STATISTICS USING TECHNOLOGY



Kathryn Kozak
Coconino Community College

LibreTexts™

Coconino Community College

Statistics Using Technology

Kathryn Kozak

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

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

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



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

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

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

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

This text was compiled on 06/17/2026

TABLE OF CONTENTS

Licensing

Preface

1: Statistical Basics

- 1.1: What is Statistics?
- 1.2: Sampling Methods
- 1.3: Experimental Design
- 1.4: How Not to Do Statistics

2: Graphical Descriptions of Data

- 2.1: Qualitative Data
- 2.2: Quantitative Data
- 2.3: Other Graphical Representations of Data

3: Examining the Evidence Using Graphs and Statistics

- 3.1: Measures of Center
- 3.2: Measures of Spread
- 3.3: Ranking

4: Probability

- 4.1: Empirical Probability
- 4.2: Theoretical Probability
- 4.3: Conditional Probability
- 4.4: Counting Techniques

5: Discrete Probability Distributions

- 5.1: Basics of Probability Distributions
- 5.2: Binomial Probability Distribution
- 5.3: Mean and Standard Deviation of Binomial Distribution

6: Continuous Probability Distributions

- 6.1: Uniform Distribution
- 6.2: Graphs of the Normal Distribution
- 6.3: Finding Probabilities for the Normal Distribution
- 6.4: Assessing Normality
- 6.5: Sampling Distribution and the Central Limit Theorem

7: One-Sample Inference

- 7.1: Basics of Hypothesis Testing
- 7.2: One-Sample Proportion Test
- 7.3: One-Sample Test for the Mean

8: Estimation

- 8.1: Basics of Confidence Intervals
- 8.2: One-Sample Interval for the Proportion
- 8.3: One-Sample Interval for the Mean

9: Two-Sample Interference

- 9.1: Two Proportions
- 9.2: Paired Samples for Two Means
- 9.3: Independent Samples for Two Means
- 9.4: Which Analysis Should You Conduct?

10: Regression and Correlation

- 10.1: Regression
- 10.2: Correlation
- 10.3: Inference for Regression and Correlation

11: Chi-Square and ANOVA Tests

- 11.1: Chi-Square Test for Independence
- 11.2: Chi-Square Goodness of Fit
- 11.3: Analysis of Variance (ANOVA)

12: Appendix- Critical Value Tables

- 12.1: Critical Values for t-Interval
- 12.2: Normal Critical Values for Confidence Levels

Index

Index

Detailed Licensing

Licensing

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

Preface

I hope you find this book useful in teaching statistics. When writing this book, I tried to follow the GAISE Standards (GAISE recommendations. (2014, January 05). Retrieved from www.amstat.org/education/gais...mendations.pdf), which are

1. Emphasis statistical literacy and develop statistical understanding.
2. Use real data.
3. Stress conceptual understanding, rather than mere knowledge of procedure.
4. Foster active learning in the classroom.
5. Use technology for developing concepts and analyzing data.

To this end, I ask students to interpret the results of their calculations. I incorporated the use of technology for most calculations. Because of that you will not find me using any of the computational formulas for standard deviations or correlation and regression since I prefer students understand the concept of these quantities. Also, because I utilize technology you will not find the standard normal table, Student's t-table, binomial table, chi-square distribution table, and F-distribution table in the book. The only tables I provided were for critical values for confidence intervals since they are more difficult to find using technology. Another difference between this book and other statistics books is the order of hypothesis testing and confidence intervals. Most books present confidence intervals first and then hypothesis tests. I find that presenting hypothesis testing first and then confidence intervals is more understandable for students. Lastly, I have deemphasized the use of the z-test. In fact, I only use it to introduce hypothesis testing, and never utilize it again. You may also notice that when I introduced hypothesis testing and confidence intervals, proportions were introduced before means. However, when two sample tests and confidence intervals are introduced I switched this order. This is because usually many instructors do not discuss the proportions for two samples. However, you might try assigning problems for proportions without discussing it in class. After doing two samples for means, the proportions are similar. Lastly, to aid student understanding and interest, most of the homework and examples utilize real data. Again, I hope you find this book useful for your introductory statistics class. I want to make a comment about the mathematical knowledge that I assumed the students possess. The course for which I wrote this book has a higher prerequisite than most introductory statistics books. However, I do feel that students can read and understand this book as long as they have had basic algebra and can substitute numbers into formulas. I do not show how to create most of the graphs, but most students should have been exposed to them in high school. So I hope the mathematical level is appropriate for your course.

The technology that I utilized for creating the graphs was Microsoft Excel, and I utilized the TI-83/84 graphing calculator for most calculations, including hypothesis testing, confidence intervals, and probability distributions. This is because these tools are readily available to my students. Please feel free to use any other technology that is more appropriate for your students. Do make sure that you use some technology. Statistics Using Technology iv

Acknowledgments

I would like to thank the following people for taking their valuable time to review the book. Their comments and insights improved this book immensely.

- Jane Tanner, Onondaga Community College
- Rob Farinelli, College of Southern Maryland
- Carrie Kinnison, retired engineer
- Sean Simpson, Westchester Community College
- Kim Sonier, Coconino Community College
- Jim Ham, Delta College
- David Straayer, Tacoma Community College
- Kendra Feinstein, Tacoma Community College
- Students of Coconino Community College Students
- Tacoma Community College

I also want to thank Coconino Community College for granting me a sabbatical so that I would have the time to write the book. Lastly, I want to thank my husband Rich and my son Dylan for supporting me in this project. Without their love and support, I would not have been able to complete the book.

New to the Second Edition

The additions to this edition mostly involve adding the commands to create graphs, compute descriptive statistics, finding probabilities, and computing inferential analysis using the open source software R. Another change involve adding an example at the end of chapter 3 that shows analyzing a data set using graphical and numerical descriptions. Another major change was adding a section 9.4 that gives some insight into which inferential analysis should be completed based on a series of questions that should be asked. Lastly, minor explanations were made and corrections were made where necessary.

On a personal note, I wanted to thank my brother, John Matic, his wife Jenelle, and their children Hannah and Eli for their hospitality when writing the first edition. In addition to allowing my family access to their home, John provided numerous examples and data sets for business applications in this book. I inadvertently left this thank you out of the first edition of the book, and for that I apologize. His help and his family's hospitality were invaluable to me.

CHAPTER OVERVIEW

1: Statistical Basics

1.1: What is Statistics?
1.2: Sampling Methods
1.3: Experimental Design
1.4: How Not to Do Statistics

This page titled 1: Statistical Basics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

1.1: What is Statistics?

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 the 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.

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. It is important that you notice the difference and utilize the statistical definitions.

The first thing to decide in a statistical study is whom 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 want 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 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 snap shot 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

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 is a snapshot of the population, and there is more information than is in the picture. The idea is to try to collect a sample that gives you an accurate picture, but you will never know for sure if your picture is the correct picture. Unlike previous mathematics classes where there was always one right answer, in statistics there can be many answers, and you don't know which are right.

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, then you want to calculate a number from the population. This is known as a parameter. As mentioned already, you can't really 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 statistic 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.

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.

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.

Example 1.1.1 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 are questioned.

