

USING EXCEL FOR INTRODUCTORY STATISTICS



Kurt Sheu
North Park University

Acknowledgements

Original source: OpenStax. *Introductory Statistics*, 2nd edition.

Portions of course content, including the instructional materials and examples, are adapted from *Introductory Statistics (2nd edition)* by Rice University (OpenStax), which is licensed under a [Creative Commons Attribution 4.0 International License](#).

The adapted content has been restructured and selectively modified for a one-semester course format.

Additional materials—such as video tutorials on applying statistical methods using Microsoft Excel—were independently developed by Professor Kurt Sheu and student Blake Paciencia at North Park University. These supplementary resources are original to the course and not part of the OpenStax publication.

Using Excel for Introductory Statistics

Kurt Sheu

North Park University

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 borne 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 were 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 adaptations 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 10/06/2026

TABLE OF CONTENTS

Acknowledgements

Licensing

Goals

Chapter One: Introduction

Introduction

Key Points About Introductory Statistics

Example Chapter 1.3

Chapter Two: Descriptive Statistics

Introduction

Example 2.8

Example 2.14

Example 2.16 and 2.18

Example 2.25

Example 2.28

Example 2.32

Chapter Three: Probability

Terminology

Example- Try It 3.23

Chapter Four: Discrete Random Variable

Example 4.4

Binomial Distribution

Example 4.14

Chapter Five: Continuous Random Variables

Example 5.4

Try It 5.4

Chapter Six: The Normal Distribution

Example 6.8

Example 6.10

Chapter Seven: The Central Limit Theorem

7.1: Sample Limit Theorem for Sample Means (Average)

Example 7.1

Example 7.2

Example 7.3

Chapter Eight: Confidence Interval

8.1: A Single Population Mean using the Normal Distribution

8.2: A Single Population Mean Using the Student T-Distribution

Example 8.8

Chapter Nine: Hypothesis Testing with one sample

Example 9.2

Example 9.16

Example 9.20

Chapter Ten: Hypothesis Testing with Two Samples

Example 10.1

Example 10.2

Homework 80

Chapter Eleven: Chi-Square Distribution

Goodness-of-Fit Test

Example 11.2

Try It 11.3

Tests of Independence

Example 11.6

Chapter Twelve: Linear Regression and Correlation

Introduction

Example 12.6

Understanding Slope

Homework 79

Chapter Thirteen: ANOVA

Introduction

The Null and Alternative Hypotheses

Example 13.3

How to use one-way ANOVA in Excel

Index

Glossary

Detailed Licensing

Licensing

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

Goals

Goals:

1. How to download Excel on PCs, iPad, and Chromebook (see note in the end of manual)
2. Find all functions (e.g. average, minimum, standard deviation, variance) and how to use them
3. Use the Data Analysis Tool
4. Use Excel to solve common homework problems

CHAPTER ONE: INTRODUCTION

Introduction

Key Points About Introductory Statistics

Example Chapter 1.3

Introduction

Statistics is the **science concerned with developing and studying methods for collecting, analyzing**, interpreting and presenting empirical data. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

In statistics, basic terms include: data, population, sample, variable, parameter, statistic, mean, median, mode, standard deviation, frequency distribution, histogram, correlation, and hypothesis testing; essentially describing the group being studied (population), a smaller representative subset (sample), the characteristics being measured (variables), and the numerical summaries used to describe the data (statistics and parameters) across different data analysis methods.

Basic areas covered in an introductory statistics course typically include data collection and presentation, descriptive statistics (like mean, median, mode, and standard deviation), probability, sampling methods, hypothesis testing, confidence intervals, correlation, regression analysis, ANOVA, and basic probability distributions like the normal distribution, focusing on interpreting and analyzing data using statistical methods.

Key Points About Introductory Statistics

Descriptive statistics:

Summarizing data using measures of central tendency (mean, median, mode) and variability (range, standard deviation).

Data visualization:

Presenting data through graphs and charts (histograms, box plots, scatterplots)

Probability:

Understanding the likelihood of events occurring, including basic probability rules and probability distributions

Sampling methods:

Selecting a representative sample from a population to make inferences about the larger group

Hypothesis testing:

Using statistical tests to evaluate claims about population parameters based on sample data

Confidence intervals:

Estimating a range within which a population parameter is likely to fall based on a sample

Correlation and Regression:

Examining relationships between variables, including linear regression to predict one variable based on another

Example Chapter 1.3

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3. Please build a table of data, frequency, relative frequency and cumulative frequency. (Use Excel)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER TWO: DESCRIPTIVE STATISTICS

Introduction

Example 2.8

Example 2.14

Example 2.16 and 2.18

Example 2.25

Example 2.28

Example 2.32

Introduction

Once you have collected data, what will you do with it? In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "**Descriptive Statistics.**" You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A **histogram** consists of contiguous (adjoining) rectangles. The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stem plot) can give you the shape of the data, the center, and the spread of the data.

Example 2.8

Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC college. The number of books is discrete data, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4; 4

5; 5; 5; 5; 5

6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



In chapter 2, has measures of location {min, first quartile (Q_1), medium, third quartile (Q_3), max, Interquartile and percentiles}; measures of center {mean, medium and mode}; and measures of spread {range, variance and standard deviation}.

This page titled [Example 2.8](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Example 2.14

Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

Find:

- a. The interquartile range: $Q_3 - Q_1$
- b. Any outliers in either set

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Example 2.16 and 2.18

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- Find the percentile for 58.
- Find the percentile for 25.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Example 2.25

The data values shown. 10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



This page titled [Example 2.25](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Example 2.28

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mean, median and mode.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Example 2.32

In a fifth-grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER THREE: PROBABILITY

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity.

Terminology

Example- Try It 3.23

Terminology

Experiment: An **experiment** is a planned operation carried out under controlled conditions. A result of an experiment is called an **outcome**.

The **sample space** of an experiment is the set of all possible outcomes. An **event** is any combination of outcomes. The **probability** of any outcome is the **long-term relative frequency** of that outcome. The probability of an event A is written $P(A)$. Sample size is n .

1. $0 \leq P(A) \leq 1$
2. $\sum_{k=1} P(E_k) = 1$, all events E_k are exclusive
3. *Compliment rule:* $P(E^c) = 1 - P(E)$
4. *Addition rule:* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. *Conditional probability:* $P(B) = \frac{P(A \cap B)}{P(A)}$
6. *Multiplication rule:* $P(A \text{ and } B) = P(A)P(B)$
7. *Independent:* if $P(A \text{ and } B) = P(A)P(B)$, or $P(A|B) = P(A)$
8. *Mutually exclusive if* $P(A \text{ and } B) = 0$

Example- Try It 3.23

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Shorts	Totals
Overweight	18	28	14	
Typical weight	20	51	28	
Underweight	12	25	9	
Totals				

Introductory Statistics, 2nd edition' by Rice University, licensed under a Creative Commons Attribution 4.0 International License. Available at: <https://creativecommons.org/licenses/by/4.0/>

- Find the total for each row and column
- Find the probability that randomly chosen individual from this group is Tall
- Find the probability that randomly chosen individual from this group is Overweight and Tall
- Find the probability that randomly chosen individual from this group is Tall given that the individual is Overweight.
- Find the probability that randomly chosen individual from this group is Overweight given that the individual is Tall.
- Find the probability that randomly chosen individual from this group is Tall and Underweight.
- Are the events Overweight and Tall independent?

