 0.01$.

Observed Data Values

ASD	None	Less than 2 months	2 to 6 months	Over 6 months	Total
Yes	241	198	164	215	818
No	20	25	27	44	116
Total	261	223	191	259	934

Table 11.2.1: Length of Breast Feeding and ASD Status

(Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006.)

Solution

1. Identify the claim and state the hypotheses.

H_0 : Autism spectrum disorder and length of breastfeeding are independent (Claim)

H_1 : Autism spectrum disorder and length of breastfeeding are dependent

2. Find the critical value using the χ^2 distribution table .

   In this problem, there are 2 rows and 4 columns. Thus, df = (2 – 1)(4 – 1) = 1 · 3 = 3. Using df = 3 and α = 0.01, the critical value is 11.345.

3. Compute the test statistic.

   Compute the expected values for each cell using the observed value in the table above using $\text{Expected Value} = \frac{\text{Row Total} \cdot \text{Column Total}}{\text{Grand Total}}$.

Expected Data Values

ASD	None	Less Than 2 Months	2 to 6 Months	Over 6 Months	Total
Yes	$\frac{818 \cdot 261}{934}$ = 228.585	$\frac{818 \cdot 223}{934}$ = 195.304	= 167.278	$\frac{818 \cdot 259}{934}$ = 226.833	818
No	$\frac{116 \cdot 261}{934}$ = 32.415	$\frac{116 \cdot 223}{934}$ = 27.696	$\frac{116 \cdot 191}{934}$ = 23.722	$\frac{116 \cdot 259}{934}$ = 32.167	116
Total	261	223	191	259	934

Table 11.2.2: Computation of Expected Values Per Cell

It will be helpful to make a table for the expected counts and another one for each of the $\frac{(O-E)^2}{E}$ values to aid in computing the test statistic.

Chi-Square Values

ASD	None	Less Than 2 Months	2 to 6 Months	Over 6 Months	Total
Yes	$\frac{(241-228.585)^2}{228.585}$ = 0.6743	$\frac{(198-195.304)^2}{195.304}$ = 0.0372	$\frac{(164-167.278)^2}{167.278}$ = 0.0642	$\frac{(215-226.833)^2}{226.833}$ = 0.6173	818
No	$\frac{(20-32.415)^2}{32.415}$ = 4.7552	$\frac{(25-27.696)^2}{27.696}$ = 0.2624	$\frac{(27-23.722)^2}{23.722}$ = 0.4531	$\frac{(44-32.167)^2}{32.167}$ = 4.3529	116
Total	261	223	191	259	934

Table 11.2.3: Computation of $\frac{(O-E)^2}{E}$ Per Cell

The test statistic is the sum of all eight $\frac{(O-E)^2}{E}$ values: $\chi^2 = \sum \frac{(O-E)^2}{E} = 11.217$.

4. Decide whether to reject or not reject the null hypothesis (H_0).

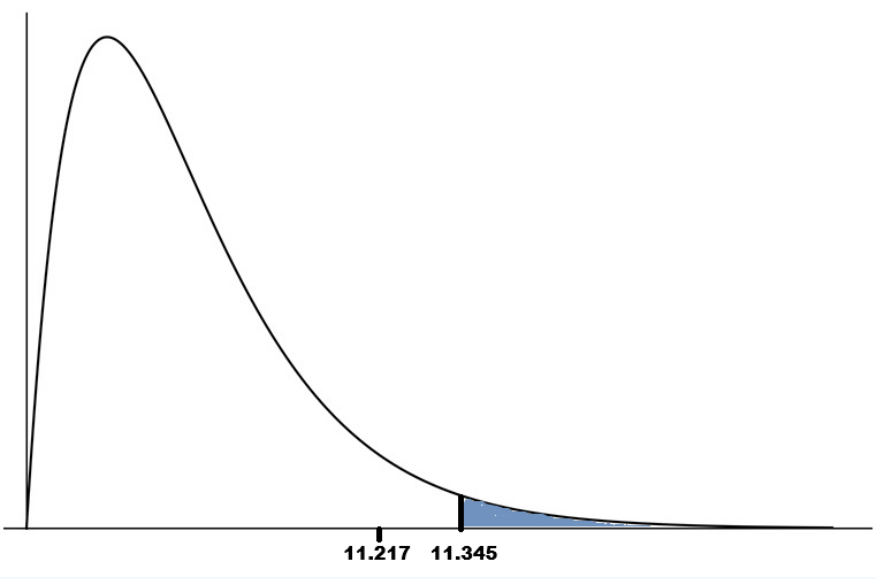


Figure 11.2.1: Test Point Falls in Noncritical Region

Since the test point falls into the non-critical region, the result is not to reject the null hypothesis (H_0).

5. Summarize the results.

   There is not enough evidence to show a relationship between autism spectrum disorder and breastfeeding.

Example 11.2.2

A community college researcher in southern California wants to know: Is there a dependent relationship between a student's gender and the type of school they plan to transfer to? To find out, the counselor surveys 800 students, recording their gender (male or female) and the type of four-year school they intend to transfer to: UC (University of California), Cal State, Out-of-State, or Private. Test the researcher's claim using the p-value method with an $\alpha = 0.05$. The observed values are provided in the contingency table below.

Observed Data Values

	UC	Cal State	Out - of - State	Private	Total
Male	116	168	63	47	394
Female	104	132	77	93	406
Total	220	300	140	140	800

Table 11.2.4: Student's Gender and Transfer School Type

Solution

1. Identify the claim and state the hypotheses.

   H_0: Gender and preferred transfer school are independent

   H_1: Gender and preferred transfer school are not independent (Claim)

2. Compute the test statistic and p-value using a TI-84+ calculator.
   a. Enter the Observed Data into a Matrix
      i. Press **[2nd]** then **[X^{-1}]** to open the **MATRIX** menu.

ii. Scroll right to **[EDIT]**, then select **[A]** by pressing **[1]**.
iii. Enter the matrix dimensions as **2 × 4** (2 rows for gender, 4 columns for school type).
iv. Enter the observed values row by row: Row 1 (Male): 116, 168, 63, 47, and Row 2 (Female): 104, 132, 77, 93.]



Figure 11.2.2: Select and Enter Values into a Matrix in TI-84+ Calculator

b. Perform the χ^2 test.
   i. Press **[STAT]** and scroll right to **TESTS**.
   ii. Scroll down and select **C:χ^2 Test.**
   iii. Set the following: **Observed Matrix: [A]** and **Expected Matrix: [B]** (the calculator will compute this).
   iv. Scroll to **Calculate** and press **[ENTER]**.

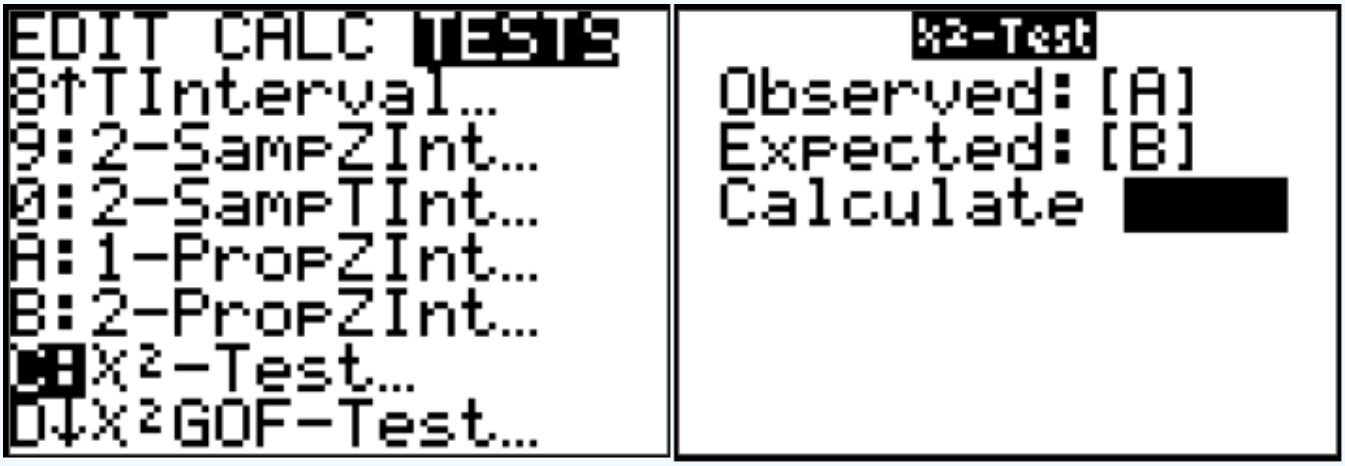


Figure 11.2.3: Select χ^2 Test in TI-84+ Calculator and Enter Matrices

c. Read the results.
   i. The test point is χ^2 = 21.314 (rounded to three place values)
   ii. The p-value = 9.061E-5, which is rewritten by moving the decimal places 5 places to the left to get 0.000091.

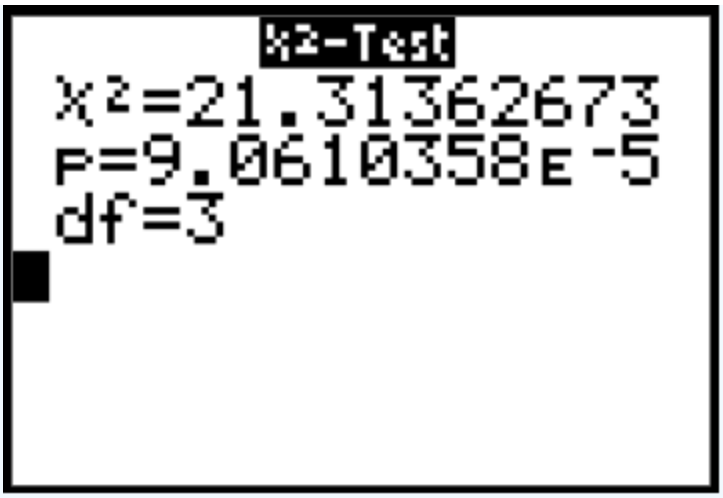


Figure 11.2.4: Output of χ^2 Test in TI-84+ Calculator

3. Decide whether to reject or not reject the null hypothesis (H_0).

   Since the p-value = 0.000091 < α = 0.05, the result is to reject the null hypothesis.

4. Summarize the results.

   There is enough evidence to support the claim that gender and preferred transfer school are related.

Example 11.2.3

A researcher is curious whether the size of a company influences whether it offers dental insurance to its employees. To explore this, data were collected from a sample of small, medium, and large companies, recording how many in each category provide dental coverage. Use an α = 0.10.

Based on the results, is there evidence of an independent relationship between company size and dental insurance coverage?

Use the data below to perform a chi-square test for independence and determine whether the type of company is associated with offering dental benefits.

Observed Data Values

Dental Insurance:	Small Company	Medium Company	Large Company	Total
Yes	21	25	19	65
No	46	39	10	95
Total	67	64	29	160

Table 11.2.5: Size of Company and Dental Insurance

Solution

1. Identify the claim and state the hypotheses.

H_0 : Dental insurance coverage and company size are independent (claim)

H_1 : Dental insurance coverage and company size are dependent

2. Find the critical value using the χ^2 distribution table .

In this problem, there are 2 rows and 3 columns. Thus, df = (2 – 1)(3 – 1) = 1 · 2 = 2. Using df = 2 and α = 0.10, the critical value is 4.605.

3. Compute the test statistic.

   Compute the expected values for each cell using the observed value in the table above using $\text{Expected Value} = \frac{\text{Row Total} \cdot \text{Column Total}}{\text{Grand Total}}$.