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

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

Example 1.1.2 stating definitions for qualitative variable

A parking control officer records the manufacturer of every 5th car in the college parking lot in order 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 – set of all names of the manufacturer of cars in the college parking lot.

Sample – set of recorded names of the manufacturer of the cars in college parking lot

Parameter – proportion of each car type calculated from population

Statistic – proportion of each car type calculated from sample

Example 1.1.3 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 the heights of plants when the new plant food is used

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 1.1.4 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 new treatment. This is quantitative data since you will have a number.

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

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

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

Parameter 2 – average life expectancy of all cancer patients given 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

There are different types of quantitative variables, called discrete or continuous. The difference is in how many values can the data have. If you can actually 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.

Definition 1.1.10

Discrete data can only take on particular values like integers. Discrete data are usually things you count.

Definition 1.1.11

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

Example 1.1.5 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 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, 2nd, 3rd), and size of a drink (small, medium, large).

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.

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 arithmetic on this data. Examples of this are height, weight, distance, and time.

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 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 are 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 really 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 numbers, and not numbers that were assigned.

Example 1.1.6 measurement scale

State which measurement scale each is.

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 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 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.

Homework

1. Suppose you want to know how Arizona workers age 16 or older travel to work. To estimate the percentage of people who use the different modes of travel, you take a sample containing 500 Arizona workers age 16 or older. State the individual, variable, population, sample, parameter, and statistic.
2. You wish to estimate the mean cholesterol levels of patients two days after they had a heart attack. To estimate the mean you collect data from 28 heart patients. State the individual, variable, population, sample, parameter, and statistic.
3. Print-O-Matic would like to estimate their mean salary of all employees. To accomplish this they collect the salary of 19 employees. State the individual, variable, population, sample, parameter, and statistic.
4. To estimate the percentage of households in Connecticut which use fuel oil as a heating source, a researcher collects information from 1000 Connecticut households about what fuel is their heating source. State the individual, variable, population, sample, parameter, and statistic.
5. The U.S. Census Bureau needs to estimate the median income of males in the U.S., they collect incomes from 2500 males. State the individual, variable, population, sample, parameter, and statistic.
6. The U.S. Census Bureau needs to estimate the median income of females in the U.S., they collect incomes from 3500 females. State the individual, variable, population, sample, parameter, and statistic.
7. Eyeglassmatic manufactures eyeglasses and they would like to know the percentage of each defect type made. They review $25,891$ defects and classify each defect that is made. State the individual, variable, population, sample, parameter, and statistic.
8. The World Health Organization wishes to estimate the mean density of people per square kilometer, they collect data on 56 countries. State the individual, variable, population, sample, parameter, and statistic
9. State the measurement scale for each.
    a. Cholesterol level
    b. Defect type
    c. Time of first class
    d. Opinion on a 5 point scale, with 5 being strongly agree and 1 being strongly disagree
10. State the measurement scale for each.
    a. Temperature in degrees Celsius
    b. Ice cream flavors available

c. Pain levels on a scale from 1 to 10, 10 being the worst pain ever
d. Salary of employees

Answer

1. See solutions

3. See solutions

5. See solutions

7. See solutions

9.

a. ratio
b. nominal
c. interval
d. ordinal

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

1.2: Sampling Methods

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, there are several techniques that 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 actually 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 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 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 opinions of people, and people who live in the same region may have similar opinions. As you can see each of the sampling techniques have pluses and minuses. Include convenience

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

Ways to select a simple random sample:

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 computer to randomly select the individuals that will be measured.

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.

Definition 1.2.2

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

For example:

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.

Definition 1.2.3

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

For example:

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.

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.

For example:

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.

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. It 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. This sample will not result in a representative sample, and should be avoided.

Definition 1.2.5

Convenience sample is one where the researcher picks individuals to be included that are easy for the researcher to collect.

An example of a convenience sample is if you want to know the opinion of people about the criminal justice system, and you stand on a street corner near the county court house, and questioning the first 10 people who walk by. The people who walk by the county court house are most likely involved in some fashion with the criminal justice system, and their opinion would not represent the opinions of all individuals.

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

Definition 1.2.6

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

Example 1.2.3 sampling type

Banner Health is a several state nonprofit chain of hospitals. Management wants to assess the incident of complications after surgery. They wish to use a sample of surgery patients. Several sampling techniques are described below. Categorize each technique as 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 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 facilities to record any complications from 20 surgeries this week and send in the results.

Solution

a. This is a stratified sample since the patients where separated into different stratum and then random samples were taken from each strata. The problem with this is that some types of surgeries may have more chances for 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 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.

Homework

1. Researchers want to collect cholesterol levels of U.S. patients who had a heart attack two days prior. The following are different sampling techniques that the researcher could use. Classify each as simple random sample, stratified sample, systematic sample, cluster sample, or convenience sample.
   a. The researchers randomly select 5 hospitals in the U.S. then measure the cholesterol levels of all the heart attack patients in each of those hospitals.
   b. The researchers list all of the heart attack patients and measure the cholesterol level of every 25th person on the list.
   c. The researchers go to one hospital on a given day and measure the cholesterol level of the heart attack patients at that time.
   d. The researchers list all of the heart attack patients. They then measure the cholesterol levels of randomly selected patients.
   e. The researchers divide the heart attack patients based on race, and then measure the cholesterol levels of randomly selected patients in each race grouping.
2. The quality control officer at a manufacturing plant needs to determine what percentage of items in a batch are defective. The following are different sampling techniques that could be used by the officer. Classify each as simple random sample, stratified sample, systematic sample, cluster sample, or convenience sample.
   a. The officer lists all of the batches in a given month. The number of defective items is counted in randomly selected batches.
   b. The officer takes the first 10 batches and counts the number of defective items.
   c. The officer groups the batches made in a month into which shift they are made. The number of defective items is counted in randomly selected batches in each shift.
   d. The officer chooses every 15th batch off the line and counts the number of defective items in each chosen batch.

e. The officer divides the batches made in a month into which day they were made. Then certain days are picked and every batch made that day is counted to determine the number of defective items.

3. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a simple random sample.
4. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a stratified sample.
5. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a systematic sample.
6. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a cluster sample.
7. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a convenience sample.

Answer

1.

a. Cluster sample
b. Systematic sample
c. Convenience sample
d. Simple random sample
e. Stratified sample

3. See solutions

5. See solutions

7. See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Experimental Design

The section is an introduction to experimental design. This is how to actually design an experiment or a survey so that they are statistical sound. Experimental design is a very involved process, so this is just a small introduction.

Guidelines for planning a statistical study

1. . Identify the individuals that you are interested in. Realize that you can only make conclusions for these individuals. As an example, if you use a fertilizer on a certain genus of plant, you can't say how the fertilizer will work on any other types of plants. 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 that you can measure, and make sure that 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 in order for you know what conclusions you can make and what individuals you are making the conclusions about.
4. Specify the method for taking measurements or making observations.
5. Determine if you are taking a census or 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 future.

There are two types of 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.

Definition 1.3.2

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

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 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.

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

In an experiment, there are different options.