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



This page titled [Example- Try It 3.23](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

CHAPTER FOUR: DISCRETE RANDOM VARIABLE

A discrete probability distribution function (PDF) has two characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

$$\text{Expected Value} = \mu = \sum (x \cdot P(x))$$

$$\text{Standard Deviation} = \sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)}$$

[Example 4.4](#)

[Binomial Distribution](#)

[Example 4.14](#)

Example 4.4

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

Table 4.6

x	$p(x)$	$x \cdot p(x)$	$(x - \mu)^2 \cdot p(x)$
0	2/50	0	$(0 - 2.1)^2 \cdot 0.04 = 0.1764$
1	11/50	11/50	$(1 - 2.1)^2 \cdot 0.22 = 0.2662$
2	23/50	46/50	$(2 - 2.1)^2 \cdot 0.46 = 0.0046$
3	9/50	27/50	$(3 - 2.1)^2 \cdot 0.18 = 0.1458$
4	4/50	16/50	$(4 - 2.1)^2 \cdot 0.08 = 0.2888$
5	1/50	5/50	$(5 - 2.1)^2 \cdot 0.02 = 0.1682$

Table 4.6

Introductory Statistics, 2nd edition by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

Add the values in the third column of the table to find the expected value of X :

$$\mu = \text{Expected Value} = 105/50 = 2.1$$

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



This page titled [Example 4.4](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Binomial Distribution

There are three characteristics of a binomial experiment.

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p , of a success and probability, q , of a failure remain the same.

The formula is
$$P(x) = C_x^n p^x (1 - p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

The mean, μ , and variance, σ^2 , for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ , is then

$$\sigma = \sqrt{npq}$$

Technology

TI 83&84	Excel
2 nd Vars	Function (Statistical)
Binompdf $P(X =)$ or Binomcdf $P(X \leq)$	BINOM.DIST
Trials (n)	Number _s (number of success)
Probability (p)	Trials (n)
X: number of success	Probability _s (p)
Calculate	Cumulative(false as $p(X=)$ or $p(X \leq)$ true)

This page titled [Binomial Distribution](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Example 4.14

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

- a. What values does x take on?
- b. What is the probability distribution? Find the following probabilities:
 - i. the probability that two pages feature signature artists
 - ii. the probability that at most six pages feature signature artists
 - iii. the probability that more than three pages feature signature artists.
- c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Solution:

- a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
- b. $X \sim B(100, 8/560)$

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER FIVE: CONTINUOUS RANDOM VARIABLES

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are **measured**, not **counted**.
- The entire area under the curve and above the x-axis is equal to **one**.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d . $P(c < x < d)$ is the area under the curve, above the x -axis, to the right of c and the left of d .
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x -axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0).
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

In general, calculus is needed to find the area under the curve for many probability density functions. There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the situation in the best way.

The **uniform distribution** is a continuous probability distribution and is concerned with events that are **equally likely to occur**.

[Example 5.4](#)

[Try It 5.4](#)

Example 5.4

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

- What is the probability that a person waits fewer than 12.5 minutes?
- On the average, how long must a person wait?
- Ninety percent of the time, the time a person must wait falls below what value?

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



Try It 5.4

The total duration of baseball games in the major league in a typical season is uniformly distributed between 447 hours and 521 hours inclusive.

- Find a and b and describe what they represent.
- Write the distribution
- Find the mean
- What is the probability that duration of games for a team for a single season is between 480 and 500 hours?
- What is the 65th percentile for the duration of games for a team in a single season?

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER SIX: THE NORMAL DISTRIBUTION

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

The normal distribution has two parameters (two numerical descriptive measures): **the mean (μ)** and the **standard deviation (σ)**. If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing $X \sim N(\mu, \sigma)$

Example 6.8

Example 6.10

Example 6.8

The final exam scores in a statistics class were normally distributed with a mean 63 and a standard deviation of five.

- Find the probability that a randomly selected student scored more than 65 on the exam.
- Find the probability that a randomly selected student scored less than 85
- Find the probability that a randomly selected student scored between 64 and 75 on the exam
- Find the 90^{th} percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above the k .)
- Find the 70^{th} percentile (that is, find the score k that has 70% of the scores below k and 30% of the scores above the k .)

Solution:

Let $X =$ A score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$, and $\sigma = 5$

[illegible]

zsON+WGnze1cgvjfI1ZAoWdN75OsqMWaomgCyA2PWIIIf64X8Bfy+ClVmtvmUppvmA7XoceJdjQQ9VSzWHh+FAQV5DNAFzQf
mINTUWs2mAwPgChNWIU/+IRWaIMUQYVZGBQBAQA7" width="369">

b. $P(X < 84) = \text{Norm.Dist}(84, 63, 5, \text{true}) = 1(\text{rounded})$

c. $P(64 < X < 75) = \text{Norm.Dist}(75, 63, 5, \text{True}) - \text{Norm.Dist}(64, 63, 5, \text{True}) = 0.4125$

d. $90^{\text{th}} \text{percentil} = \text{Norm.Inv}(0.9, 63, 5) = 69.4$

e. $70^{\text{th}} \text{percentil} = \text{Norm.Inv}(0.7, 63, 5) = 65.6$

Example 6.10

(Please watch the videos below)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER SEVEN: THE CENTRAL LIMIT THEOREM

In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing finite samples size n from a population with a known mean, μ , and a known standard deviation, σ . The first alternative says that if we collect samples of size n with a "large enough n ," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape.

The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

The size of the sample, n , that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**

7.1: Sample Limit Theorem for Sample Means (Average)

Example 7.1

Example 7.2

Example 7.3

7.1: Sample Limit Theorem for Sample Means (Average)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

- a. μ_X = the mean of X
- b. σ_X = the standard deviation of X

If you draw random samples of size n , then as n increases, the random variable $\bar{X}$ which consists of sample means, tends to be **normally distributed** and

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

The **central limit theorem** for sample means says that if you repeatedly draw samples of a given size (such as repeatedly rolling ten dice) and calculate their means, those means tend to follow a normal distribution (the sampling distribution). As **sample sizes increase**, the distribution of means **more closely follows the normal** distribution. The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. Standard deviation is the square root of variance, so the standard deviation of the sampling distribution is the standard deviation of the original distribution divided by the square root of n . The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n , the distribution of the random variable $\bar{X}$, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n , the **sample size**, increases.

The random variable $\bar{X}$ has a different z-score associated with it from that of the random variable X . The mean $\bar{X}$ is the value of $\bar{X}$ in one sample.

$$Z = \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

μ_X is the average of both X and $\bar{X}$. $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} =$ standard deviation of $\bar{X}$ and is called the **standard error of the mean**.

This page titled 7.1: Sample Limit Theorem for Sample Means (Average) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Kurt Sheu.

Example 7.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

- Find the probability that the **sample mean** is between 85 and 92.
- Find the probability that the **sample mean** is less than 85

Solution

a. $\sigma_{\underline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$, $P(85 < \underline{X} < 92) = \text{Norm.Dist}(92, 90, 3, \text{true}) - \text{Norm.Dist}(85, 90, 3, \text{true}) = 0.6997$

b. $P(\underline{X} < 85) = \text{Norm.Dist}(85, 90, 3, \text{true}) = 0.0478$

Example 7.2

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size $n = 50$** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution

It is 0.9977.