Expected Data Values

Dental Insurance:	Small Company	Medium Company	large Company	Total
Yes	$\frac{65 \cdot 67}{160}$ = 27.219	$\frac{65 \cdot 64}{160}$ = 26	$\frac{65 \cdot 29}{160}$ = 11.781	65
No	$\frac{95 \cdot 67}{160}$ = 39.781	$\frac{95 \cdot 64}{160}$ = 38	$\frac{95 \cdot 29}{160}$ = 17.219	95
Total	67	64	29	160

Table 11.2.6: Computation of Expected Values Per Cell

It will be helpful to make a table for the expected counts and another one for each of the $\frac{(O-E)^2}{E}$ values to aid in computing the test statistic.

Chi-Square Data Values

Dental Insurance:	Small Company	Medium Company	Large Company	Total
Yes	$\frac{(21-27.219)^2}{27.219}$ = 1.421	$\frac{(25-26)^2}{26}$ = 0.038	$\frac{(19-11.781)^2}{11.781}$ = 4.424	65
No	$\frac{(46-39.781)^2}{39.781}$ = 0.972	$\frac{(39-38)^2}{38}$ = 0.026	$\frac{(10-17.219)^2}{17.219}$ = 3.027	95
Total	67	64	29	160

Table 11.2.7: Computation of $\frac{(O-E)^2}{E}$ Values Per Cell

The test statistic is the sum of all six $\frac{(O-E)^2}{E}$ values: $\chi^2 = \sum \frac{(O-E)^2}{E} = 9.908$.

Remember that the TI-84+ calculator can be used to find both the test statistic and the p-value. The results are shown below. These values are more accurate than those from manual calculations because the calculator avoids rounding errors.

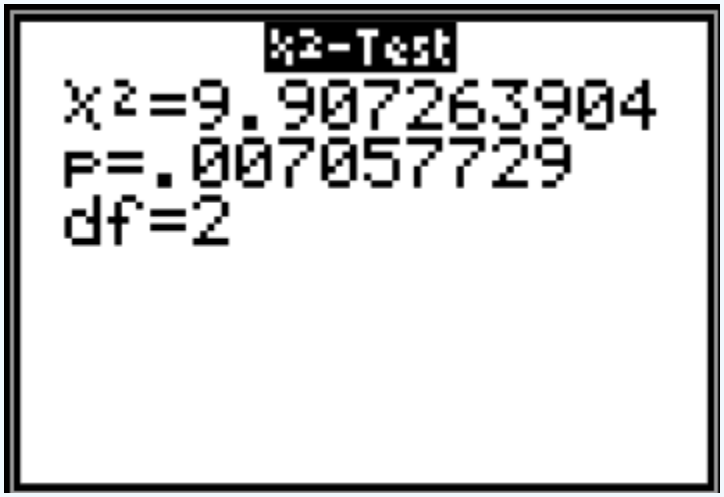


Figure 11.2.5: Output of χ^2 Test in TI-84+ Calculator

4. Decide whether to reject or not reject the null hypothesis (H_0).

- Traditional method: since the test point = 9.908 > Critical Value = 4.605, reject the null hypothesis (H_0).
- P-value Method: since the p-value = 0.0071 < α = 0.10, reject the null hypothesis (H_0).

5. Summarize the results.

There is enough evidence to support the claim that there is a relationship between dental insurance coverage and company size.

Authors

"11.2: Test for Independence" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY-SA 4.0

Attributions

"11.1: Chi-Square Test for Independence" by Kathryn Kozak is licensed under CC BY-SA 4.0

"10.3: Test for Independence" by Rachel Webb is licensed under CC BY-SA 4.0

This page titled 11.2: Test for Independence is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan via source content that was edited to the style and standards of the LibreTexts platform.

- 10.3: Test for Independence by Rachel Webb is licensed CC BY-SA 4.0. Original source: https://mostlyharmlessstat.wixsite.com/webpage.

11.3: Chapter 11 Formulas

Chapter 11 Formulas

χ^2 Test Statistic for Goodness of Fit

The χ^2 test of goodness of fit is used to determine whether the observed frequencies in a categorical variable match the expected frequencies based on a specific theoretical distribution. It assesses how well the observed data fit the expected pattern. The test uses degrees of freedom given by $df = k - 1$, where k is the number of categories.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- χ^2 is the chi-square test statistic
- O is the observed frequency in each category
- E is the expected frequency in each category
- $\sum$ indicates that the calculation is summed over all categories
- df is the degrees of freedom
- k is the number of distinct categories being tested

Expected Value for the χ^2 Test of Independence

The expected value in a χ^2 test for independence is the frequency that would be expected in each cell of a contingency table if the row and column variables were statistically independent. These expected values are used to compare with the observed values to compute the chi-square test statistic. This calculation is done for each cell in the table to evaluate how far the observed values deviate from what would be expected under the assumption of independence.

$$E = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

χ^2 Test Statistic for Independence

The chi-square test for independence is used to determine whether there is a significant association between two categorical variables in a contingency table. It compares the observed frequencies in each cell to the frequencies we would expect if the two variables were independent. The degrees of freedom for the test are calculated using the formula $df = (r - 1)(c - 1)$, where $r$is the number of rows and c is the number of columns in the table.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- χ^2 is the chi-square test statistic
- O is the observed frequency in a cell
- E is the expected frequency in the same cell
- $\sum$ indicates the sum over all cells in the contingency table
- $df = (r - 1)(c - 1)$ is the degrees of freedom
- r is the number of rows
- c is the number of columns

Authors

"11.3: Chapter 11 Formulas" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 11.3: Chapter 11 Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

11.4: Chapter 11 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

χ^2 **Distribution**: a right-skewed probability distribution used primarily in tests of categorical data, such as the goodness-of-fit test and the test for independence. It is defined by degrees of freedom and becomes more symmetric as the degrees of freedom increase.

Degrees of Freedom: the number of independent values that can vary in an analysis. For a goodness-of-fit test, it is calculated as the number of categories minus 1. For a test of independence, it is calculated as (number of rows – 1) × (number of columns – 1).

Expected Frequency: the theoretical count of observations in each category or cell, assuming the null hypothesis is true. It is used to compare with the observed frequency.

Goodness-of-Fit Test: a χ^2 test determines if a sample distribution matches an expected or theoretical distribution.

Independence (Test of): a χ^2test used to determine whether two categorical variables are independent or related based on data collected in a contingency table.

Observed Frequency: The actual count of observations collected in each category or cell from the sample data.

Test Statistic: A value computed from sample data that is used in hypothesis testing. Chi-square tests measure the discrepancy between observed and expected frequencies.

Key Symbols

χ^2 – chi-square test statistic.

O – observed frequency.

E – expected frequency.

df – degrees of freedom.

p – p-value.

α – significance level.

k – number of categories (for goodness-of-fit).

r – number of rows (in a contingency table).

c – number of columns (in a contingency table).

n – total sample size.

Authors

"11.4: Chapter 11 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 11.4: Chapter 11 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

12: ANALYSIS OF VARIANCE (ANOVA)

12.1: Analysis of Variance (ANOVA)
12.2: Chapter 12 Formulas
12.3: Chapter 12 - Key Terms and Symbols

This page titled 12: Analysis of Variance (ANOVA) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

12.1: Analysis of Variance (ANOVA)

Learning Objectives

- Define ANOVA (Analysis of Variance) and explain its purpose in statistical analysis.
- Use ANOVA to compare the means of three or more groups.
- Interpret ANOVA results to determine whether at least one group mean differs significantly from the others.
- Distinguish between variation between groups and within groups in ANOVA.
- Identify appropriate situations for using ANOVA, such as experiments, surveys, or studies with multiple categories or treatments.

Introduction

For this lesson, we will compare the differences between three or more population means using the F-test. The hypotheses testing will be the following:

$$H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

H_1 : At least one of the means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal

With an F-test, we can compare all the means at the same time. The F-test value is a ratio between two different variances. If the variances are similar, the F-value should be close to 1. If the variances are different, then the F-value will not equal one but it will be greater or lower in value.

$$F = \frac{s_1^2}{s_2^2}$$

For this section, we will compare the variance between the three groups with the variances within the three groups.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

In this case, for the groups to show evidence of different means, the variance between groups should be greater than the variance within groups. Below is an illustration, to help understand this concept.

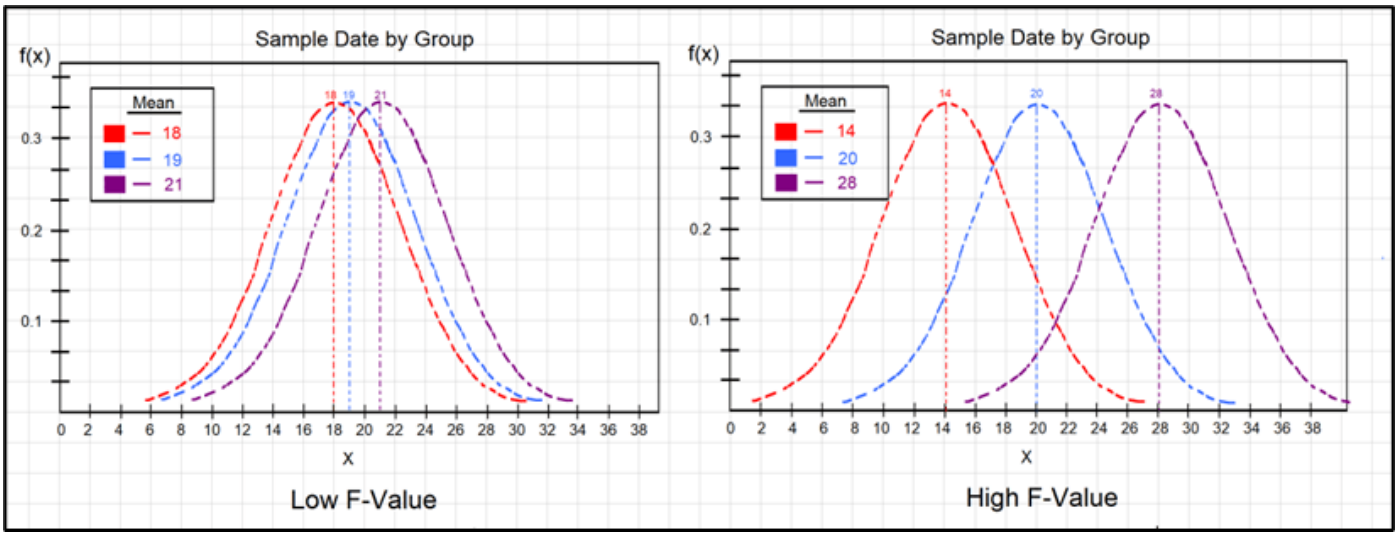


Figure 12.1.1: Low F-Value and High F-Value Comparison

If there is enough evidence to show a significant difference between these variances, then we can say we have enough statistical evidence to conclude there is a difference between the population means based on a certain α-level. The method we discussed above is called analysis of variance or ANOVA.

Assumptions Needed to Perform an ANOVA Test:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).
4. From each population, the samples must be simple random samples.

ANOVA follows the same steps for hypothesis testing (P-value method).

1. State the hypotheses H_0 and H_1.
2. Find the p-value.
3. Compare the p-value with α.
4. Determine if there is enough statistical evidence to reject H_o (or support H_1).
5. Summarize the result.

Application of Analysis of Variance (ANOVA) Using P-Value Method

Now let us look at an example to better understand the ANOVA method and how we can use it to compare the difference between three population means and make a conclusion based on statistical evidence.

Example 12.1.1 ANOVA Test on Weights of Men in Different Cities

Groups of men from three different cities in the United States are to be tested for population mean weight. The entries in the table below are the weights for the different groups. Is there enough evidence to conclude the difference between the mean weights of the three groups with $\alpha = 0.05$?

Data Values

Los Angeles	Austin	New York
216	202	170

Los Angeles	Austin	New York
198	213	165
222	244	182
184	228	197
191	210	201

Table 12.1.1: Three Samples of Weights of Men from the Cities of Los Angeles, Austin, and New York.