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 really taking the treatment. Note, not every experiment needs a placebo, such when using

animals or plants. Also, you can't always use a placebo or no treatment. As an 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 because of 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 full-time students to get the treatment and a certain number part-time students to get the treatment. This way some of each type of student gets the treatment and some do not.

Rigorously Controlled Design:

Carefully assign subjects to different treatment groups, so that 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. This type of design is hard to implement since you don't know how many differentiations you would use, and should be avoided.

Matched Pairs Design:

The treatments are given to two groups that can be matched up with each other in some ways. One example would be to measure the effectiveness of a muscle relaxer cream on the right arm and the left arm of individuals, and then for each individual you can match up their right arm measurement with their left arm. Another example of this would be before and after experiments, such as weight before and weight after a diet.

No matter which experiment type you conduct, you should also consider the following:

Replication:

Repetition of an experiment on more than one subject so you can make sure 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:

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:

Double-blind study is where neither the individual nor the researcher knows who is getting which treatment or who is getting the treatment and who is getting 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 that you are collecting the data over. There are three types of time periods that you can consider.

Cross-Sectional Study:

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

Retrospective (or Case-Control) Study:

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

Prospective (or Longitudinal or Cohort) Study:

Data collected in the future from groups sharing common factors.

Homework

1. You want to determine if cinnamon reduces a person's insulin sensitivity. You give patients who are insulin sensitive a certain amount of cinnamon and then measure their glucose levels. Is this an observation or an experiment? Why?
2. You want to determine if eating more fruits reduces a person's chance of developing cancer. You watch people over the years and ask them to tell you how many servings of fruit they eat each day. You then record who develops cancer. Is this an observation or an experiment? Why?

3. A researcher wants to evaluate whether countries with lower fertility rates have a higher life expectancy. They collect the fertility rates and the life expectancies of countries around the world. Is this an observation or an experiment? Why?
4. To evaluate whether a new fertilizer improves plant growth more than the old fertilizer, the fertilizer developer gives some plants the new fertilizer and others the old fertilizer. Is this an observation or an experiment? Why?
5. A researcher designs an experiment to determine if a new drug lowers the blood pressure of patients with high blood pressure. The patients are randomly selected to be in the study and they randomly pick which group to be in. Is this a randomized experiment? Why or why not?
6. Doctors trying to see if a new stint works longer for kidney patients, asks patients if they are willing to have one of two different stints put in. During the procedure the doctor decides which stent to put in based on which one is on hand at the time. Is this a randomized experiment? Why or why not?
7. A researcher wants to determine if diet and exercise together helps people lose weight over just exercising. The researcher solicits volunteers to be part of the study, randomly picks which volunteers are in the study, and then lets each volunteer decide if they want to be in the diet and exercise group or the exercise only group. Is this a randomized experiment? Why or why not?
8. To determine if lack of exercise reduces flexibility in the knee joint, physical therapists ask for volunteers to join their trials. They then randomly select the volunteers to be in the group that exercises and to be in the group that doesn't exercise. Is this a randomized experiment? Why or why not?
9. You collect the weights of tagged fish in a tank. You then put an extra protein fish food in water for the fish and then measure their weight a month later. Are the two samples matched pairs or not? Why or why not?
10. A mathematics instructor wants to see if a computer homework system improves the scores of the students in the class. The instructor teaches two different sections of the same course. One section utilizes the computer homework system and the other section completes homework with paper and pencil. Are the two samples matched pairs or not? Why or why not?
11. A business manager wants to see if a new procedure improves the processing time for a task. The manager measures the processing time of the employees then trains the employees using the new procedure. Then each employee performs the task again and the processing time is measured again. Are the two samples matched pairs or not? Why or why not?
12. The prices of generic items are compared to the prices of the equivalent named brand items. Are the two samples matched pairs or not? Why or why not?
13. A doctor gives some of the patients a new drug for treating acne and the rest of the patients receive the old drug. Neither the patient nor the doctor knows who is getting which drug. Is this a blind experiment, double blind experiment, or neither? Why?
14. One group is told to exercise and one group is told to not exercise. Is this a blind experiment, double blind experiment, or neither? Why?
15. The researchers at a hospital want to see if a new surgery procedure has a better recovery time than the old procedure. The patients are not told which procedure that was used on them, but the surgeons obviously did know. Is this a blind experiment, double blind experiment, or neither? Why?
16. To determine if a new medication reduces headache pain, some patients are given the new medication and others are given a placebo. Neither the researchers nor the patients know who is taking the real medication and who is taking the placebo. Is this a blind experiment, double blind experiment, or neither? Why?
17. A new study is underway to track the eating and exercise patterns of people at different time periods in the future, and see who is afflicted with cancer later in life. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?
18. To determine if a new medication reduces headache pain, some patients are given the new medication and others are given a placebo. The pain levels of a patient are then recorded. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?
19. To see if there is a link between smoking and bladder cancer, patients with bladder cancer are asked if they currently smoke or if they smoked in the past. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?
20. The Nurses Health Survey was a survey where nurses were asked to record their eating habits over a period of time, and their general health was recorded. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?
21. Consider a question that you would like to answer. Describe how you would design your own experiment. Make sure you state the question you would like to answer, then determine if an experiment or an observation is to be done, decide if the question needs one or two samples, if two samples are the samples matched, if this is a randomized experiment, if there is any blinding, and if this is a cross-sectional, retrospective, or prospective study.

Answer

1. Experiment

3. Observation

5. No, see solutions

7. No, see solutions

9. Yes, see solutions

11. Yes, see solutions

13. Double blind, see solutions

15. Blind, see solutions

17. Prospective, see solutions

19. Retrospective, see solutions

21. See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

1.4: How Not to Do Statistics

Many studies are conducted and conclusions are made. However, there are occasions where the study is not conducted in the correct manner 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 bias in it. Bias is where a study may have a certain slant or preference for a certain result. The following are 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 that the results are wrong, but you should scrutinize them on your own to make sure they are sound. As an example if 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 their profits rely on people buying their food, there may be 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. An example of this is when you give fertilizer to some plants and no fertilizer to others, but the no fertilizer plants also are placed in a location that doesn't receive direct sunlight. 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 that you can.

Overgeneralization

Overgeneralization is where you do a study on one group and then try to say that it will happen on 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 stresses using individuals from different groups.

Cause and Effect

Cause and effect is where people decide that one variable causes the other just because the variables are related or 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 number of drownings at the beach and ice cream sales. This does not mean that ice cream sales increasing causes people to drown. Most likely the cause for both increasing 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. As an example, if you take a sample of 5 people's height 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 through 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 significance, which means that a statistical test shows there is evidence to show what you are trying to prove. However, in practice it doesn't mean much or there are other issues to consider. As an example, suppose you find that a new drug for high blood pressure does reduce the blood pressure of patients. When you look at the improvement it actually doesn't amount to a large difference. Even though statistically there is a change, it may not be worth marketing the product because it really isn't that big of a change. Another consideration is that you find the blood pressure medication does improve 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 is 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 believe 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 believe 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 question can also cause hidden bias. An example of this is if you were asked if there should be a fine for texting while driving, but proceeding that question is the question asking if you text while drive. By asking a person if they actually partake in the activity, that person now personalizes the question and that might affect how they answer the next question of creating the fine.