$\text{Norm.Dist}(2.3, 2, 0.5/\sqrt{50}, \text{true}) - \text{Norm.Dist}(1.8, 2, 0.5/\sqrt{50}, \text{true})$

Example 7.3

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

- What are the mean and standard deviation for the sample mean ages of tablet users?
- What does the distribution look like?
- Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
- Find the 95th percentile for the sample mean age (to one decimal place).

Solution:

- Since the sample mean tends to target the population mean, we have $\mu_{\bar{X}} = 34$. The sample standard deviation is given by $\sigma_{\bar{X}} = 1.5$
- The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.
- The probability that the sample mean age is more than 30 is 0.9962
- Let k = the 95th percentile. $k = 36.5$

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER EIGHT: CONFIDENCE INTERVAL

8.1: A Single Population Mean using the Normal Distribution

8.2: A Single Population Mean Using the Student T-Distribution

Example 8.8

This page titled [Chapter Eight: Confidence Interval](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

8.1: A Single Population Mean using the Normal Distribution

8.1 A Single Population Mean using the Normal Distribution

A confidence interval for a population mean, when the population standard deviation **is known**, is based on the conclusion of the Central Limit Theorem that the sampling distribution of the sample means follow an approximately normal distribution. Suppose

that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where *EBM* (Margin of Error) = 5.

Constructing the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**,

we need $\bar{x}$ as an estimate for μ and we need the margin of error. Here, the margin of error (*EBM*) is called the **error bound for a**

population mean (abbreviated *EBM*). The sample mean $\bar{x}$ is the **point estimate** of the unknown population mean μ .

The steps to construct and interpret the confidence interval are:

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we already know the population standard deviation σ .
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound *EBM*.
- Construct the confidence interval.

The confidence interval estimate will have the form:

(point estimate - error bound, point estimate + error bound) or,

in symbols, ($\bar{x} - EBM, \bar{x} + EBM$)

The **margin of error** (*EBM*) depends on the **confidence level** (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). **α is related to the confidence level, CL.** α is the probability that the interval does not contain the unknown population parameter. Mathematically, $\alpha + CL = 1$.

Remember to use the area to the LEFT of $\frac{Z\alpha}{2}$ (Critical Value) in this chapter the last two inputs in the Norm.Inv command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound (EBM)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

$$\bullet \text{ EBM} = \left(Z_{\frac{\alpha}{2}} \right) \left(\frac{\sigma}{\sqrt{n}} \right)$$

Interpreting the Confidence Interval

Write a sentence that interprets the estimate in the context of the situation in the problem.

If $\alpha = 0.1$ then 90% C.I., $Z_{0.05} = 1.64485 = 1.645$

If $\alpha = 0.05$, then 95% C.I., $Z_{0.025} = \text{Norm.Inv}(0.025, 0, 1) = 1.96$

If $\alpha = 0.01$, then 99% C.I. $Z_{0.005} = 2.5758$

8.2: A Single Population Mean Using the Student T-Distribution

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation *s as an estimate for σ* and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ .

If you draw a simple random sample of size n from a population that has an approximately normal distribution with mean μ

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

and unknown population standard deviation σ and calculate the t -score

then the t -scores follow a **Student's t-**

distribution with $n - 1$ degrees of freedom. The t -score has the same interpretation as the **z-score**. It measures how far $\frac{x}{\mu}$ is from its mean μ . For each sample size n , there is a different Student's t-distribution. The **degrees of freedom, $n - 1$** , come from the calculation of the sample standard deviation s .

If the population standard deviation is not known, the error bound for a population mean is:

$$\bullet \text{ EBM} = \left(t_{\frac{\alpha}{2}} \right) \left(\frac{s}{\sqrt{n}} \right)$$

- $t_{\frac{\alpha}{2}}$ is the t -score with area to the right equal to $\alpha/2$
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:

$$\left(\bar{x} - \text{EBM}, \bar{x} + \text{EBM} \right).$$

This page titled [8.2: A Single Population Mean Using the Student T-Distribution](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Example 8.8

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Solution:

The 95% confidence interval is (7.30, 9.15). We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER NINE: HYPOTHESIS TESTING WITH ONE SAMPLE

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0 : **The null hypothesis:** It is a statement of no difference between the variables—they are not related. This can often be considered the status quo and as a result if you cannot accept the null it requires some action.

H_a : **The alternative hypothesis: It is a claim** about the population that is contradictory to H_0 and what we conclude when we reject H_0 . **This is usually what the researcher is trying to prove.**

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision**. There are two options for a decision. They are "reject H_0 " if the sample information favours the alternative hypothesis or "do not reject H_0 " or "decline to reject H_0 " if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a :

H_0	H_a
equal (=)	not equal ($\neq$) (two-tailed) or greater than ($>$) (right tailed) or less than ($<$) (left-tailed)

[Example 9.2](#)

[Example 9.16](#)

[Example 9.20](#)

Example 9.2

We want to test whether the mean GPA of students in American colleges is **different** from 2.0 (out of 4.0). The null and alternative hypotheses are:

$$H_0: \mu = 2.0$$

$$H_a: \mu \neq 2.0$$

Example 9.16

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Solution 1

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

Step 1: $H_0: \mu = 65$ $H_a: \mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Step 2: $\alpha = 0.05$

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t .

Step 3: $t_0 = \frac{67 - 65}{s/\sqrt{10}} = 1.9781$

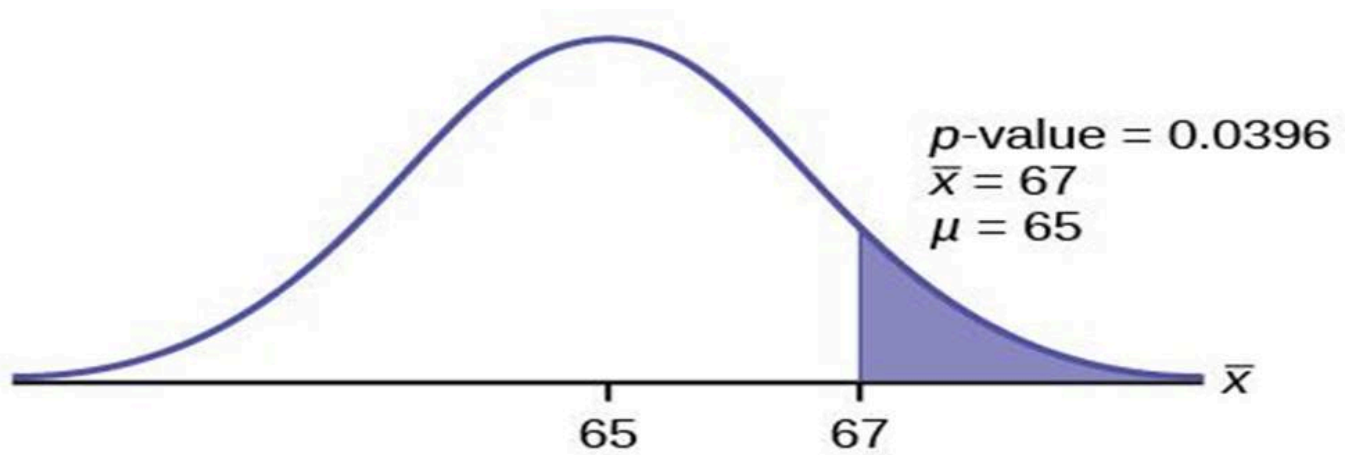
Use t_{df} . Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p -value using the Student's t -distribution:

Step 4: $p\text{-value} = T.\text{Dist.RT}(1.9781, 9) = 0.0396$

$p\text{-value} = P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p -value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.



Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

Compare α and the p -value:

Since $\alpha = 0.05$ and $p\text{-value} = 0.0396$, $p\text{-value} < \alpha$

Make a decision: Since $p\text{-value} < \alpha$, reject H_0 .

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Step 5

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p -value can easily be calculated.

Example 9.20

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98; 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Solution 1

Let's follow a five-step process to answer this statistical question.

1. State the Question: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
 - a. $H_0: \mu = 1$
 - b. $H_a: \mu > 1$

$$\alpha = 0.05$$

- 2.
3. Plan: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's t-distribution. Assume the underlying population is normal.
4. Do the calculations: We will input the sample
5. State the Conclusions: Since the p -value ($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



CHAPTER TEN: HYPOTHESIS TESTING WITH TWO SAMPLES

10.1 Two Population Means with Unknown Standard Deviations

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:
 - if the sample sizes are small, the distributions are important (should be normal)
 - if the sample sizes are large, the distributions are not important (need not be normal)

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples.

In order to account for the variation, we take the difference of the sample means, $\underline{X_1} - \underline{X_2}$, and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of the

difference in sample means, $\underline{X_1} - \underline{X_2}$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

The test statistic (*t*-score) is calculated as follows:

$$t_0 = \frac{(\underline{X_1} - \underline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where:

- s_1 and s_2 , the sample standard deviations, are estimates of σ_1 and σ_2 , respectively.
- σ_1 and σ_2 are the unknown population standard deviations.

- $\underline{x_1}$ and $\underline{x_2}$ are the sample means. μ_1 and μ_2 are the population means.

The number of **degrees of freedom (df)** requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The *df* are not always a whole number. The test statistic calculated previously is approximated by the Student's *t*-distribution with *df* as follows:

Degrees of freedom

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right) \left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right) \left(\frac{(s_2)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are five or larger, the Student's t approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

NOTE

It is not necessary to compute this by hand. A calculator or computer easily computes it.

[Example 10.1](#)

[Example 10.2](#)

[Homework 80](#)

Example 10.1

The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in Table 10.1. Each populations has a normal distribution.

Table 10.1

Table 10.1

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

Solution

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population means.

Random variable: $\bar{X}_g - \bar{X}_b$ = difference in the sample mean amount of time girls and boys play sports each day.

$H_0: \mu_g = \mu_b$ $H_0: \mu_g - \mu_b = 0$

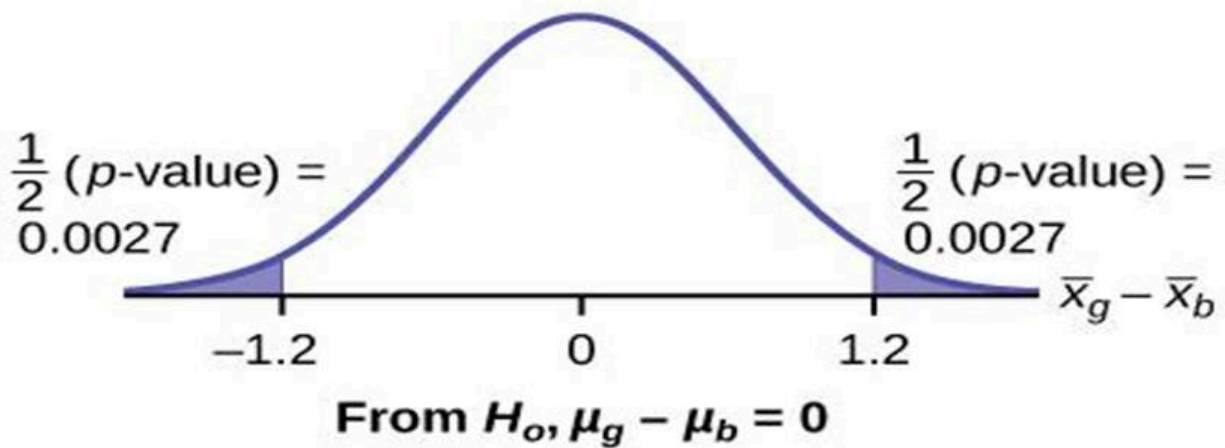
$H_a: \mu_g \neq \mu_b$ $H_a: \mu_g - \mu_b \neq 0$

$$t_0 = \frac{3.2 - 2}{\sqrt{\frac{1}{16} + \frac{0.866^2}{9}}} = 3.1424$$

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. **Do not pool the variances.**

Calculate the p-value using a Student's t-distribution: $p\text{-value} = 0.0054 = T.\text{Dist.}2T(3.1424, 18.8)$

Graph:



Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

$$s_g = 0.866$$

$$s_b = 1$$

$$\text{So, } \frac{x_g}{s_g} - \frac{x_b}{s_b} = 2 - 3.2 = -1.2$$

Half the p -value is below -1.2 and half is above 1.2 .

Make a decision: Since $p\text{-value} < \alpha$, reject H_0 . This means you reject $\mu_g = \mu_b$. The means are different.

Conclusion: There is sufficient evidence that the means are different at the 0.05 level of significant

The words "**the same**" tell you H_0 has an "=". Since there are no other words to indicate H_a , assume it says, "**is different.**" This is a two-tailed test.

Example 10.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes**, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

1. a. Is this a test of two means or two proportions?

Answer: Two Means

1. b. Are the populations standard deviations known or unknown?

Answer: Unknown

1. c. Which distribution do you use to perform the test?

Answer: Student's t

1. d. What is the random variable?

Answer: $\underline{X_A} - \underline{X_B}$

1. e. What are the null and alternative hypothesis (in words and symbols)?

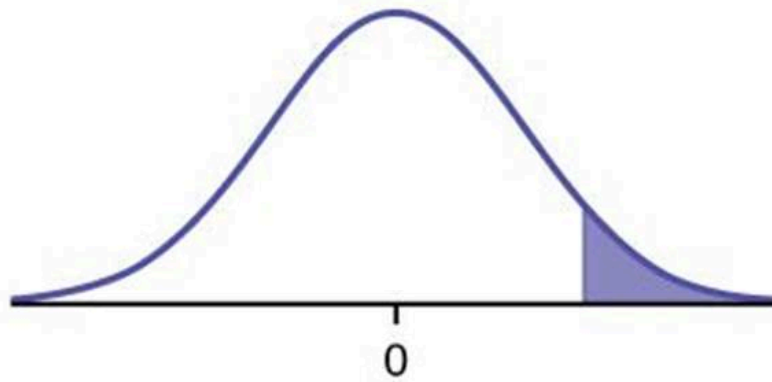
Answer: $H_0: \mu_A - \mu_B = 0$

And $H_1: \mu_A - \mu_B \neq 0$



1. f. Is this test right-tailed, left-tailed or two-tailed?

Answer: right-tailed test



$$\bar{X}_A - \bar{X}_B = 0.5^*$$

$$\text{Note: } \bar{X}_A - \bar{X}_B = 4 - 3.5 = 0.5$$

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

1. g. What is the p-value?

Answer: 0.1928

1. h. Do you reject or not reject the null hypothesis?