Attribution: "Introductory Statistics" by Barbara Illowsky and Susan Dean, OpenStax is licensed under CC BY 4.0

Solution

1. Please state the hypotheses.

   $H_o : \mu_1 = \mu_2 = \mu_3$

   H_1 : At least one of the means is not equal

2. Find the P-value from our TI-84 graphing calculator.
   a. Press the **[STAT]** button.
   b. On the screen, highlight **[EDIT]**, select 1: Edit..., and press the **[ENTER]** button.
   c. Put your data from each group (Los Angeles, Austin, and New York) in columns **L_1**, **L_2**, and **L_3**. If everything is done correctly, it should look like the image below:

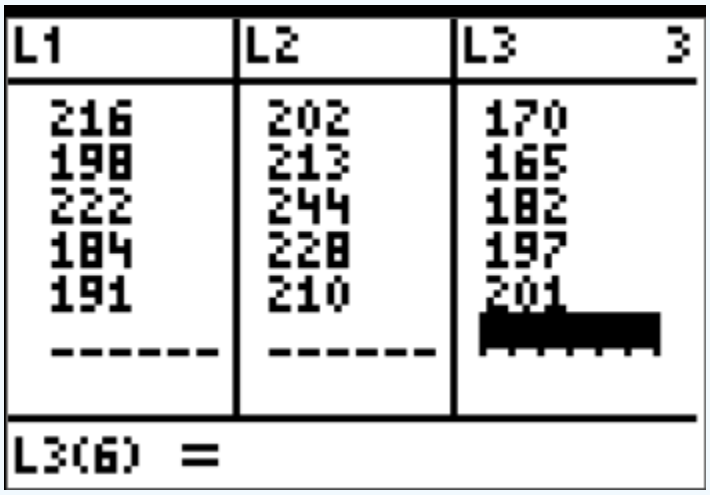


Figure 12.1.1: Men's weight from Los Angeles is in L1, Austin is in L2, and New York is in L3.

   d. Press the **[STAT]** button again.
   e. On the screen, highlight **[TESTS]**, scroll down, and select **H:ANOVA(** . Then, press the **[ENTER]** button.
   f. It should put ANOVA(on the screen.
   g. To enter **L_1** on the screen, press **[2ND]**, then press the 1 button.
   h. Press **[,]** (comma) button. It is above the 7 button.
   i. To enter **L_2** on the screen, press **[2ND]** then press the 2 button.
   j. Press **[,]** (comma) button.
   k. To enter **L_3** on the screen, press **[2ND]** then press the 3 button. Finally, press the right parenthesis **[ )]** above the 9 button.
   l. If the instructions in Step 2 are done correctly, then the screen should look like the following:



Figure 12.1.2: Calculator Screen Shows ANOVA(L1, L2, L3)

l. For the last step, press the ENTER button. If everything is done correctly, then your screen should show the following image on the next page.

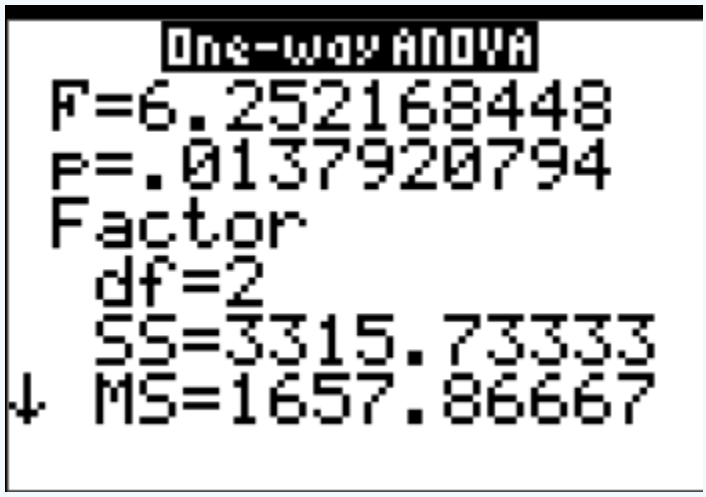


Figure 12.1.3: Calculator Screen Displays the P-value for ANOVA Test

m. The screen shows that the P-value = 0.0137920794

3. Compare the p-value with α and determine if we have enough evidence to reject H_o (or support H_1 if that is the claim). If not given, then we assume it is equal to 0.05.

- If your p-value is less than α, we reject H_o (or support H_1)
- If your p-value is more than α, we do not reject H_o (or not support H_1)

Since p-value = 0.0138 < α = 0.05, the result is to reject H_0.

4. Determine if there is enough statistical evidence to reject H_o (or support H_1).

   Since the claim is H_1, write the summary in terms of H_1. Since the p-value is less than our α, the result of the test supports H_1.

5. Summarize the results.

   There is enough statistical evidence to support the claim that there is a difference between the mean weight of men in the three cities based on an α-level of 0.05.

Hypothesis Testing for ANOVA Using Traditional Method

1. State the hypotheses H_o and H_1.
2. Determine critical value in F-distribution table using α, degrees of freedom of the numerator = k – 1 (k = number of classes/groups), and degrees of freedom of the denominator = N – k (N = total number of data values)
3. Compute the test point.
4. Determine if there is enough statistical evidence to reject H_o (or support H_1).
5. Summarize the result.

Example 12.1.2

A nutritionist conducted a 12-week study to compare the effectiveness of three popular weight loss diets: Keto, Intermittent Fasting, and Mediterranean. Participants were randomly assigned to one of the three diets, and their weight loss (in pounds) at the end of the study was recorded. The goal is to determine whether there is a statistically significant difference in average weight loss among the three diets using a one-way ANOVA test with a significance level of α = 0.05.

Data Values

Keto	Intermittent Fasting	Mediterranean
12	8	6
15	11	5
14	9	7
10	10	6
13	7	4

Table 12.1.3: Three Samples of Pounds Loss for People on Three Diets: Keto, Intermittent Fasting, and Mediterranean.

1. Please state the hypotheses.

   $H_o : \mu_1 = \mu_2 = \mu_3$

   H_1 : At least one of the means is not equal (Claim)

2. Determine the critical value.

- k = 3, as there are 3 different diets.
- N =15, as there are 15 given data values.
- Degrees of Freedom of the numerator k - 1 = 3 - 1 = 2
- Degrees of freedom of the denominator = N - k = 15 - 3= 12
- Look up the critical value in the F-distribution with α = 0.05.
- The value is 3.89.

3. Use a graphing calculator or other technology to compute the test point. Below is the output of the TI-84+ found using the same steps as example 1 above.

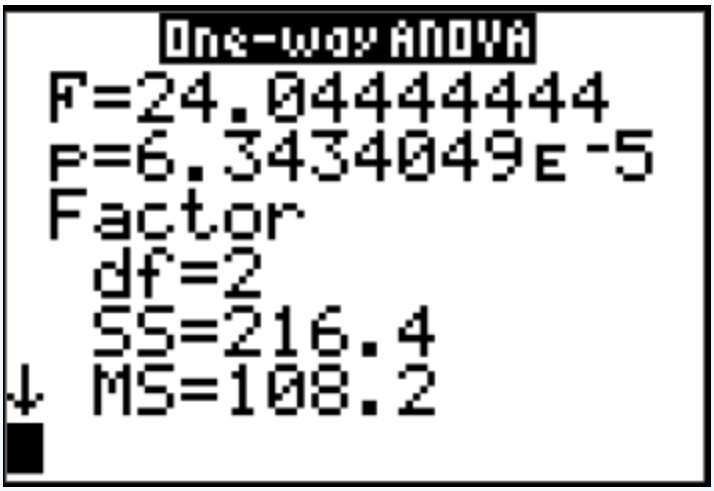


Figure 12.1.4: Calculator Screen Displays the Test Point for ANOVA Test

The test point is F = 24.04, rounded to two decimal places.

4. Decide whether to reject H_0 or not.

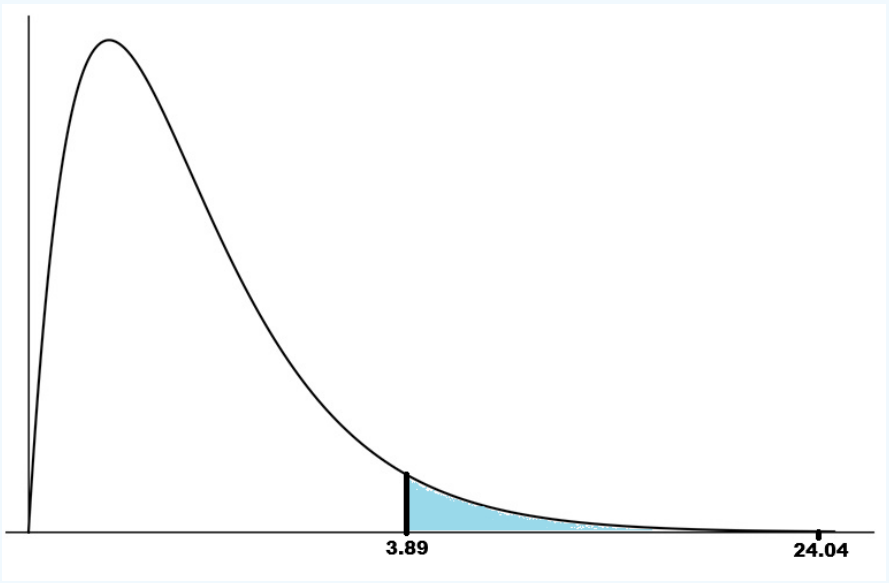


Figure 12.1.5: Reject the Null Hypothesis as the Test Point Falls in the Critical Region

Since the test point F = 24.04 > critical value = 3.89, the result is to reject H_0.

5. Summarize the results.

There is enough evidence to support the claim that there is significant weight loss among the three diets.

Exercise 12.1.1

Citrus College wants to analyze whether the mean GPA of students differs across three majors: Art, STEM, and Nursing. A random sample of 10 students from each major was selected, and their GPAs were recorded. Conduct a one-way ANOVA test at a 5% significance level to determine if there is a statistically significant difference in the mean GPA among the three majors.

Data Values

Art	STEM	Nursing
2.80	3.40	3.50
3.00	3.60	3.70
2.90	3.80	3.60
3.10	3.50	3.80
3.20	3.70	3.40
2.70	3.40	3.50
3.00	3.60	3.90
3.30	3.80	3.70
2.90	3.50	3.60
3.10	3.60	3.80

Answer available at the end of this section.

Authors

"12.1: Analysis of Variance (ANOVA)" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

Answers

Exercise 12.1.1

This page titled 12.1: Analysis of Variance (ANOVA) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

12.2: Chapter 12 Formulas

Chapter 12 Formulas

Between-Group Variance

The between-group variance in ANOVA, denoted as S_B^2, measures the variability of group means around the overall (grand) mean. It reflects how much the means of different groups differ from one another. A larger between-group variance suggests that the group means are more spread out, which may indicate a significant effect of the independent variable.

$$S_B^2 = \frac{\sum n_i(\bar{x}_i - \bar{x}_{\text{GM}})^2}{k-1}$$

Where:

- S_B^2 is the between-group variance
- n_i is the number of observations in group i
- $\bar{x}_i$ is the mean of group i
- $\bar{x}_{\text{GM}}$ is the grand mean (mean of all data points)
- k is the number of groups

Within-Group Variance

The within-group variance in ANOVA, denoted as S_W^2, estimates the average variability within each group by pooling the individual group variances, weighted by their degrees of freedom. This form emphasizes that the total degrees of freedom within groups is the sum of $n_i - 1$ across all groups.

$$S_W^2 = \frac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)}$$

Where:

- S_W^2 is the within-group variance
- s_i^2 is the variance of group i
- n_i is the number of observations in group i
- $\sum(n_i - 1)s_i^2$ is the weighted sum of variances from each group
- $\sum(n_i - 1)$ is the total degrees of freedom within groups

Test Statistic For ANOVA

The test statistic for ANOVA is used to determine whether there are statistically significant differences between the means of three or more independent groups. It is calculated as the ratio of the between-group variance to the within-group variance. A larger test statistic value indicates greater variation between the group means relative to the variation within the groups, which may suggest that at least one group mean is different from the others.

$$F = \frac{S_B^2}{S_W^2}$$

Where:

- F is the ANOVA test statistic
- S_B^2 is the between-group variance
- S_W^2 is the within-group variance

Authors

"12.2: Chapter 12 Formulas" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 12.2: Chapter 12 Formulas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

12.3: Chapter 12 - Key Terms and Symbols

Glossary of Key Terms and Symbols

Key Terms

Alternative Hypothesis: The statement that at least one group mean is different from the others. It contrasts with the null hypothesis in ANOVA.