Non-response

Non-response is where you send out a survey but not everyone returns the survey. You can calculate the response rate by dividing the number of returns by the number of surveys sent. Most response rates are around 30-50%. A response rate less than 30% is very poor 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

Voluntary response is where people are asked to respond via phone, email or online. The problem with these is that only people who really care about the topic are likely to call or email. 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 response survey the individuals become involved on their own choosing.

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 use 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 have the same teacher teach both classes. The control group would utilize traditional paper and pencil homework and the treatment group would utilize the 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 the one variable did cause the change in the other variable.

a. Cinnamon was giving to a group of people who have diabetes, and then their blood glucose levels were measured a time period later. All other factors for each person were kept the same. Their glucose levels went down. 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, then 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 tanning spray, then you cannot say that increased use causes 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 humans 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 humans. 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. Just because a drug is safe to use on one species doesn't mean it is safe to use for all species. In fact, ibuprofen is toxic to cats.
b. No. Just because one age group can use it doesn't mean it is safe to use for all age groups. In fact, there has been a link between giving a child under the age of 19 aspirin when they have a fever and Reye's syndrome.

Homework

1. Suppose there is a study where a researcher conducts an experiment to show that deep breathing exercises helps to lower blood pressure. The researcher takes two groups of people and has one group to perform deep breathing exercises and a series of aerobic exercises every day and the other group was asked to refrain from any exercises. The researcher found that the group performing the deep breathing exercises and the aerobic exercises had lower blood pressure. Discuss any issue with this study.
2. Suppose a car dealership offers a low interest rate and a longer payoff period to customers or a high interest rate and a shorter payoff period to customers, and most customers choose the low interest rate and longer payoff period, does that mean that most customers want a lower interest rate? Explain.
3. Over the years it has been said that coffee is bad for you. When looking at the studies that have shown that coffee is linked to poor health, you will see that people who tend to drink coffee don't sleep much, tend to smoke, don't eat healthy, and tend to not exercise. Can you say that the coffee is the reason for the poor health or is there a lurking variable that is the actual cause? Explain.
4. When researchers were trying to figure out what caused polio, they saw a connection between ice cream sales and polio. As ice cream sales increased so did the incident of polio. Does that mean that eating ice cream causes polio? Explain your answer.
5. There is a positive correlation between having a discussion of gun control, which usually occur after a mass shooting, and the sale of guns. Does that mean that the discussion of gun control increases the likelihood that people will buy more guns? Explain.
6. There is a study that shows that people who are obese have a vitamin D deficiency. Does that mean that obesity causes a deficiency in vitamin D? Explain.
7. A study was conducted that shows that polytetrafluoroethylene (PFOA) (Teflon is made from this chemical) has an increase risk of tumors in lab mice. Does that mean that PFOA's have an increased risk of tumors in humans? Explain.
8. Suppose a telephone poll is conducted by contacting U.S. citizens via landlines about their view of gay marriage. Suppose over 50% of those called do not support gay marriage. Does that mean that you can say over 50% of all people in the U.S. do not support gay marriage? Explain.
9. Suppose that it can be shown to be statistically significant that a smaller percentage of the people are satisfied with your business. The percentage before was 87% and is now 85%. Do you change how you conduct business? Explain?
10. You are testing a new drug for weight loss. You find that the drug does in fact statistically show a weight loss. Do you market the new drug? Why or why not?
11. There was an online poll conducted about whether the mayor of Auckland, New Zealand, should resign due to an affair. The majority of people participating said he should. Should the mayor resign due to the results of this poll? Explain.
12. An online poll showed that the majority of Americans believe that the government covered up events of 9/11. Does that really mean that most Americans believe this? Explain.

13. A survey was conducted at a college asking all employees if they were satisfied with the level of security provided by the security department. Discuss how the results of this question could be biased.
14. An employee survey says, “Employees at this institution are very satisfied with working here. Please rate your satisfaction with the institution.” Discuss how this question could create bias.
15. A survey has a question that says, “Most people are afraid that they will lose their house due to economic collapse. Choose what you think is the biggest issue facing the nation today.
    a. Economic collapse
    b. Foreign policy issues
    c. Environmental concerns.” Discuss how this question could create bias.
16. A survey says, “Please rate the career of Roberto Clemente, one of the best right field baseball players in the world.” Discuss how this question could create bias.

Answer

See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

2: Graphical Descriptions of Data

In chapter 1, you were introduced to the concepts of population, which again is a collection of all the measurements from the individuals of interest. Remember, in most cases you can't collect the entire population, so you have to take a sample. Thus, you collect data either through a sample or a census. Now you have a large number of data values. What can you do with them? No one likes to look at just a set of numbers. One thing is to organize the data into a table or graph. Ultimately though, you want to be able to use that graph to interpret the data, to describe the distribution of the data set, and to explore different characteristics of the data. The characteristics that will be discussed in this chapter and the next chapter are:

1. Center: middle of the data set, also known as the average.
2. Variation: how much the data varies.
3. Distribution: shape of the data (symmetric, uniform, or skewed).
4. Qualitative data: analysis of the data
5. Outliers: data values that are far from the majority of the data.
6. Time: changing characteristics of the data over time.

This chapter will focus mostly on using the graphs to understand aspects of the data, and not as much on how to create the graphs. There is technology that will create most of the graphs, though it is important for you to understand the basics of how to create them.

2.1: Qualitative Data
2.2: Quantitative Data
2.3: Other Graphical Representations of Data

This page titled 2: Graphical Descriptions of Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

2.1: Qualitative Data

Remember, qualitative data are words describing a characteristic of the individual. There are several different graphs that are used for qualitative data. These graphs include 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 Excel can make both of them. The first step for either graph is to make a **frequency or relative frequency table**. A frequency table is a summary of the data with counts of how often a data value (or category) occurs.

Example 2.1.1

Suppose you have the following data for which type of car students at a college drive?

Ford, Chevy, Honda, Toyota, Toyota, Nissan, Kia, Nissan, Chevy, Toyota, Honda, Chevy, Toyota, Nissan, Ford, Toyota, Nissan, Mercedes, Chevy, Ford, Nissan, Toyota, Nissan, Ford, Chevy, Toyota, Nissan, Honda, Porsche, Hyundai, Chevy, Chevy, Honda, Toyota, Chevy, Ford, Nissan, Toyota, Chevy, Honda, Chevy, Saturn, Toyota, Chevy, Chevy, Nissan, Honda, Toyota, Toyota, Nissan

Solution

A listing of data is too hard to look at and analyze, so you need to summarize it. First you need to decide the categories. In this case it is relatively easy; just use the car type. However, there are several cars that only have one car in the list. In that case it is easier to make a category called other for the ones with low values. Now just count how many of each type of cars there are. For example, there are 5 Fords, 12 Chevys, and 6 Hondas. This can be put in a frequency distribution:

Table 2.1.1: Frequency Table for Type of Car Data

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

The total of the frequency column should be the number of observations in the data.

Since raw numbers are not as useful to tell other people it is better to create a third column that gives the relative frequency of each category. This is just the frequency divided by the total. As an example for Ford category:

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

This can be written as a decimal, fraction, or percent. You now have a relative frequency distribution:

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

Category	Frequency	Relative Frequency
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 errors.