Answer: Do not reject

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

Homework 80

At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Table 10.27

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Table 10.27

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Solution:

Step 1 $H_0: \mu_R - \mu_J = 0$ and $H_a: \mu_R - \mu_J \neq 0$

Step 2 $\alpha = 0.05$

Step 3: List all data in two columns in Excel

Go to “data analysis”, choose “t-test: paired two sample for means” t-test=1.507

Step 4: p-value=0.1754

Step 5: p-value > α , do not reject H_0

There is not sufficient evidence to claim the mean difference between their jumping and relaxed times are not zero.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



1. Copy-and-paste the chart from the problem into Excel (or type it).
2. Go to the data tab – click Data analysis – click the t-Test: Paired Two Sample for Means tab.
3. For Variable 1 range: highlight Relaxed Time (seconds) column.
4. For Variable 2 range: highlight Jumping Time (seconds) column.
5. For the hypothesized mean difference look back at the question to find; then type into the box.
6. Make sure the alpha matches up with the alpha given in the question. (If not given use 0.05).
7. Click OK then look for P(T<=t) two tail.
8. Compare p-value to alpha – figure out if we reject or fail to reject the null hypothesis.

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1 ; and arrow to $\neq \mu_2$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

CHAPTER ELEVEN: CHI-SQUARE DISTRIBUTION

Goodness-of-Fit Test

Example 11.2

Try It 11.3

Tests of Independence

Example 11.6

This page titled [Chapter Eleven: Chi-Square Distribution](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:

$$\sum_k \frac{(O - E)^2}{E}$$

where:

- O = **observed values** (data)
- E = **expected values** (from theory= n *percent or common sense)
- k = the number of different data cells or categories
- The degree of freedom is $df = (\text{number of categories} - 1) = (k-1)$
- **The goodness-of-fit test is almost always right-tailed.** If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note

- The expected value for each cell needs to be at least five in order for you to use this test.

Example 11.2

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in Table 11.7. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at 5% significance level.

Table 11.7

Table 11.7

	Monday	Tuesday	Wednesday	Thursday	Friday
# of absences	15	12	9	9	15

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

Solution:

The null and alternative hypotheses are:

H_0 : The absent days occur with equal frequencies, that is, they fit a uniform distribution.

H_a : The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then out of 60 absent days (the total in the sample: $15+12+9+9+15=60$), there would be 12 absences on each weekday. These numbers are the expected (**E**) values. The values in the table are the observed (**O**).

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

Solution table

	Monday	Tuesday	Wednesday	Thursday	Friday
Observed	15	12	9	9	15
Expected	12	12	12	12	12

We find $\sum_k \frac{(O-E)^2}{E} = \frac{(15-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(15-12)^2}{12} = 3.0$

This is the χ^2 test statistic. The df is the number of cells - 1 = 5 - 1 = 4.

[illegible]

Conclusion: AT 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

USING EXCEL

1. Input the data

2. Find the difference between the Observed and expected values

Calculate the Chi-Square value

$$\sum_k \frac{(O - E)^2}{E}$$

and corresponding p-value= CHISQ.DIST.RT(3,4)=0.55788

3. Conclusion: Do not reject H_0

Try It 11.3

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in [Table 11.12](#).

Table 11.2

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

Table 11.12

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

A random sample of 1,000 students from the Eastern United States resulted in the data in [Table 11.13](#).

Table 11.13

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

Table 11.13

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#)

At the 1% significance level, does it appear that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the p -value?

Note: The key to solve this question is to change the percent in Table 11.12 to expected values such as 180, 250, 300, 180, and 90. Frequencies are the observations. Then use the formula.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



1. Copy-and-paste both charts from the problem into Excel, making sure to leave a space in between them.
2. In the space between the two charts label the space Expected Frequency.
3. In each row, multiply the percentage by the sample (1000).
4. In the blank space next to the second chart, label the space $(O-E)^2/E$.
5. Create a formula for each row: the observed frequency minus the expected frequency squared; divided by the expected frequency.
6. Find the sum of all the rows under the label $(O-E)^2/E$. This is your X^2 .
7. Look for the formula CHISQ.DIST.RT and input your X^2 and your degrees of freedom.
8. The answer it gives is your P-Value; compare the p -value and the alpha given in the question – do you reject or fail to reject the null hypothesis?

Tests of Independence

Tests of independence involve using a **contingency table** of observed (data) values.

The test statistic for a test of independence is:

$$\sum_{i,j} \frac{(O - E)^2}{E}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

A test of independence determines whether two factors are independent or not. You first encountered the term independence in [Probability Topics](#). As a review, consider the following example.

Example 11.6

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In [Table 11.15](#) is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Table 11.15

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Table 11.15 Number of Hours Worked Per Week by Volunteer Type (Observed) The table contains **observed (O)** values (data).Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

Is the number of hours volunteered **independent** of the type of volunteer?

USING EXCEL

1. List all data in matrix A (Observed)
2. Add column total and row total and overall total.
3. Use (the row total multiply column total)/overall total to get each cell for new matrix B (Expected)
4. Then use $(\text{Obs}-\text{Exp})^2/\text{Exp}$ to get final matrix C
5. Sum all element of Matrix C to get value X
6. Find $df=(\text{\#of row}-1)\times(\text{\#of column}-1)$
7. Use $\text{Chisq.Distri.RT}(X, df) = p\text{-value}$
8. If $p\text{-value} < \alpha$ reject null hypothesis

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



1. Copy-and-paste the chart from the problem into Excel – label this chart as observed.

2. Copy-and-paste it again right next to the previous chart – label this chart as expected.
3. Delete the data within the second chart and for each open section, create the formula: (row total x column total)/total. This will get you the expected result for each section.
4. Under those two charts copy-and-paste the chart again – label this chart as $(O-E)^2/E$.
5. Once again, delete the data inside this chart and enter into each blank section the formula (observed-expected)^2/ expected. The previous two charts will give you the observed and expected inputs for each section.
6. Next, find the sum of the row totals for the $(O-E)^2/E$ chart. The answer is your X^2 .
7. Look for the formula CHISQ.DIST.RT and input your X^2 and your degrees of freedom.
8. The answer it gives is your P-Value; compare the p-value and the alpha given in the question – do you reject or fail to reject the null hypothesis?

CHAPTER TWELVE: LINEAR REGRESSION AND CORRELATION

Introduction

Example 12.6

Understanding Slope

Homework 79

This page titled [Chapter Twelve: Linear Regression and Correlation](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by Kurt Sheu.

Introduction

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y . The most common and easiest way is a **scatter plot**. A scatter plot shows the **direction** of a **relationship** between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If **x is the independent variable and y the dependent variable**, then we can use a regression line to **predict y** for a given value of x .