Between-Group Variance: A measure of variability between the means of different groups. It reflects how much the group means differs from the grand mean.

Degrees of Freedom: The number of independent values that can vary in an analysis. ANOVA has degrees of freedom for between-group, within-group, and total variation.

Dependent Variable: The variable being measured or tested in the experiment. It is affected by changes in the independent variable.

F-Distribution: A probability distribution used in ANOVA to compare variances. It is skewed right and used to determine critical values in F-tests.

F-Ratio: The test statistic in ANOVA, calculated by dividing the between-group variance by the within-group variance.

F-Test: A statistical test that compares the variances of two or more groups to determine if at least one group mean is different.

Grand Mean: The overall mean of all the data points from all groups combined.

Independent Variable: The variable that is manipulated or categorized to observe its effect on the dependent variable. It defines the groups being compared.

Level of Significance: The probability threshold (commonly denoted as alpha, such as 0.05) used to decide whether to reject the null hypothesis.

Mean Square: The variance estimate in ANOVA, calculated by dividing the sum of squares by the corresponding degrees of freedom. Includes between-group and within-group mean squares.

Null Hypothesis: The statement that all group means are equal. It is tested against the alternative hypothesis in ANOVA.

One-Way ANOVA: An analysis of variance method used to compare means across one independent variable with two or more levels or groups.

Within-Group Variance: A measure of variability within each group. It reflects differences among individual scores in the same group.

Key Symbols

$\bar{x}$ – sample mean of a group

$\bar{x}_{GM}$ – grand mean (mean of all data across all groups)

k – number of groups or levels of the independent variable

n – number of observations per group

N – total number of observations across all groups

S_B^2 – Between-group variance.

S_W^2 – within group variance

dfn – degrees of freedom of the numerator $(k-1)$

dfd – degrees of freedom of the denominator $(N-k)$

F – F-ratio or test statistic (S_B^2/S_W^2)

α – significance level (probability of Type I error)

p – p-value.

Authors

"12.3: Chapter 12 - Key Terms and Symbols" by Toros Berberyan, Tracy Nguyen, and Alfie Swan is licensed under CC BY 4.0

This page titled 12.3: Chapter 12 - Key Terms and Symbols is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13: STATISTICAL TABLES

13.1: The Binomial Distribution Table for n = 2 to n = 10
13.2: The Binomial Distribution Table for n = 11 to n = 15
13.3: The Binomial Distribtution Table for n = 16 to n = 18
13.4: The Binomial Distribution Table for n = 19 to n = 20
13.5: The Standard Normal Distriubtion Table for Negative Z-Values
13.6: The Standard Normal Distribution Table for Positive Z-Values
13.7: The t-Distribution Table for Degrees of Freedom = n - 1
13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2
13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1
13.10: The F-Distribution for Alpha = 0.005
13.11: The F-Distribution Table for Alpha = 0.01
13.12: The F-Distribution Table for Alpha = 0.025
13.13: The F-Distribution Table for Alpha = 0.05
13.14: The F-Distribution Table for Alpha = 0.10

This page titled 13: Statistical Tables is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.1: The Binomial Distribution Table for n = 2 to n = 10

The Binomial Distribution Table for n = 2 to n = 10

	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8
	0.640	0.490	0.360	0.250	0.160	0.090	0.040
	0.320	0.420	0.480	0.500	0.480	0.420	0.320
	0.040	0.090	0.160	0.250	0.360	0.490	0.640
	0.512	0.343	0.216	0.125	0.064	0.027	0.008
	0.384	0.441	0.432	0.375	0.288	0.189	0.096
	0.096	0.189	0.288	0.375	0.432	0.441	0.384
	0.008	0.027	0.064	0.125	0.216	0.343	0.512
	0.410	0.240	0.130	0.063	0.026	0.008	0.002
	0.410	0.412	0.346	0.250	0.154	0.076	0.026
	0.154	0.265	0.346	0.375	0.346	0.265	0.154
	0.026	0.076	0.154	0.250	0.346	0.412	0.410
	0.002	0.008	0.026	0.063	0.130	0.240	0.410
	0.328	0.168	0.078	0.031	0.010	0.002	0.000
	0.410	0.360	0.259	0.156	0.077	0.028	0.006
	0.205	0.309	0.346	0.313	0.230	0.132	0.051
	0.051	0.132	0.230	0.313	0.346	0.309	0.205
	0.006	0.028	0.077	0.156	0.259	0.360	0.410
	0.000	0.002	0.010	0.031	0.078	0.168	0.328
	0.262	0.118	0.047	0.016	0.004	0.001	0.000
	0.393	0.303	0.187	0.094	0.037	0.010	0.002
	0.246	0.324	0.311	0.234	0.138	0.060	0.015
	0.082	0.185	0.276	0.313	0.276	0.185	0.082
	0.015	0.060	0.138	0.234	0.311	0.324	0.246
	0.002	0.010	0.037	0.094	0.187	0.303	0.393
	0.000	0.001	0.004	0.016	0.047	0.118	0.262
	0.210	0.082	0.028	0.008	0.002	0.000	0.000
	0.367	0.247	0.131	0.055	0.017	0.004	0.000
	0.275	0.318	0.261	0.164	0.077	0.025	0.004
	0.115	0.227	0.290	0.273	0.194	0.097	0.029
	0.029	0.097	0.194	0.273	0.290	0.227	0.115
	0.004	0.025	0.077	0.164	0.261	0.318	0.275
	0.000	0.004	0.017	0.055	0.131	0.247	0.367
	0.000	0.000	0.002	0.008	0.028	0.082	0.210
	0.168	0.058	0.017	0.004	0.001	0.000	0.000
	0.336	0.198	0.090	0.031	0.008	0.001	0.000

	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8
	0.294	0.296	0.209	0.109	0.041	0.010	0.001
	0.147	0.254	0.279	0.219	0.124	0.047	0.009
	0.046	0.136	0.232	0.273	0.232	0.136	0.046
	0.009	0.047	0.124	0.219	0.279	0.254	0.147
	0.001	0.010	0.041	0.109	0.209	0.296	0.294
	0.000	0.001	0.008	0.031	0.090	0.198	0.336
	0.000	0.000	0.001	0.004	0.017	0.058	0.168
	0.134	0.040	0.010	0.002	0.000	0.000	0.000
	0.302	0.156	0.060	0.018	0.004	0.000	0.000
	0.302	0.267	0.161	0.070	0.021	0.004	0.000
	0.176	0.267	0.251	0.164	0.074	0.021	0.003
	0.066	0.172	0.251	0.246	0.167	0.074	0.017
	0.017	0.074	0.167	0.246	0.251	0.172	0.066
	0.003	0.021	0.074	0.164	0.251	0.267	0.176
	0.000	0.004	0.021	0.070	0.161	0.267	0.302
	0.000	0.000	0.004	0.018	0.060	0.156	0.302
	0.000	0.000	0.000	0.002	0.010	0.040	0.134
	0.107	0.028	0.006	0.001	0.000	0.000	0.000
	0.268	0.121	0.040	0.010	0.002	0.000	0.000
	0.302	0.233	0.121	0.044	0.011	0.001	0.000
	0.201	0.267	0.215	0.117	0.042	0.009	0.001
	0.088	0.200	0.251	0.205	0.111	0.037	0.006
	0.026	0.103	0.201	0.246	0.201	0.103	0.026
	0.006	0.037	0.111	0.205	0.251	0.200	0.088
	0.001	0.009	0.042	0.117	0.215	0.267	0.201
	0.000	0.001	0.011	0.044	0.121	0.233	0.302
	0.000	0.000	0.002	0.010	0.040	0.121	0.268
	0.000	0.000	0.000	0.001	0.006	0.028	0.107

Table 13.1.1: Binomial Probabilitiy Table for n =2 to n = 10

Author

"13.1: The Binomial Distribution Table for n = 2 to n = 10" by Alfie Swan is in the Public Domain, CC0

This page titled 13.1: The Binomial Distribution Table for n = 2 to n = 10 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.2: The Binomial Distribution Table for n = 11 to n = 15

The Binomial Distribution Table for n = 11 to n = 15

	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8
	0.314	0.086	0.020	0.004	0.000	0.000	0.000	0.000
	0.384	0.236	0.093	0.027	0.005	0.001	0.000	0.000
	0.213	0.295	0.200	0.089	0.027	0.005	0.001	0.000
	0.071	0.221	0.257	0.177	0.081	0.023	0.004	0.000
	0.016	0.111	0.220	0.236	0.161	0.070	0.017	0.002
	0.002	0.039	0.132	0.221	0.226	0.147	0.057	0.016
	0.000	0.010	0.057	0.147	0.226	0.221	0.132	0.039
	0.000	0.002	0.017	0.070	0.161	0.236	0.220	0.111
	0.000	0.000	0.004	0.023	0.081	0.177	0.257	0.221
	0.000	0.000	0.001	0.005	0.027	0.089	0.200	0.295
	0.000	0.000	0.000	0.001	0.005	0.027	0.093	0.236
	0.000	0.000	0.000	0.000	0.000	0.004	0.020	0.086
	0.282	0.069	0.014	0.002	0.000	0.000	0.000	0.000
	0.377	0.206	0.071	0.017	0.003	0.000	0.000	0.000
	0.230	0.283	0.168	0.064	0.016	0.002	0.000	0.000
	0.085	0.236	0.240	0.142	0.054	0.012	0.001	0.000
	0.021	0.133	0.231	0.213	0.121	0.042	0.008	0.001
	0.004	0.053	0.158	0.227	0.193	0.101	0.029	0.003
	0.000	0.016	0.079	0.177	0.226	0.177	0.079	0.016
	0.000	0.003	0.029	0.101	0.193	0.227	0.158	0.053
	0.000	0.001	0.008	0.042	0.121	0.213	0.231	0.133
	0.000	0.000	0.001	0.012	0.054	0.142	0.240	0.236
	0.000	0.000	0.000	0.002	0.016	0.064	0.168	0.283
	0.000	0.000	0.000	0.000	0.003	0.017	0.071	0.206
	0.000	0.000	0.000	0.000	0.000	0.002	0.014	0.069
	0.254	0.055	0.010	0.001	0.000	0.000	0.000	0.000
	0.367	0.179	0.054	0.011	0.002	0.000	0.000	0.000
	0.245	0.268	0.139	0.045	0.010	0.001	0.000	0.000
	0.100	0.246	0.218	0.111	0.035	0.006	0.001	0.000
	0.028	0.154	0.234	0.184	0.087	0.024	0.003	0.000
	0.006	0.069	0.180	0.221	0.157	0.066	0.014	0.001
	0.001	0.023	0.103	0.197	0.209	0.131	0.044	0.006
	0.000	0.006	0.044	0.131	0.209	0.197	0.103	0.023
	0.000	0.001	0.014	0.066	0.157	0.221	0.180	0.069
	0.000	0.000	0.003	0.024	0.087	0.184	0.234	0.154