Now that you have the frequency and relative frequency table, it would be good to display this data using a graph. There are several different types of graphs that can be used: bar chart, pie chart, and Pareto charts.

Bar graphs or charts consist of the frequencies on one axis and the categories on the other axis. Then you draw rectangles for each category with a height (if frequency is on the vertical axis) or length (if frequency is on the horizontal axis) that is equal to the frequency. All of the rectangles should be the same width, and there should be equally width gaps between each bar.

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

Put the frequency on the vertical axis and the category on the horizontal axis.

Then just draw a box above each category whose height is the frequency.

All graphs are drawn using R. The command in R to create a bar graph is:

variable<-c(type in percentages or frequencies for each class with commas in between values)

barplot(variable,names.arg=c("type in name of 1st category", "type in name of 2nd category",…,"type in name of last category"),

ylim=c(0,number over max), xlab="type in label for x-axis", ylab="type in label for y-axis",ylim=c(0,number above maximum y value), main="type in title", col="type in a color") – creates a bar graph of the data in a color if you want.

For this example the command would be:

car<-c(5, 12, 6, 12, 10, 5)

barplot(car, names.arg=c("Ford", "Chevy", "Honda", "Toyota", "Nissan", "Other"), xlab="Type of Car", ylab="Frequency", ylim=c(0,12), main="Type of Car Driven by College Students", col="blue")

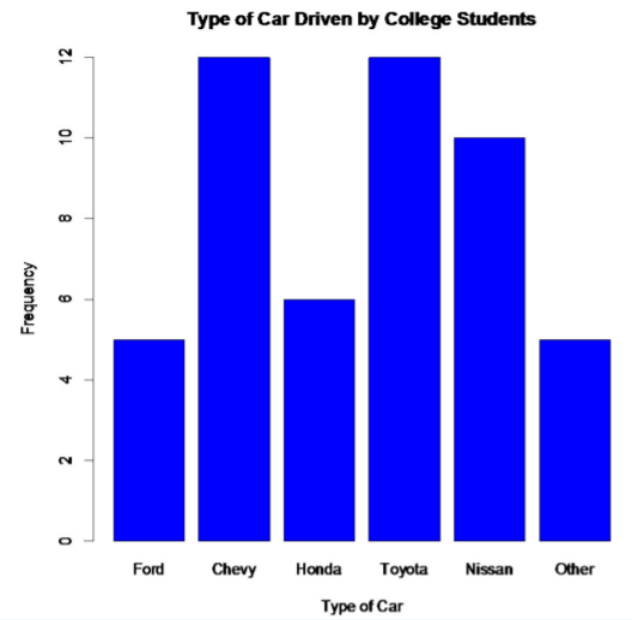


Figure for Type of Car Data

Notice from the graph, you can see that Toyota and Chevy are the more popular car, with Nissan not far behind. Ford seems to be the type of car that you can tell was the least liked, though the cars in the other category would be liked less than a Ford.

Some key features of a bar graph:

- Equal spacing on each axis.
- 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 frequency axis and the categories should be listed on the category axis.
- The bars don't touch.

You can also draw a bar graph using relative frequency on the vertical axis. This is useful when you want to compare two samples with different sample sizes. The relative frequency graph and the frequency graph should look the same, except for the scaling on the frequency axis.

Using R, the command would be:

```
car<-c(0.1, 0.24, 0.12, 0.24, 0.2, 0.1)

barplot(car, names.arg=c("Ford", "Chevy", "Honda", "Toyota", "Nissan", "Other"), xlab="Type of Car", ylab="Relative Frequency", main="Type of Car Driven by College Students", col="blue", ylim=c(0,.25))
```

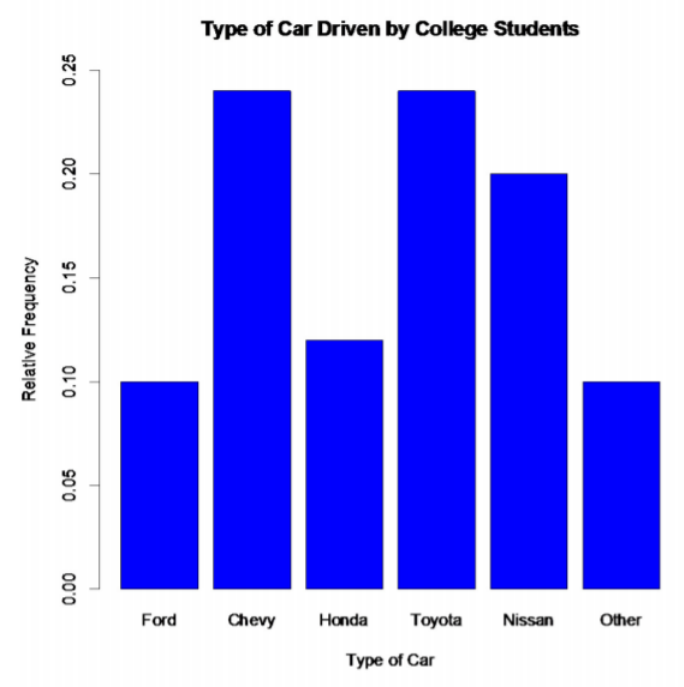


Figure for Type of Car Data

Another type of graph for qualitative data is a pie chart. A pie chart is where you have a circle and you divide pieces of the circle into pie shapes that are proportional to the size of the relative frequency. There are 360 degrees in a full circle. Relative frequency is just the percentage as a decimal. All you have to do to find the angle by multiplying the relative frequency by 360 degrees. Remember that 180 degrees is half a circle and 90 degrees is a quarter of a circle

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

Solution

Then you multiply each relative frequency by 360° to obtain the angle measure for each category.

Table 2.1.3: Pie Chart Angles for Type of Car Data

Category	Relative Frequency	Angle (in degrees (°))
Ford	0.10	36.0

Category	Relative Frequency	Angle (in degrees (°))
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 chart 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.

You can use R to graph the pie chart. In R, the commands would be:

> pie(variable,labels=c("type in name of 1st category", "type in name of 2nd category",…,"type in name of last category"),main="type in title", col=rainbow(number of categories)) – creates a pie chart with a title and rainbow of colors for each category.