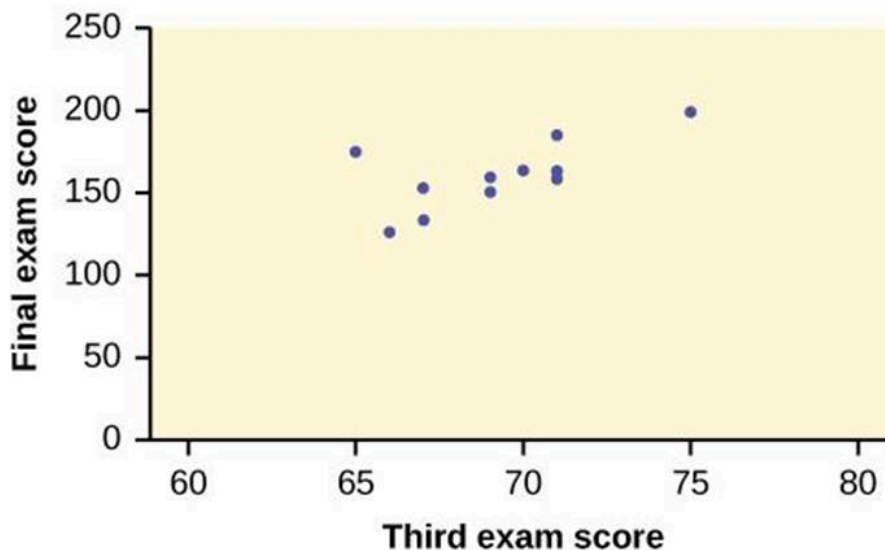
Example 12.6

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table 12.6

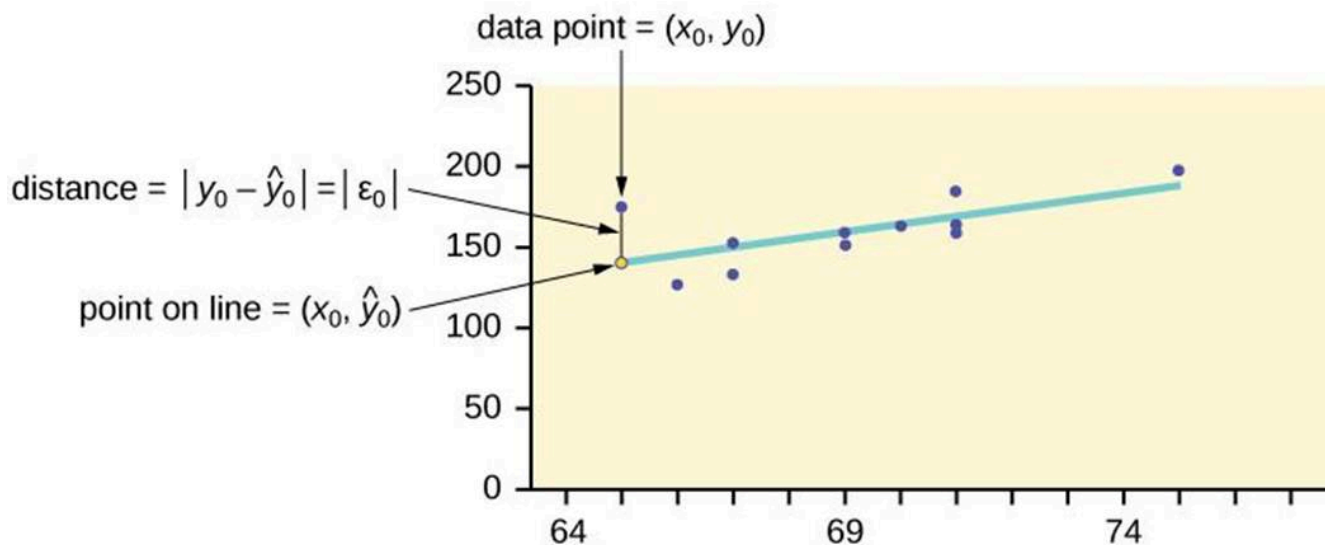
x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



The third exam score, x , is the independent variable and the final exam score, y , is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

The $\hat{y}$ is read "**y hat**" and is the **estimated value of y**. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \epsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The absolute value of residual

measures the vertical distance between the actual value of y and $\hat{y}$. Therefore, there are 11 ϵ values. If you square each ϵ and add, you get $\epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_{11}^2 = \sum_{i=1}^{11} \epsilon_i^2$. This is called the **Sum of Squared Errors (SSE)**. Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx, \text{ where } a = \bar{y} - b\bar{x}, \text{ and } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

The sample means of the x values and the y values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$

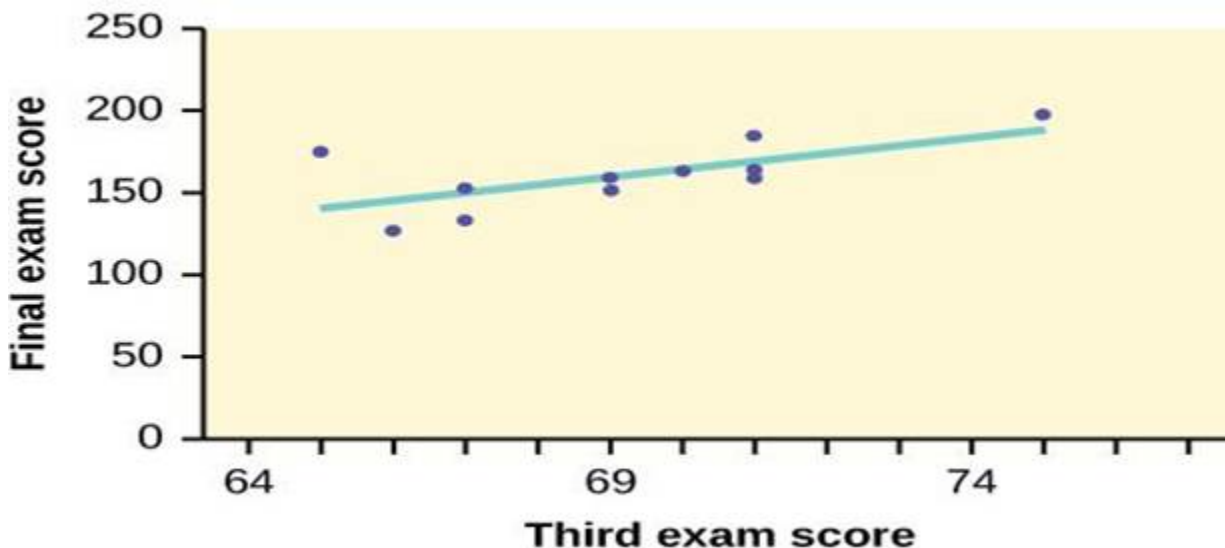
$$b = r \frac{s_y}{s_x}$$

The slope b can be written as $r \frac{s_y}{s_x}$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient, which is discussed in the next section.

The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least-squares regression line**.

THIRD EXAM vs FINAL EXAM EXAMPLE: The graph of the line of best fit for the third-exam/final-exam example is as follows:

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation:

$$\hat{y} = -173.51 + 4.83x$$

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x -values in the sample data, **but not necessarily for x -values outside that domain**. You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x -values in the sample data,

which are between 65 and 75.

Understanding Slope

The slope of the line, b , describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE SLOPE: The slope of the line is $b = 4.83$.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y .

The correlation coefficient is calculated as

$$r = \frac{n \sum (xy) - (\sum x)(\sum y)}{\sqrt{\left[n \sum x^2 - (\sum x)^2 \right] \left[n \sum y^2 - (\sum y)^2 \right]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y , then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of r is always between -1 and $+1$: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y . Values of r close to -1 or to $+1$ indicate a stronger linear relationship between x and y .
- If $r = 0$ there is likely no linear correlation. It is important to view the scatterplot, however, because data that exhibit a curved or horizontal pattern may have a correlation of 0 .
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (**positive correlation**).
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (**negative correlation**).
- The sign of r is the same as the sign of the slope, b , of the best-fit line.

The variable r^2 is called the **coefficient of determination** and is the square of the correlation coefficient, but is usually stated

as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2 when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.

- $1 - r^2$ when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
 - Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
 - Therefore, approximately 56% of the variation ($1 - 0.44 = 0.56$) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Homework 79

Homework #79

The percent of female wage and salary workers who are paid hourly rates is given in [Table 12.30](#) for the years 1979 to 1992.

Table 12.30

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](#).

- Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- Find the correlation coefficient. Is it significant?