	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8
	0.000	0.000	0.001	0.006	0.035	0.111	0.218	0.24
	0.000	0.000	0.000	0.001	0.010	0.045	0.139	0.26
	0.000	0.000	0.000	0.000	0.002	0.011	0.054	0.17
	0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.05
	0.229	0.044	0.007	0.001	0.000	0.000	0.000	0.00
	0.356	0.154	0.041	0.007	0.001	0.000	0.000	0.00
	0.257	0.250	0.113	0.032	0.006	0.001	0.000	0.00
	0.114	0.250	0.194	0.085	0.022	0.003	0.000	0.00
	0.035	0.172	0.229	0.155	0.061	0.014	0.001	0.00
	0.008	0.086	0.196	0.207	0.122	0.041	0.007	0.00
	0.001	0.032	0.126	0.207	0.183	0.092	0.023	0.00
	0.000	0.009	0.062	0.157	0.209	0.157	0.062	0.00
	0.000	0.002	0.023	0.092	0.183	0.207	0.126	0.03
	0.000	0.000	0.007	0.041	0.122	0.207	0.196	0.08
	0.000	0.000	0.001	0.014	0.061	0.155	0.229	0.17
	0.000	0.000	0.000	0.003	0.022	0.085	0.194	0.25
	0.000	0.000	0.000	0.001	0.006	0.032	0.113	0.25
	0.000	0.000	0.000	0.000	0.001	0.007	0.041	0.15
	0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.04
	0.206	0.035	0.005	0.000	0.000	0.000	0.000	0.00
	0.343	0.132	0.031	0.005	0.000	0.000	0.000	0.00
	0.267	0.231	0.092	0.022	0.003	0.000	0.000	0.00
	0.129	0.250	0.170	0.063	0.014	0.002	0.000	0.00
	0.043	0.188	0.219	0.127	0.042	0.007	0.001	0.00
	0.010	0.103	0.206	0.186	0.092	0.024	0.003	0.00
	0.002	0.043	0.147	0.207	0.153	0.061	0.012	0.00
	0.000	0.014	0.081	0.177	0.196	0.118	0.035	0.00
	0.000	0.003	0.035	0.118	0.196	0.177	0.081	0.01
	0.000	0.001	0.012	0.061	0.153	0.207	0.147	0.04
	0.000	0.000	0.003	0.024	0.092	0.186	0.206	0.10
	0.000	0.000	0.001	0.007	0.042	0.127	0.219	0.18
	0.000	0.000	0.000	0.002	0.014	0.063	0.170	0.25
	0.000	0.000	0.000	0.000	0.003	0.022	0.092	0.23
	0.000	0.000	0.000	0.000	0.000	0.005	0.031	0.13
	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.03

Table 13.2.1: The Binomial Distribution Table for n = 10 to n = 15

Author

"13.2: The Binomial Distribution Table for n = 11 to n = 15" by Alfie Swan is in the Public Domain, CC0

This page titled 13.2: The Binomial Distribution Table for n = 11 to n = 15 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.3: The Binomial Distribtution Table for n = 16 to n = 18

The Binomial Distribution Table for n = 16 to n = 18

p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	
0.440	0.185	0.028	0.003	0.000	0.000	0.000	0.000	0.000	
0.371	0.329	0.113	0.023	0.003	0.000	0.000	0.000	0.000	
0.146	0.275	0.211	0.073	0.015	0.002	0.000	0.000	0.000	
0.036	0.142	0.246	0.146	0.047	0.009	0.001	0.000	0.000	
0.006	0.051	0.200	0.204	0.101	0.028	0.004	0.000	0.000	
0.001	0.014	0.120	0.210	0.162	0.067	0.014	0.001	0.000	
0.000	0.003	0.055	0.165	0.198	0.122	0.039	0.006	0.000	
0.000	0.000	0.020	0.101	0.189	0.175	0.084	0.019	0.001	
0.000	0.000	0.006	0.049	0.142	0.196	0.142	0.049	0.006	
0.000	0.000	0.001	0.019	0.084	0.175	0.189	0.101	0.020	
0.000	0.000	0.000	0.006	0.039	0.122	0.198	0.165	0.055	
0.000	0.000	0.000	0.001	0.014	0.067	0.162	0.210	0.120	
0.000	0.000	0.000	0.000	0.004	0.028	0.101	0.204	0.200	
0.000	0.000	0.000	0.000	0.001	0.009	0.047	0.146	0.246	
0.000	0.000	0.000	0.000	0.000	0.002	0.015	0.073	0.211	
0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.023	0.113	
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.028	
0.418	0.167	0.023	0.002	0.000	0.000	0.000	0.000	0.000	
0.374	0.315	0.096	0.017	0.002	0.000	0.000	0.000	0.000	
0.158	0.280	0.191	0.058	0.010	0.001	0.000	0.000	0.000	
0.041	0.156	0.239	0.125	0.034	0.005	0.000	0.000	0.000	
0.008	0.060	0.209	0.187	0.080	0.018	0.002	0.000	0.000	
0.001	0.017	0.136	0.208	0.138	0.047	0.008	0.001	0.000	
0.000	0.004	0.068	0.178	0.184	0.094	0.024	0.003	0.000	
0.000	0.001	0.027	0.120	0.193	0.148	0.057	0.009	0.000	
0.000	0.000	0.008	0.064	0.161	0.185	0.107	0.028	0.002	
0.000	0.000	0.002	0.028	0.107	0.185	0.161	0.064	0.008	
0.000	0.000	0.000	0.009	0.057	0.148	0.193	0.120	0.027	
0.000	0.000	0.000	0.003	0.024	0.094	0.184	0.178	0.068	
0.000	0.000	0.000	0.001	0.008	0.047	0.138	0.208	0.136	
0.000	0.000	0.000	0.000	0.002	0.018	0.080	0.187	0.209	
0.000	0.000	0.000	0.000	0.000	0.005	0.034	0.125	0.239	
0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.058	0.191	
0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.017	0.096	
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.023	

p=0.05	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	
0.397	0.150	0.018	0.002	0.000	0.000	0.000	0.000	0.000	
0.376	0.300	0.081	0.013	0.001	0.000	0.000	0.000	0.000	
0.168	0.284	0.172	0.046	0.007	0.001	0.000	0.000	0.000	
0.047	0.168	0.230	0.105	0.025	0.003	0.000	0.000	0.000	
0.009	0.070	0.215	0.168	0.061	0.012	0.001	0.000	0.000	
0.001	0.022	0.151	0.202	0.115	0.033	0.004	0.000	0.000	
0.000	0.005	0.082	0.187	0.166	0.071	0.015	0.001	0.000	
0.000	0.001	0.035	0.138	0.189	0.121	0.037	0.005	0.000	
0.000	0.000	0.012	0.081	0.173	0.167	0.077	0.015	0.001	
0.000	0.000	0.003	0.039	0.128	0.185	0.128	0.039	0.003	
0.000	0.000	0.001	0.015	0.077	0.167	0.173	0.081	0.012	
0.000	0.000	0.000	0.005	0.037	0.121	0.189	0.138	0.035	
0.000	0.000	0.000	0.001	0.015	0.071	0.166	0.187	0.082	
0.000	0.000	0.000	0.000	0.004	0.033	0.115	0.202	0.151	
0.000	0.000	0.000	0.000	0.001	0.012	0.061	0.168	0.215	
0.000	0.000	0.000	0.000	0.000	0.003	0.025	0.105	0.230	
0.000	0.000	0.000	0.000	0.000	0.001	0.007	0.046	0.172	
0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.013	0.081	
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.018	

Table 13.3.1: The Binomial Distribution for n = 16 to 18

Author

"13.3: The Binomial Distribtution Table for n = 16 to n = 18" by Alfie Swan is in the Public Domain, CC0

This page titled 13.3: The Binomial Distribtution Table for n = 16 to n = 18 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.4: The Binomial Distribution Table for n = 19 to n = 20

The Binomial Distribution Table for n = 19 to n = 20

5	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	
7	0.135	0.014	0.001	0.000	0.000	0.000	0.000	0.000	
7	0.285	0.068	0.009	0.001	0.000	0.000	0.000	0.000	
9	0.285	0.154	0.036	0.005	0.000	0.000	0.000	0.000	
3	0.180	0.218	0.087	0.017	0.002	0.000	0.000	0.000	
1	0.080	0.218	0.149	0.047	0.007	0.001	0.000	0.000	
2	0.027	0.164	0.192	0.093	0.022	0.002	0.000	0.000	
9	0.007	0.095	0.192	0.145	0.052	0.008	0.001	0.000	
9	0.001	0.044	0.153	0.180	0.096	0.024	0.002	0.000	
9	0.000	0.017	0.098	0.180	0.144	0.053	0.008	0.000	
9	0.000	0.005	0.051	0.146	0.176	0.098	0.022	0.001	
9	0.000	0.001	0.022	0.098	0.176	0.146	0.051	0.005	
9	0.000	0.000	0.008	0.053	0.144	0.180	0.098	0.017	
9	0.000	0.000	0.002	0.024	0.096	0.180	0.153	0.044	
9	0.000	0.000	0.001	0.008	0.052	0.145	0.192	0.095	
9	0.000	0.000	0.000	0.002	0.022	0.093	0.192	0.164	
9	0.000	0.000	0.000	0.001	0.007	0.047	0.149	0.218	
9	0.000	0.000	0.000	0.000	0.002	0.017	0.087	0.218	
9	0.000	0.000	0.000	0.000	0.000	0.005	0.036	0.154	
9	0.000	0.000	0.000	0.000	0.000	0.001	0.009	0.068	
9	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.014	
3	0.122	0.012	0.001	0.000	0.000	0.000	0.000	0.000	
7	0.270	0.058	0.007	0.000	0.000	0.000	0.000	0.000	
9	0.285	0.137	0.028	0.003	0.000	0.000	0.000	0.000	
9	0.190	0.205	0.072	0.012	0.001	0.000	0.000	0.000	
3	0.090	0.218	0.130	0.035	0.005	0.000	0.000	0.000	
2	0.032	0.175	0.179	0.075	0.015	0.001	0.000	0.000	
9	0.009	0.109	0.192	0.124	0.037	0.005	0.000	0.000	
9	0.002	0.055	0.164	0.166	0.074	0.015	0.001	0.000	
9	0.000	0.022	0.114	0.180	0.120	0.035	0.004	0.000	
9	0.000	0.007	0.065	0.160	0.160	0.071	0.012	0.000	
9	0.000	0.002	0.031	0.117	0.176	0.117	0.031	0.002	
9	0.000	0.000	0.012	0.071	0.160	0.160	0.065	0.007	
9	0.000	0.000	0.004	0.035	0.120	0.180	0.114	0.022	
9	0.000	0.000	0.001	0.015	0.074	0.166	0.164	0.055	
9	0.000	0.000	0.000	0.005	0.037	0.124	0.192	0.109	

5	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5	p=0.6	p=0.7	p=0.8	
9	0.000	0.000	0.000	0.001	0.015	0.075	0.179	0.175	
9	0.000	0.000	0.000	0.000	0.005	0.035	0.130	0.218	
9	0.000	0.000	0.000	0.000	0.001	0.012	0.072	0.205	
9	0.000	0.000	0.000	0.000	0.000	0.003	0.028	0.137	
9	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.058	
9	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.012	

Figure 13.4.1: The Binomial Distribution for n = 19 to n = 20

Author

"13.4: The Binomial Distribution Table for n = 19 to n = 20" by Alfie Swan is in the Public Domain, CC0

This page titled 13.4: The Binomial Distribution Table for n = 19 to n = 20 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.5: The Standard Normal Distriubtion Table for Negative Z-Values

The Left Cumulative Areas Under the Standard Normal Distribution for the Given Negative Z-Value

Areas Under the Standard Normal Distribution

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Note: If the z-value is less than –3.49, then use the approximate area of 0.0001.