For this example, the commands would be:

> car<-c(5, 12, 6, 12, 10, 5)
>
> pie(car, labels=c("Ford, 10%", "Chevy, 24%", "Honda, 12%", "Toyota, 24%", "Nissan, 20%", "Other, 10%"), main="Type of Car Driven by College Students", col=rainbow(6))

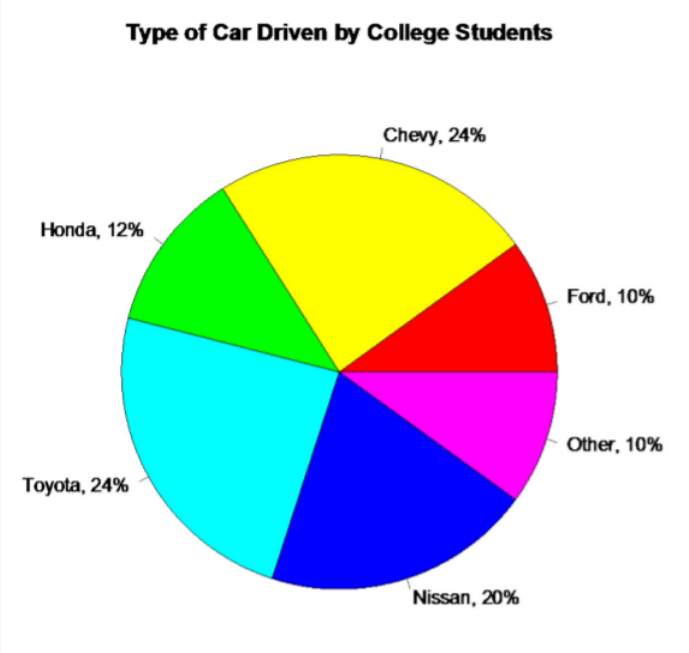


Figure 2.1.3: Pie Chart for Type of Car Data

As you can see from the graph, Toyota and Chevy are more popular, while the cars in the other category are liked the least. Of the cars that you can determine from the graph, Ford is liked less than the others.

Pie charts are useful for comparing sizes of categories. Bar charts show similar information. It really doesn't matter which one you use. It really is a personal preference and also what information you are trying to address. However, pie charts are best when you only have a few categories and the data can be expressed as a percentage. The data doesn't have to be percentages to draw the pie chart, but if a data value can fit into multiple categories, you cannot use a pie chart. As an example, if you are asking people about

what their favorite national park is, and you say to pick the top three choices, then the total number of answers can add up to more than 100% of the people involved. So you cannot use a pie chart to display the favorite national park.

A third type of qualitative data graph is a **Pareto chart**, which is just a bar chart with the bars sorted with the highest frequencies on the left. Here is the Pareto chart for the data in Example 2.1.1.

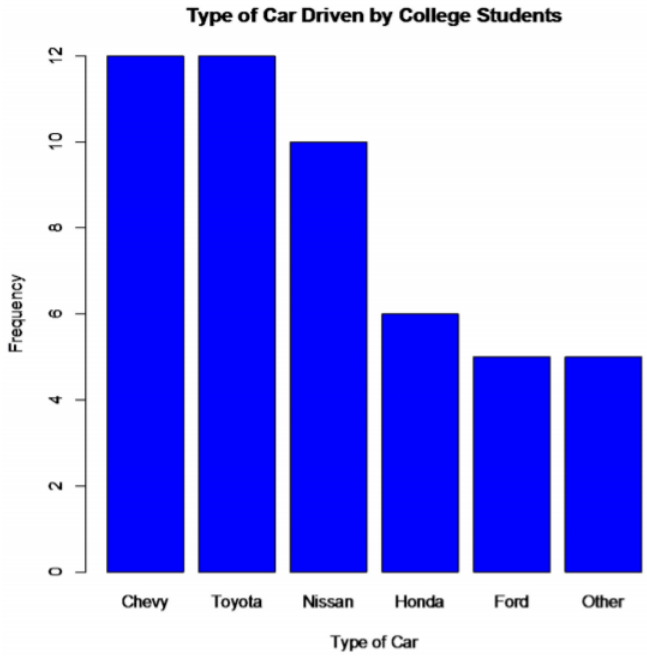


Figure 2.1.4: Pareto Chart for Type of Car Data

The advantage of Pareto charts is that you can visually see the more popular answer to the least popular. 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 type of questions like these.

There are many other types of graphs that can be used on qualitative data. There are spreadsheet software packages that will create most of them, and it is better to look at them to see what can be done. It depends on your data as to which may be useful. The next example illustrates one of these types known as a multiple bar graph.

Example 2.1.4 multiple bar graph

In the Wii Fit game, you can do four different types of exercises: yoga, strength, aerobic, and balance. The Wii system keeps track of how many minutes you spend on each of the exercises everyday. The following graph is the data for Dylan over one week time period. Discuss any indication you can infer from the graph.

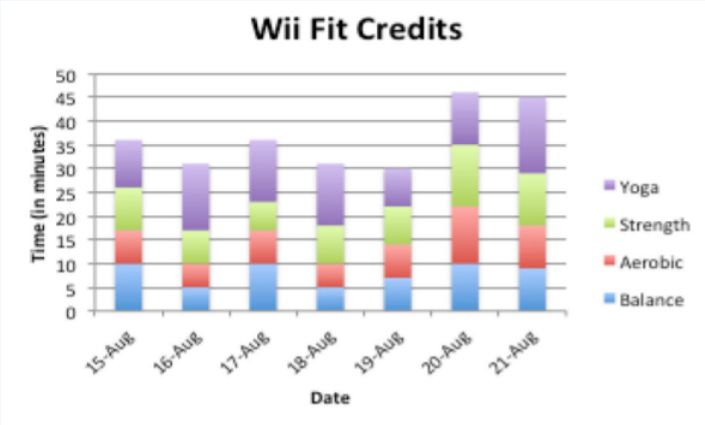


Figure 2.1.5: Multiple Bar Chart for Wii Fit Data

Solution

It appears that Dylan spends more time on balance exercises than on any other exercises on any given day. He seems to spend less time on strength exercises on a given day. There are several days when the amount of exercise in the different categories is almost equal.

The usefulness of a multiple bar graph is the ability to compare several different categories over another variable, in Example 2.1.4 the variable would be time. This allows a person to interpret the data with a little more ease.

Homework

1. Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in Example 2.1.4.

Activity	Grind	Multicoat	Assemble	Make frames	Receive finished	Unknown
Number of lenses	18872	12105	4333	25880	26991	1508

Table 2.1.4*: Data for Eyeglassomatic*

Grind means that they ground the lenses and put them in frames, multicoat means that they put tinting or scratch resistance coatings on lenses and then put them in frames, assemble means that they receive frames and lenses from other sources and put them together, make frames means that they make the frames and put lenses in from other sources, receive finished means that they received glasses from other source, and unknown means they do not know where the lenses came from. Make a bar chart and a pie chart of this data. State any findings you can see from the graphs.

2. To analyze how Arizona workers ages 16 or older travel to work the percentage of workers using carpool, private vehicle (alone), and public transportation was collected. Create a bar chart and pie chart of the data in Example 2.1.5. State any findings you can see from the graphs.

Table 2.1.5: Data of Travel Mode for Arizona Workers

Transportation type	Percentage
Carpool	11.6%
Private Vehicle (Alone)	75.8%
Public Transportation	2.0%
Other	10.6%

3. The number of deaths in the US due to carbon monoxide (CO) poisoning from generators from the years 1999 to 2011 are in table #2.1.6 (Hinatov, 2012). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Table 2.1.6: Data of Number of Deaths Due to CO Poisoning

Region	Number of Deaths from CO While Using a Generator
Urban Core	401
Sub-Urban	97
Large Rural	86
Small Rural/Isolated	111

4. In Connecticut households use gas, fuel oil, or electricity as a heating source. Example 2.1.7 shows the percentage of households that use one of these as their principle heating sources ("Electricity usage," 2013), ("Fuel oil usage," 2013), ("Gas usage," 2013). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Table 2.1.7: Data of Household Heating Sources

Heating Source	Percentage
Electricity	15.3%

Heating Source	Percentage
Fuel Oil	46.3%
Gas	35.6%
Other	2.85

5. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made during the time period of January 1 to March 31. Example 2.1.8 gives the defect and the number of defects. Create a Pareto chart of the data and then describe what this tells you about what causes the most defects.