- e. Find the estimated percents for 1991 and 1988.
- f. Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
- g. Are there any outliers in the data?
- h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
- i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution:

- a. Check student's solution.
- b. yes
- c. $\hat{y} = -266.8863 + 0.1656x$
- d. 0.9448; Yes
- e. 62.8233; 62.3265
- f. yes
- g. no; (1987, 62.7)
- h. 72.5937; no
- i. slope = 0.1656.

As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



1. Copy-and-paste the chart from the problem into Excel.
2. Go to Insert, highlight all your data, and click on XY Scatter (it's the image of the scatter plot), then in the pop-up click on the first scatter plot.
3. Interpret the scatter plot to find a relationship between the two variables.
4. Go to Data – click Data Analysis – and click on the Regression tab.
5. Highlight the Percent of workers paid hourly rates for the Input Y Range and highlight the Year for the Input X Range.
6. Find the intercept and the X Variable 1 under the Coefficients column.
7. Find the Multiple R (correlation coefficient).
8. Answer the rest of the questions using the Regression analysis and the regression line.

CHAPTER THIRTEEN: ANOVA

Introduction

The Null and Alternative Hypotheses

Example 13.3

How to use one-way ANOVA in Excel

Introduction

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests **comparing averages between more than two groups**, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the F distribution, used for one-way ANOVA, and the test of two variances. This is just a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference **among several group means**. The test uses **variances** to help determine if the means are equal or not. To perform a one-way ANOVA test, there are five basic **assumptions** to be fulfilled:

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. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups:

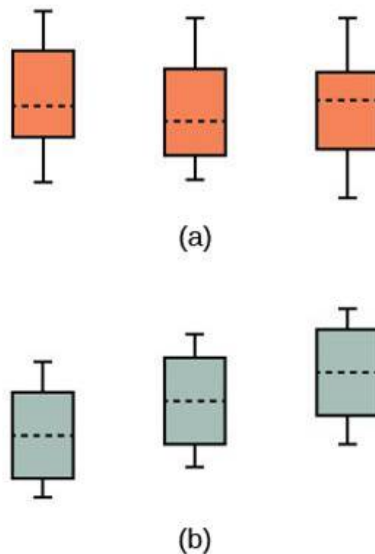
$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

H_a : At least two of the group means $\mu_1, \mu_2, \mu_3, \dots, \mu_k$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H_0: \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



The distribution used for the hypothesis test is a new one. It is called the **F distribution**, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom: one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

To calculate the F ratio, two estimates of the variance are made:

1. Variance between samples: An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation**.
2. Variance within samples: An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance).

When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation**.

- $SS_{between}$ = the **sum of squares** that represents the variation among the different samples

- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in [Descriptive Statistics](#).

$$SST = SSG + SSE = \sum n_i(\underline{x}_i - \underline{x})^2 + \sum (x - \underline{x}_i)^2 = \sum (x - \underline{x})^2$$

MS means "mean square." $MS_{between}$ is the variance between groups, and MS_{within} is the variance within groups.

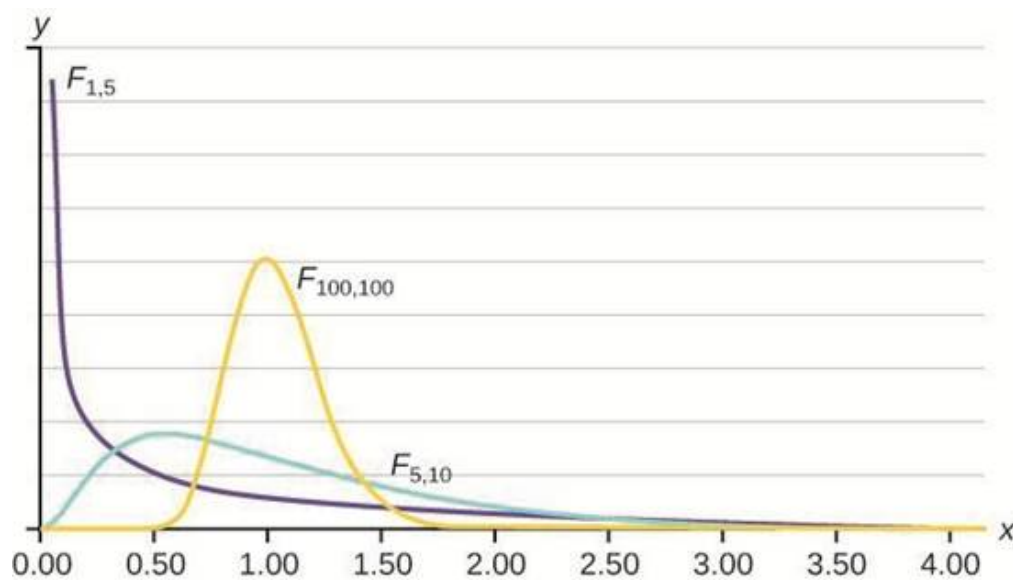
Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$

$$MS_{between} = \frac{SS_{between}}{k-1}, \quad MS_{within} = \frac{SS_{within}}{n-k}$$

Here are some facts about the F distribution.

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of dfs .
3. The F statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
5. Other uses for the F distribution include comparing two variances and two-way Analysis of Variance. Two-Way Analysis is beyond the scope of this chapter.



Introductory Statistics, 2nd edition' by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

This page titled [The Null and Alternative Hypotheses](#) is shared under a [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by Kurt Sheu.

Example 13.3

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in **Table 13.7**.

Table 13.7

	Sorority 1	Sorority 2	Sorority 3	Sorority 4
	2.17	2.63	2.63	3.79
	1.85	1.77	3.78	3.45
	2.83	3.25	4.00	3.08
	1.69	1.86	2.55	2.26
	3.33	2.21	2.45	3.18

Introductory Statistics, 2nd edition' by Rice University, licensed under a Creative Commons Attribution 4.0 International License.

Using a significance level of 1%, is there a difference in mean grades among the sororities?

Solution:

Let $\mu_1, \mu_2, \mu_3, \mu_4$ be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : Not all of the means $\mu_1, \mu_2, \mu_3, \mu_4$ are equal.

Distribution for the test: $F_{3,16}$

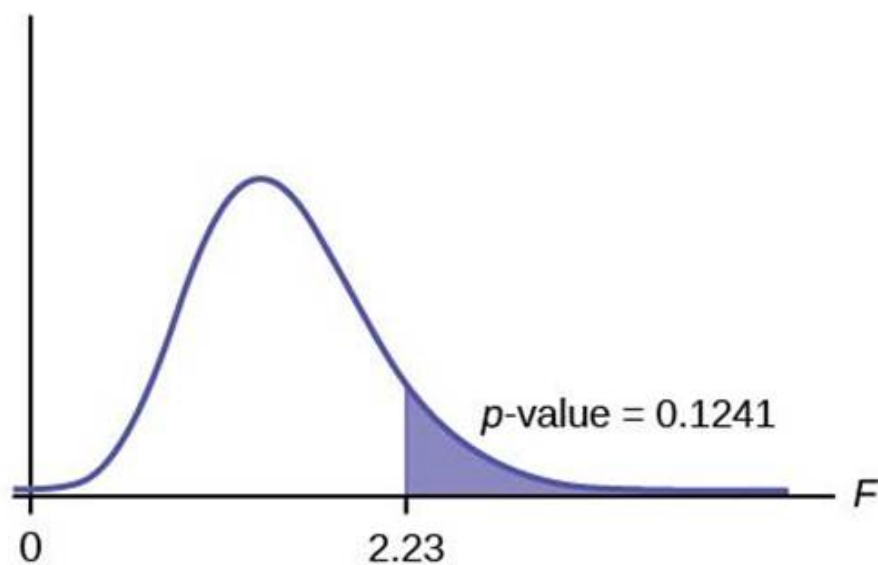
where $k = 4$ groups and $n = 20$ samples in total

$$df(num) = k - 1 = 4 - 1 = 3$$

$$df(denom) = n - k = 20 - 4 = 16$$

Calculate the test statistic: $F = 2.23$

Graph:



Introductory Statistics, 2nd edition by Rice University, licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Figure 13.6

Probability statement: $p\text{-value} = P(F > 2.23) = 0.1241$

Compare α and the $p\text{-value}$: $\alpha = 0.01$

$p\text{-value} = 0.1241$

$\alpha < p\text{-value}$

Make a decision: Since $\alpha < p\text{-value}$, you cannot reject H_0 .