Author

"13.5: The Standard Normal Distriubtion Table for Negative Z-Values" by Alfie Swan is in the Public Domain, CC0

This page titled 13.5: The Standard Normal Distriubtion Table for Negative Z-Values is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.6: The Standard Normal Distribution Table for Positive Z-Values

The Left Cumulative Areas Under the Standard Normal Distribution for the Given Positive Z-Value

Areas Under the Standard Normal Distribution

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note: If the z-value is greater than 3.49, then use the approximate area of 0.9999.

Author

"13.6: The Standard Normal Distribution Table for Positive Z-Values" by Alfie Swan is in the Public Domain, CC0

This page titled 13.6: The Standard Normal Distribution Table for Positive Z-Values is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.7: The t-Distribution Table for Degrees of Freedom = n - 1

The t-Distribution for Degrees of Freedom = n - 1 (Where n is the Sample Size)

e for Degrees of Freedom = n – 1

Confidence Intervals	80%	90%	95%	98%	
One Tail (α)	0.10	0.05	0.03	0.01	
eedom Two Tails (α)	0.20	0.10	0.05	0.02	
	3.078	6.314	12.706	31.821	
	1.886	2.920	4.303	6.965	
	1.638	2.353	3.182	4.541	
	1.533	2.132	2.776	3.747	
	1.476	2.015	2.571	3.365	
	1.440	1.943	2.447	3.143	
	1.415	1.895	2.365	2.998	
	1.397	1.860	2.306	2.896	
	1.383	1.833	2.262	2.821	
	1.372	1.812	2.228	2.764	
	1.363	1.796	2.201	2.718	
	1.356	1.782	2.179	2.681	
	1.350	1.771	2.160	2.650	
	1.345	1.761	2.145	2.624	
	1.341	1.753	2.131	2.602	
	1.337	1.746	2.120	2.583	
	1.333	1.740	2.110	2.567	
	1.330	1.734	2.101	2.552	
	1.328	1.729	2.093	2.539	
	1.325	1.725	2.086	2.528	
	1.323	1.721	2.080	2.518	
	1.321	1.717	2.074	2.508	
	1.319	1.714	2.069	2.500	
	1.318	1.711	2.064	2.492	
	1.316	1.708	2.060	2.485	
	1.315	1.706	2.056	2.479	
	1.314	1.703	2.052	2.473	
	1.313	1.701	2.048	2.467	
	1.311	1.699	2.045	2.462	
	1.310	1.697	2.042	2.457	
	1.309	1.696	2.040	2.453	
	1.309	1.694	2.037	2.449	
	1.308	1.692	2.035	2.445	

Confidence Intervals		80%	90%	95%	98%	
		1.307	1.691	2.032	2.441	
		1.306	1.690	2.030	2.438	
		1.306	1.688	2.028	2.434	
		1.305	1.687	2.026	2.431	
		1.304	1.686	2.024	2.429	
		1.304	1.685	2.023	2.426	
		1.303	1.684	2.021	2.423	
		1.301	1.679	2.014	2.412	
		1.299	1.676	2.009	2.403	
		1.297	1.673	2.004	2.396	
		1.296	1.671	2.000	2.390	
		1.295	1.669	1.997	2.385	
		1.294	1.667	1.994	2.381	
		1.293	1.665	1.992	2.377	
		1.292	1.664	1.990	2.374	
		1.292	1.663	1.988	2.371	
		1.291	1.662	1.987	2.368	
		1.291	1.661	1.985	2.366	
		1.290	1.660	1.984	2.364	
		1.286	1.653	1.972	2.345	
		1.284	1.650	1.968	2.339	
		1.284	1.649	1.966	2.336	
		1.283	1.648	1.965	2.334	
;		1.282	1.645	1.960	2.326	

Table 13.7.1: The t-Value Table for Degrees of Freedom = n – 1

Author

"13.7: The t-Distribution Table for Degrees of Freedom = n - 1" by Alfie Swan is in the Public Domain, CC0

This page titled 13.7: The t-Distribution Table for Degrees of Freedom = n - 1 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2

13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2

Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2

rees of Freedom	α = 0.1	α = 0.05	α = 0.01	α = 0.0
1	0.988	0.997	0.999	0.99
2	0.900	0.950	0.990	0.99
3	0.805	0.878	0.959	0.97
4	0.729	0.811	0.917	0.94
5	0.669	0.754	0.875	0.90
6	0.621	0.707	0.834	0.87
7	0.582	0.666	0.798	0.83
8	0.549	0.632	0.765	0.80
9	0.521	0.602	0.735	0.77
10	0.497	0.576	0.708	0.75
11	0.476	0.553	0.684	0.72
12	0.458	0.532	0.661	0.70
13	0.441	0.514	0.641	0.68
14	0.426	0.497	0.623	0.66
15	0.412	0.482	0.606	0.64
16	0.400	0.468	0.590	0.63
17	0.389	0.456	0.575	0.61
18	0.378	0.444	0.561	0.60
19	0.369	0.433	0.549	0.58
20	0.360	0.423	0.537	0.57
21	0.352	0.413	0.526	0.56
22	0.344	0.404	0.515	0.55
23	0.337	0.396	0.505	0.54
24	0.330	0.388	0.496	0.53
25	0.323	0.381	0.487	0.52
26	0.317	0.374	0.479	0.51
27	0.311	0.367	0.471	0.50
28	0.306	0.361	0.463	0.49
29	0.301	0.355	0.456	0.49
30	0.296	0.349	0.449	0.48
31	0.291	0.344	0.442	0.47
32	0.287	0.339	0.436	0.47
33	0.283	0.334	0.430	0.46
34	0.279	0.329	0.424	0.45
35	0.275	0.325	0.418	0.45

rees of Freedom	α = 0.1	α = 0.05	α = 0.01	α = 0.0
36	0.271	0.320	0.413	0.44(
37	0.267	0.316	0.408	0.44
38	0.264	0.312	0.403	0.43
39	0.260	0.308	0.398	0.43(
40	0.257	0.304	0.393	0.42
41	0.254	0.301	0.389	0.42(
42	0.251	0.297	0.384	0.41(
43	0.248	0.294	0.380	0.41
44	0.246	0.291	0.376	0.40
45	0.243	0.288	0.372	0.40
46	0.240	0.285	0.368	0.39
47	0.238	0.282	0.365	0.39
48	0.235	0.279	0.361	0.39
49	0.233	0.276	0.358	0.38
50	0.231	0.273	0.354	0.38

Table 13.8.1: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2

Author

"13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2" by Alfie Swan is in the Public Domain, CC0

This page titled 13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1

The Chi-Square Distribution for Degrees of Freedom n - 1

dom n - 1

α=0.99	α=0.975	α=0.95	α=0.90	α=0.10	α=0.05	α=0.025
0.000	0.001	0.004	0.016	2.706	3.841	5.024
0.020	0.051	0.103	0.211	4.605	5.991	7.378
0.115	0.216	0.352	0.584	6.251	7.815	9.348
0.297	0.484	0.711	1.064	7.779	9.488	11.143
0.554	0.831	1.145	1.610	9.236	11.070	12.833
0.872	1.237	1.635	2.204	10.645	12.592	14.449
1.239	1.690	2.167	2.833	12.017	14.067	16.013
1.646	2.180	2.733	3.490	13.362	15.507	17.535
2.088	2.700	3.325	4.168	14.684	16.919	19.023
2.558	3.247	3.940	4.865	15.987	18.307	20.483
3.053	3.816	4.575	5.578	17.275	19.675	21.920
3.571	4.404	5.226	6.304	18.549	21.026	23.337
4.107	5.009	5.892	7.042	19.812	22.362	24.736
4.660	5.629	6.571	7.790	21.064	23.685	26.119
5.229	6.262	7.261	8.547	22.307	24.996	27.488
5.812	6.908	7.962	9.312	23.542	26.296	28.845
6.408	7.564	8.672	10.085	24.769	27.587	30.191
7.015	8.231	9.390	10.865	25.989	28.869	31.526
7.633	8.907	10.117	11.651	27.204	30.144	32.852
8.260	9.591	10.851	12.443	28.412	31.410	34.170
8.897	10.283	11.591	13.240	29.615	32.671	35.479
9.542	10.982	12.338	14.041	30.813	33.924	36.781
10.196	11.689	13.091	14.848	32.007	35.172	38.076
10.856	12.401	13.848	15.659	33.196	36.415	39.364
11.524	13.120	14.611	16.473	34.382	37.652	40.646
12.198	13.844	15.379	17.292	35.563	38.885	41.923
12.879	14.573	16.151	18.114	36.741	40.113	43.195
13.565	15.308	16.928	18.939	37.916	41.337	44.461
14.256	16.047	17.708	19.768	39.087	42.557	45.722
14.953	16.791	18.493	20.599	40.256	43.773	46.979
15.655	17.539	19.281	21.434	41.422	44.985	48.232
16.362	18.291	20.072	22.271	42.585	46.194	49.480
17.074	19.047	20.867	23.110	43.745	47.400	50.725
17.789	19.806	21.664	23.952	44.903	48.602	51.966
18.509	20.569	22.465	24.797	46.059	49.802	53.203

α=0.99	α=0.975	α=0.95	α=0.90	α=0.10	α=0.05	α=0.025
19.233	21.336	23.269	25.643	47.212	50.998	54.437
19.960	22.106	24.075	26.492	48.363	52.192	55.668
20.691	22.878	24.884	27.343	49.513	53.384	56.896
21.426	23.654	25.695	28.196	50.660	54.572	58.120
22.164	24.433	26.509	29.051	51.805	55.758	59.342
22.906	25.215	27.326	29.907	52.949	56.942	60.561
23.650	25.999	28.144	30.765	54.090	58.124	61.777
24.398	26.785	28.965	31.625	55.230	59.304	62.990
25.148	27.575	29.787	32.487	56.369	60.481	64.201
25.901	28.366	30.612	33.350	57.505	61.656	65.410
26.657	29.160	31.439	34.215	58.641	62.830	66.617
27.416	29.956	32.268	35.081	59.774	64.001	67.821
28.177	30.755	33.098	35.949	60.907	65.171	69.023
28.941	31.555	33.930	36.818	62.038	66.339	70.222
29.707	32.357	34.764	37.689	63.167	67.505	71.420

Table 13.9.1: The Chi-Square Distribution for Degrees of Freedom n - 1

Author

"13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1" by Alfie Swan is in the Public Domain, CC0

This page titled 13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.10: The F-Distribution for Alpha = 0.005

The F-Distribution for Alpha = 0.005

- The first row is the degrees of freedom of the numerator = $n_1 - 1$
- The first column is the degrees of freedom of the numerator = $n_2 - 1$

Degrees of Freedom of the Numerator

Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24334	24426	24505
2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.41	199.42	199.42
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.52	43.39	43.27
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.82	20.70	20.60
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.49	13.38	13.29
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.13	10.03	9.95
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	8.10
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.94
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.15
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.59
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.16
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.84
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.57
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.36
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.18
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	4.03
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.90
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.79
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.70
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.61
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.54
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.47
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.41
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.35
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.30
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.40	3.33	3.26
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.36	3.28	3.22

Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.32	3.25	3.18
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.29	3.21	3.15
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.11
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.89
50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.90	2.82	2.76
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.68

Table 13.10.1: The F-Distribution Table for α = 0.005

Author

"13.10: The F-Distribution for Alpha = 0.005" by Alfie Swan is in the Public Domain, CC0

This page titled 13.10: The F-Distribution for Alpha = 0.005 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.11: The F-Distribution Table for Alpha = 0.01

The F-Distribution Table for Alpha = 0.01

- The first row is the degrees of freedom of the numerator = $n_1 - 1$
- The first column is the degrees of freedom of the numerator = $n_2 - 1$.