Table 2.1.8: Data of Defect Type

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

6. People in Bangladesh were asked to state what type of birth control method they use. The percentages are given in Example 2.1.9 ("Contraceptive use," 2013). Create a Pareto chart of the data and then state any findings you can from the graph.

Table 2.1.9: Data of Birth Control Type

Method	Percentage
Condom	4.50%
Pill	28.50%
Periodic Abstinence	4.90%
Injection	7.00%
Female Sterilization	5.00%
IUD	0.90%
Male Sterilization	0.70%
Withdrawal	2.90%
Other Modern Methods	0.70%
Other Traditional Methods	0.60%

7. The percentages of people who use certain contraceptives in Central American countries are displayed in *Graph 2.1.6* ("Contraceptive use," 2013). State any findings you can from the graph.

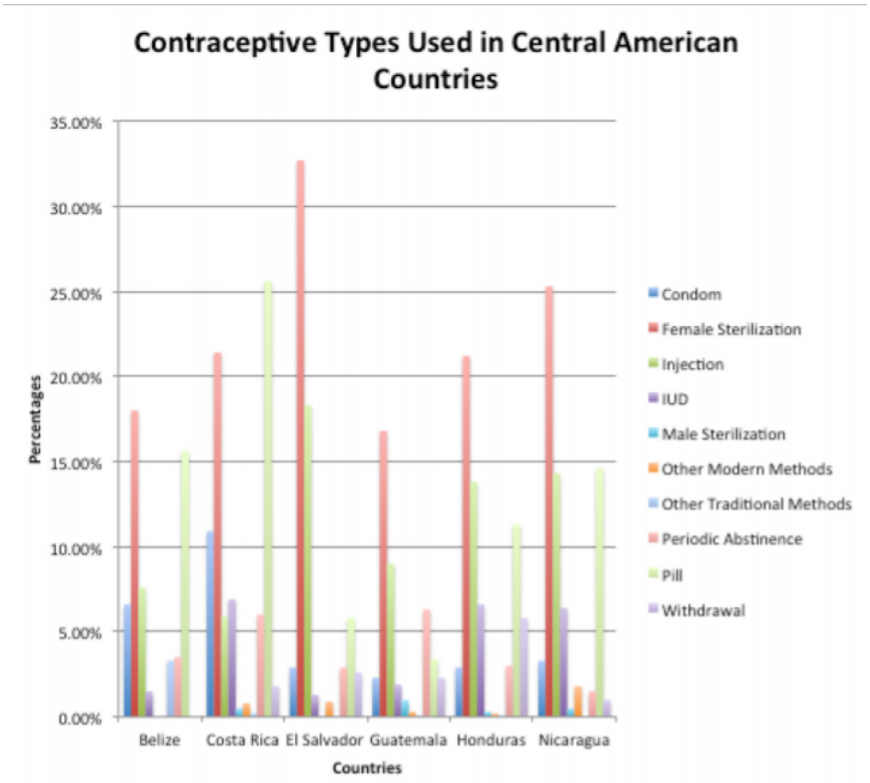


Figure 2.1.6*: Multiple Bar Chart for Contraceptive Types*

Answer

See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Quantitative Data

The graph for quantitative data looks similar to a bar graph, except there are some major differences. First, in 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 just by putting the categories in different orders. With quantitative data, the data are in specific orders, since you are dealing with numbers. With quantitative data, you can talk about a distribution, since the shape only changes a little bit depending on how many categories you set up. This is called a **frequency distribution**.

This leads to the second difference from bar graphs. In a bar graph, the categories that you made in the frequency table were determined by you. In quantitative data, the categories are numerical categories, and the numbers are determined by how many categories (or what are called classes) you choose. If two people have the same number of categories, then they will have the same frequency distribution. Whereas in qualitative data, there can be many different categories depending on the point of view of the author.

The third difference is that the categories touch with quantitative data, and there will be no gaps in the graph. The reason that bar graphs have gaps is to show that the categories do not continue on, like they do in quantitative data. Since the graph for quantitative data is different from qualitative data, it is given a new name. The name of the graph is a **histogram**. 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.

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. Usually the number of classes is between five and twenty. Five classes are used if there are a small number of data points 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 you would have to add another class just for it. If you round up, then your largest data value will fall in the last class, and there are no issues.
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. In order for the classes to actually touch, then 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 figure out the number of data points that fall in each class, go through each data value and see which class boundaries it is between. Utilizing tally marks may be helpful in counting 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 you 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. As an example, if your data have 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

Example 2.2.1 contains the amount of rent paid every month for 24 students from a statistics course. Make a relative 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 directions to say to use 7 classes.
3. Find the class width:
   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.
   $350 + 315 = 665, 665 + 315 = 980, 980 + 315 = 1295 \rightleftharpoons,$
   The upper limit is one less than the next lower limit: so for the first class the upper class limit would be $665 - 1 = 664$.
   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 class boundaries. Add 0.5 to the upper class limit for the last class's boundary.
   $350 - 0.5 = 349.5, \quad 665 - 0.5 = 664.5, \quad 980 - 0.5 = 979.5, \quad 1295 - 0.5 = 1294.5 \rightleftharpoons$
   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:
   midpoint $= \frac{\text{lower limit } + \text{ upper limit}}{2}$
   $\frac{350 + 664}{2} = 507, \frac{665 + 979}{2} = 822, \rightleftharpoons$
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.

Table 2.2.2: *Frequency Distribution for Monthly Rent*

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency
350-664	349.5-664.5	507	𝍸	4
665-979	664.5-979.5	822	𝍸 |||	8
980-1294	979.5-1294.5	1137	𝍸	5
1295-1609	1294.5-1609.5	1452	𝍸 |	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

Make sure the total of the frequencies is the same as the number of data points.

R command for a frequency distribution:

To create a frequency distribution:

summary(variable) – so you can find out the minimum and maximum.

breaks = seq(min, number above max, by = class width)

breaks – so you can see the breaks that R made.

variable.cut=cut(variable, breaks, right=FALSE) – this will cut up the data into the classes.

variable.freq=table(variable.cut) – this will create the frequency table.

variable.freq – this will display the frequency table.

For the data in Example 2.2.1, the R command would be:

rent<-c(1500, 1350, 350, 1200, 850, 900, 1500, 1150, 1500, 900, 1400, 1100, 1250, 600, 610, 960, 890, 1325, 900, 800, 2550, 495, 1200, 690) summary(rent)

Output:

Min	1st Qu.	Median	Mean	3rd Qu.	Max
350	837.5	1030.0	1082.0	1331.0	2550.0

breaks=seq(350, 3000, by = 315)
breaks

Output:
[1] 350 665 980 1295 1610 1925 2240 2555 2870
These are your lower limits of the frequency distribution. You can now write your own table.

rent.cut=cut(rent, breaks, right=FALSE)
rent.freq=table(rent.cut)

Output:
rent.cut

$[350, 665)$	$[665, 980)$	$[980, 1.3e+03)$	$[1.3e+03, 1.61e+03)$	$[1.61e+03, 1.92e+03)$	$[1.92e+03, 2.24e+03)$	$[2.24e+03, 2.56e+03)$	$[2.56e+03, 2.87e+03)$
4	8	5	6	0	0	1	0

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 just 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 \(\PageIndex{2}\): 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. On R, the command is

hist(variable, col="type in what color you want", breaks, main="type the title you want", xlab="type the label you want for the horizontal axis",
ylim=c(0, number above maximum frequency) – produces histogram with specified color and using the breaks you made for the frequency distribution.