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

How to use one-way ANOVA in Excel

1. Click the Data tab
2. Click Data Analysis
3. Select Anova: Single Factor and click OK
4. Next to Input Range click the up arrow
5. Select the data and click the down arrow
2. Click OK to run the analysis

Results for one-way ANOVA in Excel: Summary statistics

The calculator displays the F statistic, the p-value and the values for the one-way ANOVA table:

$F = 2.2303$

$p = 0.1241$ (p-value)

Factor

$df = 3$

$SS = 2.88732$

$MS = 0.96244$

Error

$df = 16$

$SS = 6.9044$

$MS = 0.431525$

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

[View this YouTube Video online](#)



1. Type the chart from the problem into Excel.
2. Go to Data – click Data Analysis – and click on the Anova: Single Factor tab.
3. Highlight the chart for the input range; ensure that columns are clicked not rows; ensure that the alpha given in the problem matches the alpha listed in the Anova pop-up window.
4. Find the p-value and compare to the alpha — figure out if we reject or fail to reject the null hypothesis.
5. Interpret the results based on the problem: what specifically are you rejecting or failing to reject?

Note:

1. North Park provides Office 365 for free. Just:

• Go to **Office.com/GetOffice365**

- Sign in with your university email
- Download and install
- Excel will activate automatically
- For I-pad Users

Download Microsoft Excel (or all-in-one Microsoft Office) from App Store

Sign in with your North Park University Office 365 account

- For Chromebook Users

Best method: Use Excel Online at office.com

Sign in with your **North Park email**

2. See videos in chapter 2

3. How to install Data Analysis tool in Excel

If you click data, you do not see the data analysis, then click File

Select Options

Select Add-Ins (from the manage drop-down list)

Check Analyst Tool Pack in the pop up that appears

Select Ok

INDEX

No index terms found.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

DETAILED LICENSING

Overview

Title: Using Excel for Introductory Statistics

Webpages: 70

All licenses found:

- [CC BY 4.0](#): 90% (63 pages)
- [Undeclared](#): 5.7% (4 pages)
- [Public Domain](#): 4.3% (3 pages)

By Page

- Using Excel for Introductory Statistics - [CC BY 4.0](#)
 - Front Matter - [CC BY 4.0](#)
 - Acknowledgements - [CC BY 4.0](#)
 - TitlePage - [CC BY 4.0](#)
 - InfoPage - [Public Domain](#)
 - Table of Contents - [Public Domain](#)
 - Goals - [CC BY 4.0](#)
 - Licensing - [Public Domain](#)
 - Chapter One: Introduction - [CC BY 4.0](#)
 - Introduction - [CC BY 4.0](#)
 - Key Points About Introductory Statistics - [CC BY 4.0](#)
 - Example Chapter 1.3 - [CC BY 4.0](#)
 - Chapter Two: Descriptive Statistics - [CC BY 4.0](#)
 - Introduction - [CC BY 4.0](#)
 - Example 2.8 - [CC BY 4.0](#)
 - Example 2.14 - [CC BY 4.0](#)
 - Example 2.16 and 2.18 - [CC BY 4.0](#)
 - Example 2.25 - [CC BY 4.0](#)
 - Example 2.28 - [CC BY 4.0](#)
 - Example 2.32 - [CC BY 4.0](#)
 - Chapter Three: Probability - [CC BY 4.0](#)
 - Terminology - [CC BY 4.0](#)
 - Example- Try It 3.23 - [CC BY 4.0](#)
 - Chapter Four: Discrete Random Variable - [CC BY 4.0](#)
 - Example 4.4 - [CC BY 4.0](#)
 - Binomial Distribution - [CC BY 4.0](#)
 - Example 4.14 - [CC BY 4.0](#)
 - Chapter Five: Continuous Random Variables - [CC BY 4.0](#)
 - Example 5.4 - [CC BY 4.0](#)
 - Try It 5.4 - [CC BY 4.0](#)
 - Chapter Six: The Normal Distribution - [CC BY 4.0](#)
 - Example 6.8 - [CC BY 4.0](#)
 - Example 6.10 - [CC BY 4.0](#)
 - Chapter Seven: The Central Limit Theorem - [CC BY 4.0](#)
 - 7.1: Sample Limit Theorem for Sample Means (Average) - [CC BY 4.0](#)
 - Example 7.1 - [CC BY 4.0](#)
 - Example 7.2 - [CC BY 4.0](#)
 - Example 7.3 - [CC BY 4.0](#)
 - Chapter Eight: Confidence Interval - [CC BY 4.0](#)
 - 8.1: A Single Population Mean using the Normal Distribution - [CC BY 4.0](#)
 - 8.2: A Single Population Mean Using the Student T-Distribution - [CC BY 4.0](#)
 - Example 8.8 - [CC BY 4.0](#)
 - Chapter Nine: Hypothesis Testing with one sample - [CC BY 4.0](#)
 - Example 9.2 - [CC BY 4.0](#)
 - Example 9.16 - [CC BY 4.0](#)
 - Example 9.20 - [CC BY 4.0](#)
 - Chapter Ten: Hypothesis Testing with Two Samples - [CC BY 4.0](#)
 - Example 10.1 - [CC BY 4.0](#)
 - Example 10.2 - [CC BY 4.0](#)
 - Homework 80 - [CC BY 4.0](#)
 - Chapter Eleven: Chi-Square Distribution - [CC BY 4.0](#)
 - Goodness-of-Fit Test - [CC BY 4.0](#)
 - Example 11.2 - [CC BY 4.0](#)
 - Try It 11.3 - [CC BY 4.0](#)
 - Tests of Independence - [CC BY 4.0](#)
 - Example 11.6 - [CC BY 4.0](#)
 - Chapter Twelve: Linear Regression and Correlation - [CC BY 4.0](#)
 - Introduction - [CC BY 4.0](#)
 - Example 12.6 - [CC BY 4.0](#)
 - Understanding Slope - [CC BY 4.0](#)
 - Homework 79 - [CC BY 4.0](#)
 - Chapter Thirteen: ANOVA - [CC BY 4.0](#)
 - Introduction - [CC BY 4.0](#)
 - The Null and Alternative Hypotheses - [CC BY 4.0](#)
 - Example 13.3 - [CC BY 4.0](#)
 - How to use one-way ANOVA in Excel - [CC BY 4.0](#)
 - Back Matter - [Undeclared](#)

- [Index - Undeclared](#)
- [Glossary - Undeclared](#)
- [Detailed Licensing - Undeclared](#)