Degrees of Freedom of the Numerator

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8
1	4052	5000	5403	5625	5764	5859	5928	5981
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82

Table 13.11.1: The F-Distribution Table for α = 0.01

Author

"13.11: The F-Distribution Table for Alpha = 0.01" by Alfie Swan is in the Public Domain, CC0

This page titled 13.11: The F-Distribution Table for Alpha = 0.01 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.12: The F-Distribution Table for Alpha = 0.025

The F-Distribution Table for Alpha = 0.025

- The first row is the degrees of freedom of the numerator = $n_1 - 1$
- The first column is the degrees of freedom of the numerator = $n_2 - 1$.

Degrees of Freedom of the Numerator

Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	648	800	864	900	922	937	948	957	963	969	973	977	980	983	985
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.42	39.43	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.30	14.28	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.71	8.68	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.80	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.92	2.89	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.79	2.75	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.68	2.65	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.57
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.60	2.56	2.53
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.50
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.53	2.50	2.47
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.44
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.48	2.44	2.41
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.39
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.51	2.47	2.43	2.39	2.36

Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.34
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.48	2.43	2.39	2.36	2.32
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.18
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.11
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.06

Table 13.12.1: The F-Distribution Table for α = 0.025

Author

"13.12: The F-Distribution Table for Alpha = 0.025" by Alfie Swan is in the Public Domain, CC0

This page titled 13.12: The F-Distribution Table for Alpha = 0.025 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.13: The F-Distribution Table for Alpha = 0.05

The F-Distribution Table for Alpha = 0.05

- The first row is the degrees of freedom of the numerator = $n_1 - 1$
- The first column is the degrees of freedom of the numerator = $n_2 - 1$.

Degrees of Freedom of the Numerator

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91	244.69
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89

Table 13.13.1: The F-Distribution Table for α = 0.05

Author

"13.13: The F-Distribution Table for Alpha = 0.05" by Alfie Swan is in the Public Domain, CC0

This page titled 13.13: The F-Distribution Table for Alpha = 0.05 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

13.14: The F-Distribution Table for Alpha = 0.10

The F-Distribution Table for Alpha = 0.10

- The first row is the degrees of freedom of the numerator = $n_1 - 1$
- The first column is the degrees of freedom of the numerator = $n_2 - 1$.

Degrees of Freedom of the Numerator

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.47	60.71	60.90	61.07	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41	9.41	9.42	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22	5.21	5.20	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90	3.89	3.88	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27	3.26	3.25	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90	2.89	2.88	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67	2.65	2.64	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50	2.49	2.48	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38	2.36	2.35	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28	2.27	2.26	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.18	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15	2.13	2.12	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10	2.08	2.07	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.07	2.05	2.04	2.02	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99	1.97	1.95	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96	1.94	1.93	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.95	1.93	1.92	1.90	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.93	1.91	1.89	1.88	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89	1.87	1.86	1.84
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.90	1.87	1.86	1.84	1.83
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86	1.84	1.83	1.81
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.87	1.84	1.83	1.81	1.80
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83	1.81	1.80	1.78
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.84	1.82	1.80	1.79	1.77
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.83	1.81	1.79	1.77	1.76
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.82	1.80	1.78	1.76	1.75

The Degrees of Freedom of the Denominator	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79	1.77	1.75	1.74
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.80	1.78	1.76	1.75	1.73
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77	1.75	1.74	1.72
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71	1.70	1.68	1.66
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.70	1.68	1.66	1.64	1.63
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66	1.64	1.62	1.60

Table 13.14.1: The F-Distribution Table for α = 0.10

Author

"13.14: The F-Distribution Table for Alpha = 0.10" by Alfie Swan is in the Public Domain, CC0

This page titled 13.14: The F-Distribution Table for Alpha = 0.10 is shared under a Public Domain license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

INDEX

B

bar graphs
2.1: Qualitative Data
Bernoulli trial
5.2: Binomial Probability Distribution
binomial experiment
5.2: Binomial Probability Distribution
binomial probability distribution
5.2: Binomial Probability Distribution
5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution

C

central limit theorem
6.3: The Central Limit Theorem
Chebyshev's theorem
3.3: Measures of Variation Version 2
Complementary events
4.2: Theoretical Probability
confidence intervals
7.1: Introduction to Confidence Intervals
confounding variable
1.4: How Not to Do Statistics
correlation
10.1: Bivariate Data and Scatter Plots
Critical Value Method
8.3: z-Test for a Mean

D

deviation
3.3: Measures of Variation Version 2

E

equally likely outcomes
4.2: Theoretical Probability

F

frequency distribution
2.2: Quantitative Data

G

goodness of fit
11.1: Goodness of Fit Test

H

hidden bias
1.4: How Not to Do Statistics
histogram
2.2: Quantitative Data

I

interval of interest
5.4: Poisson Distribution

L

law of large numbers
4.1: Empirical Probability
lurking variable
1.4: How Not to Do Statistics

M

mean
3.1: Measures of Center
median
3.1: Measures of Center
mode
3.1: Measures of Center
mutually exclusive
4.2: Theoretical Probability

N

normal distribution
6.1: Introduction to the Normal Distribution

O

overgeneralization
1.4: How Not to Do Statistics

P

Pareto charts
2.1: Qualitative Data
percentiles
3.4: Measures of Position
pie charts
2.1: Qualitative Data
Poisson distribution
5.4: Poisson Distribution

Q

quartiles
3.4: Measures of Position

R

range
3.3: Measures of Variation Version 2

S

sample variance
3.3: Measures of Variation Version 2
sampling distribution
6.3: The Central Limit Theorem

T

test for independence
11.2: Test for Independence
Theoretical Probabilities
4.2: Theoretical Probability
Type I error
8.2: Type I and II Errors
Type II error
8.2: Type I and II Errors

This page titled Index is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

DETAILED LICENSING

Overview

Title: Statistics C1000 - Introduction to Statistics

Webpages: 108

All licenses found:

- CC BY-SA 4.0: 51.9% (56 pages)
- CC BY 4.0: 32.4% (35 pages)
- Public Domain: 13.9% (15 pages)
- Undeclared: 1.9% (2 pages)

By Page

- Statistics C1000 - Introduction to Statistics - *CC BY-SA 4.0*
  - Front Matter - *CC BY-SA 4.0*
    - TitlePage - *CC BY-SA 4.0*
    - InfoPage - *CC BY-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-SA 4.0*
    - Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework) - *CC BY 4.0*
  - 1: Fundamentals of Statistics - *CC BY-SA 4.0*
    - 1.1: What is Statistics? - *CC BY-SA 4.0*
    - 1.2: Sampling Methods - *CC BY-SA 4.0*
    - 1.3: Experimental Design - *CC BY-SA 4.0*
    - 1.4: How Not to Do Statistics - *CC BY-SA 4.0*
    - 1.5: Chapter 1 - Key Terms and Symbols - *CC BY 4.0*
  - 2: Frequency Distributions, and Graphs - *CC BY-SA 4.0*
    - 2.1: Qualitative Data - *CC BY-SA 4.0*
    - 2.2: Quantitative Data - *CC BY-SA 4.0*
    - 2.3: Other Graphical Representations of Data - *CC BY-SA 4.0*
    - 2.4: Chapter 2 Formulas - *CC BY 4.0*
    - 2.5: Chapter 2 - Key Terms and Symbols - *CC BY 4.0*
  - 3: Descriptive Statistics - *CC BY-SA 4.0*
    - 3.1: Measures of Center - *CC BY-SA 4.0*
    - 3.2: Measures of Variation Version 1 - *CC BY-SA 4.0*
    - 3.3: Measures of Variation Version 2 - *CC BY-SA 4.0*
    - 3.4: Measures of Position - *CC BY-SA 4.0*
    - 3.5: Formulas for Chapter 3 - *CC BY 4.0*
    - 3.6: Chapter 3 - Key Terms and Symbols - *CC BY 4.0*
  - 4: Probability and Combinatorics - *CC BY-SA 4.0*
    - 4.1: Empirical Probability - *CC BY-SA 4.0*
    - 4.2: Theoretical Probability - *CC BY-SA 4.0*
    - 4.3: Addition Rules for Probabilities - *CC BY 4.0*
    - 4.4: Multiplication Rules and Conditional Probability - *CC BY-SA 4.0*
    - 4.5: At Least One Rule and Tree Diagrams - *CC BY 4.0*
    - 4.6: Counting Rules - *CC BY-SA 4.0*
    - 4.7: Chapter 4 Formulas - *CC BY 4.0*
    - 4.8: Chapter 4 - Key Terms and Symbols - *CC BY 4.0*
  - 5: Discrete Probability Distributions - *CC BY-SA 4.0*
    - 5.1: Basics of Probability Distributions - *CC BY-SA 4.0*
    - 5.2: Binomial Probability Distribution - *CC BY-SA 4.0*
    - 5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution - *CC BY-SA 4.0*
    - 5.4: Poisson Distribution - *CC BY 4.0*
    - 5.5: Chapter 5 Formulas - *CC BY 4.0*
    - 5.6: Chapter 5 - Key Terms and Symbols - *CC BY 4.0*
  - 6: Continuous Probability Distribution - *CC BY-SA 4.0*
    - 6.1: Introduction to the Normal Distribution - *CC BY-SA 4.0*
    - 6.2: Finding Probabilities for the Normal Distribution - *CC BY-SA 4.0*
    - 6.3: The Central Limit Theorem - *CC BY-SA 4.0*
    - 6.4: Formulas for Chapter 6 - *CC BY 4.0*
    - 6.5: Chapter 6 - Key Terms and Symbols - *CC BY 4.0*
  - 7: Confidence Interval for One Sample - *CC BY-SA 4.0*
    - 7.1: Introduction to Confidence Intervals - *CC BY-SA 4.0*
    - 7.2: Confidence Interval for the Mean Using z-values - *CC BY 4.0*
    - 7.3: Confidence Interval for the Mean Using t-

values - *CC BY-SA 4.0*
- 7.4: Confidence Interval and Sample Size for the Proportion - *CC BY-SA 4.0*
- 7.5: Formulas for Chapter 7 - *CC BY 4.0*
- 7.6: Chapter 7 - Key Terms and Symbols - *CC BY 4.0*

◦ 8: Hypothesis Testing for One Sample - *CC BY-SA 4.0*
- 8.1: Introduction to Hypothesis Testing - *CC BY-SA 4.0*
- 8.2: Type I and II Errors - *CC BY-SA 4.0*
- 8.3: z-Test for a Mean - *CC BY-SA 4.0*
- 8.4: t-Test for a Mean - *CC BY-SA 4.0*
- 8.5: z-Test for a Proportion - *CC BY-SA 4.0*
- 8.6: Formulas for Chapter 8 - *CC BY 4.0*
- 8.7: Chapter 8 Key Terms and Symbols - *CC BY 4.0*

◦ 9: Hypothesis Testing for Two Samples - *CC BY-SA 4.0*
- 9.1: z-Test for the Difference Between Two Means - *CC BY 4.0*
- 9.2: t-Test for the Difference Between Two Means - *CC BY-SA 4.0*
- 9.3: z-Test for the Difference Between Two Proportions - *CC BY-SA 4.0*
- 9.4: Confidence Intervals for the Difference Between Two Means and Proportions - *CC BY 4.0*
- 9.5: Formulas for Chapter 9 - *CC BY 4.0*
- 9.6: Chapter 9 - Key Terms and Symbols - *CC BY 4.0*

◦ 10: Correlation and Linear Regression - *CC BY-SA 4.0*
- 10.1: Bivariate Data and Scatter Plots - *CC BY-SA 4.0*
- 10.2: Correlation Coefficient - *CC BY 4.0*
- 10.3: Hypothesis Test for Correlation Using r - *CC BY 4.0*
- 10.4: Hypothesis Test for a Correlation Using t-Test - *CC BY-SA 4.0*
- 10.5: Linear Regression - *CC BY-SA 4.0*
- 10.6: Coefficient of Determination and the Standard Error of the Estimate - *CC BY 4.0*
- 10.7: Formulas for Chapter 10 - *CC BY 4.0*
- 10.8: Chapter 10 - Key Terms and Symbols - *CC BY 4.0*

◦ 11: Chi-Square Test - *CC BY-SA 4.0*
- 11.1: Goodness of Fit Test - *CC BY-SA 4.0*
- 11.2: Test for Independence - *CC BY-SA 4.0*
- 11.3: Chapter 11 Formulas - *CC BY 4.0*
- 11.4: Chapter 11 - Key Terms and Symbols - *CC BY 4.0*

◦ 12: Analysis of Variance (ANOVA) - *CC BY 4.0*
- 12.1: Analysis of Variance (ANOVA) - *CC BY 4.0*
- 12.2: Chapter 12 Formulas - *CC BY 4.0*
- 12.3: Chapter 12 - Key Terms and Symbols - *CC BY 4.0*

◦ 13: Statistical Tables - *Public Domain*
- 13.1: The Binomial Distribution Table for n = 2 to n = 10 - *Public Domain*
- 13.2: The Binomial Distribution Table for n = 11 to n = 15 - *Public Domain*
- 13.3: The Binomial Distribtution Table for n = 16 to n = 18 - *Public Domain*
- 13.4: The Binomial Distribution Table for n = 19 to n = 20 - *Public Domain*
- 13.5: The Standard Normal Distriubtion Table for Negative Z-Values - *Public Domain*
- 13.6: The Standard Normal Distribution Table for Positive Z-Values - *Public Domain*
- 13.7: The t-Distribution Table for Degrees of Freedom = n - 1 - *Public Domain*
- 13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2 - *Public Domain*
- 13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1 - *Public Domain*
- 13.10: The F-Distribution for Alpha = 0.005 - *Public Domain*
- 13.11: The F-Distribution Table for Alpha = 0.01 - *Public Domain*
- 13.12: The F-Distribution Table for Alpha = 0.025 - *Public Domain*
- 13.13: The F-Distribution Table for Alpha = 0.05 - *Public Domain*
- 13.14: The F-Distribution Table for Alpha = 0.10 - *Public Domain*

◦ Back Matter - *CC BY-SA 4.0*
- Index - *CC BY-SA 4.0*
- Glossary - *CC BY-SA 4.0*
- Detailed Licensing - *CC BY-SA 4.0*
- Detailed Licensing - *Undeclared*

This page titled Statistics C1000 - Introduction to Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.

DETAILED LICENSING

Overview

Title: Statistics C1000 - Introduction to Statistics

Webpages: 108

All licenses found:

- CC BY-SA 4.0: 51.9% (56 pages)
- CC BY 4.0: 32.4% (35 pages)
- Public Domain: 13.9% (15 pages)
- Undeclared: 1.9% (2 pages)

By Page

- Statistics C1000 - Introduction to Statistics - *CC BY-SA 4.0*
  - Front Matter - *CC BY-SA 4.0*
    - TitlePage - *CC BY-SA 4.0*
    - InfoPage - *CC BY-SA 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY-SA 4.0*
    - Getting Started with MyOpenMath: Access and Key Information (Needed To View Homework) - *CC BY 4.0*
  - 1: Fundamentals of Statistics - *CC BY-SA 4.0*
    - 1.1: What is Statistics? - *CC BY-SA 4.0*
    - 1.2: Sampling Methods - *CC BY-SA 4.0*
    - 1.3: Experimental Design - *CC BY-SA 4.0*
    - 1.4: How Not to Do Statistics - *CC BY-SA 4.0*
    - 1.5: Chapter 1 - Key Terms and Symbols - *CC BY 4.0*
  - 2: Frequency Distributions, and Graphs - *CC BY-SA 4.0*
    - 2.1: Qualitative Data - *CC BY-SA 4.0*
    - 2.2: Quantitative Data - *CC BY-SA 4.0*
    - 2.3: Other Graphical Representations of Data - *CC BY-SA 4.0*
    - 2.4: Chapter 2 Formulas - *CC BY 4.0*
    - 2.5: Chapter 2 - Key Terms and Symbols - *CC BY 4.0*
  - 3: Descriptive Statistics - *CC BY-SA 4.0*
    - 3.1: Measures of Center - *CC BY-SA 4.0*
    - 3.2: Measures of Variation Version 1 - *CC BY-SA 4.0*
    - 3.3: Measures of Variation Version 2 - *CC BY-SA 4.0*
    - 3.4: Measures of Position - *CC BY-SA 4.0*
    - 3.5: Formulas for Chapter 3 - *CC BY 4.0*
    - 3.6: Chapter 3 - Key Terms and Symbols - *CC BY 4.0*
  - 4: Probability and Combinatorics - *CC BY-SA 4.0*
    - 4.1: Empirical Probability - *CC BY-SA 4.0*
    - 4.2: Theoretical Probability - *CC BY-SA 4.0*
    - 4.3: Addition Rules for Probabilities - *CC BY 4.0*
    - 4.4: Multiplication Rules and Conditional Probability - *CC BY-SA 4.0*
    - 4.5: At Least One Rule and Tree Diagrams - *CC BY 4.0*
    - 4.6: Counting Rules - *CC BY-SA 4.0*
    - 4.7: Chapter 4 Formulas - *CC BY 4.0*
    - 4.8: Chapter 4 - Key Terms and Symbols - *CC BY 4.0*
  - 5: Discrete Probability Distributions - *CC BY-SA 4.0*
    - 5.1: Basics of Probability Distributions - *CC BY-SA 4.0*
    - 5.2: Binomial Probability Distribution - *CC BY-SA 4.0*
    - 5.3: Mean, Variance, and Standard Deviation of the Binomial Distribution - *CC BY-SA 4.0*
    - 5.4: Poisson Distribution - *CC BY 4.0*
    - 5.5: Chapter 5 Formulas - *CC BY 4.0*
    - 5.6: Chapter 5 - Key Terms and Symbols - *CC BY 4.0*
  - 6: Continuous Probability Distribution - *CC BY-SA 4.0*
    - 6.1: Introduction to the Normal Distribution - *CC BY-SA 4.0*
    - 6.2: Finding Probabilities for the Normal Distribution - *CC BY-SA 4.0*
    - 6.3: The Central Limit Theorem - *CC BY-SA 4.0*
    - 6.4: Formulas for Chapter 6 - *CC BY 4.0*
    - 6.5: Chapter 6 - Key Terms and Symbols - *CC BY 4.0*
  - 7: Confidence Interval for One Sample - *CC BY-SA 4.0*
    - 7.1: Introduction to Confidence Intervals - *CC BY-SA 4.0*
    - 7.2: Confidence Interval for the Mean Using z-values - *CC BY 4.0*
    - 7.3: Confidence Interval for the Mean Using t-

values - *CC BY-SA 4.0*
- 7.4: Confidence Interval and Sample Size for the Proportion - *CC BY-SA 4.0*
- 7.5: Formulas for Chapter 7 - *CC BY 4.0*
- 7.6: Chapter 7 - Key Terms and Symbols - *CC BY 4.0*

◦ 8: Hypothesis Testing for One Sample - *CC BY-SA 4.0*
- 8.1: Introduction to Hypothesis Testing - *CC BY-SA 4.0*
- 8.2: Type I and II Errors - *CC BY-SA 4.0*
- 8.3: z-Test for a Mean - *CC BY-SA 4.0*
- 8.4: t-Test for a Mean - *CC BY-SA 4.0*
- 8.5: z-Test for a Proportion - *CC BY-SA 4.0*
- 8.6: Formulas for Chapter 8 - *CC BY 4.0*
- 8.7: Chapter 8 Key Terms and Symbols - *CC BY 4.0*

◦ 9: Hypothesis Testing for Two Samples - *CC BY-SA 4.0*
- 9.1: z-Test for the Difference Between Two Means - *CC BY 4.0*
- 9.2: t-Test for the Difference Between Two Means - *CC BY-SA 4.0*
- 9.3: z-Test for the Difference Between Two Proportions - *CC BY-SA 4.0*
- 9.4: Confidence Intervals for the Difference Between Two Means and Proportions - *CC BY 4.0*
- 9.5: Formulas for Chapter 9 - *CC BY 4.0*
- 9.6: Chapter 9 - Key Terms and Symbols - *CC BY 4.0*

◦ 10: Correlation and Linear Regression - *CC BY-SA 4.0*
- 10.1: Bivariate Data and Scatter Plots - *CC BY-SA 4.0*
- 10.2: Correlation Coefficient - *CC BY 4.0*
- 10.3: Hypothesis Test for Correlation Using r - *CC BY 4.0*
- 10.4: Hypothesis Test for a Correlation Using t-Test - *CC BY-SA 4.0*
- 10.5: Linear Regression - *CC BY-SA 4.0*
- 10.6: Coefficient of Determination and the Standard Error of the Estimate - *CC BY 4.0*
- 10.7: Formulas for Chapter 10 - *CC BY 4.0*
- 10.8: Chapter 10 - Key Terms and Symbols - *CC BY 4.0*

◦ 11: Chi-Square Test - *CC BY-SA 4.0*
- 11.1: Goodness of Fit Test - *CC BY-SA 4.0*
- 11.2: Test for Independence - *CC BY-SA 4.0*
- 11.3: Chapter 11 Formulas - *CC BY 4.0*
- 11.4: Chapter 11 - Key Terms and Symbols - *CC BY 4.0*

◦ 12: Analysis of Variance (ANOVA) - *CC BY 4.0*
- 12.1: Analysis of Variance (ANOVA) - *CC BY 4.0*
- 12.2: Chapter 12 Formulas - *CC BY 4.0*
- 12.3: Chapter 12 - Key Terms and Symbols - *CC BY 4.0*

◦ 13: Statistical Tables - *Public Domain*
- 13.1: The Binomial Distribution Table for n = 2 to n = 10 - *Public Domain*
- 13.2: The Binomial Distribution Table for n = 11 to n = 15 - *Public Domain*
- 13.3: The Binomial Distribtution Table for n = 16 to n = 18 - *Public Domain*
- 13.4: The Binomial Distribution Table for n = 19 to n = 20 - *Public Domain*
- 13.5: The Standard Normal Distriubtion Table for Negative Z-Values - *Public Domain*
- 13.6: The Standard Normal Distribution Table for Positive Z-Values - *Public Domain*
- 13.7: The t-Distribution Table for Degrees of Freedom = n - 1 - *Public Domain*
- 13.8: The Critical Values for the Pearson Correlation Matrix for Degrees of Freedom = n -2 - *Public Domain*
- 13.9: The Chi-Square Distribution for Degrees of Freedom = n - 1 - *Public Domain*
- 13.10: The F-Distribution for Alpha = 0.005 - *Public Domain*
- 13.11: The F-Distribution Table for Alpha = 0.01 - *Public Domain*
- 13.12: The F-Distribution Table for Alpha = 0.025 - *Public Domain*
- 13.13: The F-Distribution Table for Alpha = 0.05 - *Public Domain*
- 13.14: The F-Distribution Table for Alpha = 0.10 - *Public Domain*

◦ Back Matter - *CC BY-SA 4.0*
- Index - *CC BY-SA 4.0*
- Glossary - *CC BY-SA 4.0*
- Detailed Licensing - *CC BY-SA 4.0*
- Detailed Licensing - *Undeclared*

This page titled Statistics C1000 - Introduction to Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Toros Berberyan, Tracy Nguyen, and Alfie Swan.