For this example, the command in R would be (assuming you created a frequency distribution in R as described previously):

hist(rent, col="blue", breaks, right=FALSE, main="Monthly Rent Paid by Students", ylim=c(0,8) xlab="Monthly Rent ($)")

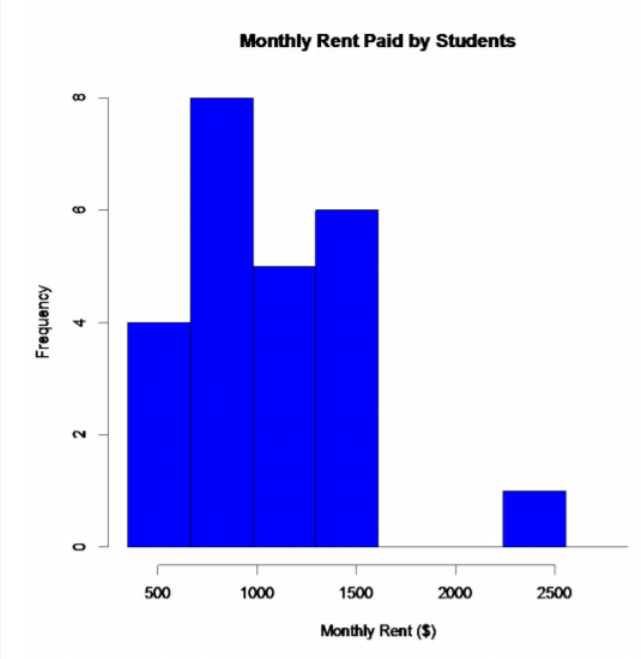


Figure 2.2.1: Histogram for Monthly Rent

If no frequency distribution was created before the histogram, then the command would be:

> hist(variable, col="type in what color you want", number of classes, main="type the title you want", xlab="type the label you want for the horizontal axis") – produces histogram with specified color and number of classes (though the number of classes is an estimate and R will create the number of classes near this value).

For this example, the R command without a frequency distribution created first would be:

> hist(rent, col="blue", 7, main="Monthly Rent Paid by Students", xlab="Monthly Rent ($)")

Notice the graph has the axes labeled, the tick marks are labeled on each axis, and there is a title.

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. There are other aspects that 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.

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 an 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 2.2.5 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 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

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. This is summarized in Example 2.2.4.

To produce cumulative frequencies in R, you need to have performed the commands for the frequency distribution. Once you have complete that, then use variable.cumfreq=cumsum(variable.freq) – creates the cumulative frequencies for the variable
cumfreq0=c(0,variable.cumfreq) – creates a cumulative frequency table for the variable.
cumfreq0 – displays the cumulative frequency table.

For this example the command would be:
rent.cumfreq=cumsum(rent.freq)

cumfreq0=c(0,rent.cumfreq)
cumfreq0

Output:

$$[350, 665, 980, 1.3e+03) \; [665, 980, 1.3e+03) \; [980, 1.3e+03) \; [1.3e+03, 1.61e+03) \; [1.61e+03, 1.92e+03) \; [1.92e+03, 2.22e+03) \; [2.22e+03, 2.56e+03) \; [2.56e+03, 2.87e+03)$$
$$0 \quad 4 \quad 12 \quad 17 \quad 23 \quad 23 \quad 23 \quad 24 \quad 24$$

Now type this into a table. See Example 2.2.4.

Table 2.2.4: Cumulative Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency	Cumulative Frequency
350-664	349.5-664.5	507	4	4
665-979	664.5-979.5	822	8	12
980-1294	979.5-1294.5	1127	5	17
1295-1609	1294.5-1609.5	1452	6	23
1610-1924	1609.5-1924.5	1767	0	23
1925-2239	1924.5-2239.5	2082	0	23
2240-2554	2239.5-2554.5	2397	1	24

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 horizontal and vertical axes that will fit the data. Then plot the points of the class upper class boundary versus the cumulative frequency. Make sure you include the point with the lowest class boundary and the 0 cumulative frequency. Then just connect the dots.

Example 2.2.6 drawing an ogive

Draw an ogive for the data in Example 2.2.1.

Solution

In R, the commands would be:
plot(breaks,cumfreq0, main="title you want to use", xlab="label you want to use", ylab="label you want to use", ylim=c(0, number above maximum cumulative frequency) – plots the ogive
lines(breaks,cumfreq0) – connects the dots on the ogive

For this example, the commands would be:
Plot(breaks,cumfreq0, main=”Cumulative Frequency for Monthly Rent”, xlab=”Monthly Rent ($)”, ylab=”Cumulative Frequency”, ylim=c(0,25))
lines(breaks,cumfreq0)

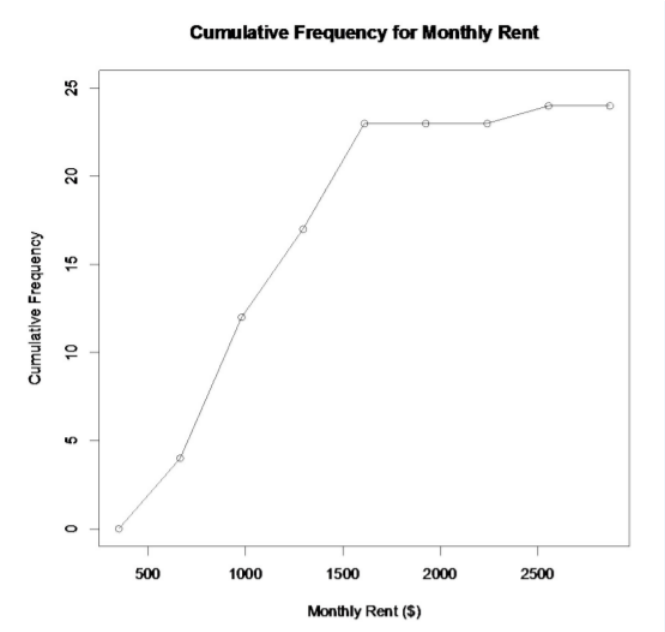


Figure 2.2.3: Ogive for Monthly Rent

The usefulness of a ogive is to allow the reader to find out how many students pay less than a certain value, and also what amount of monthly rent is paid by a certain number of students. As an example, suppose you want to know how many students pay less than $1500 a month in rent, then you can go up from the $1500 until you hit the graph and then you go over to the cumulative frequency axes to see what value corresponds to this value. It appears that around 20 students pay less than $1500. (See *Graph 2.2.4.*)

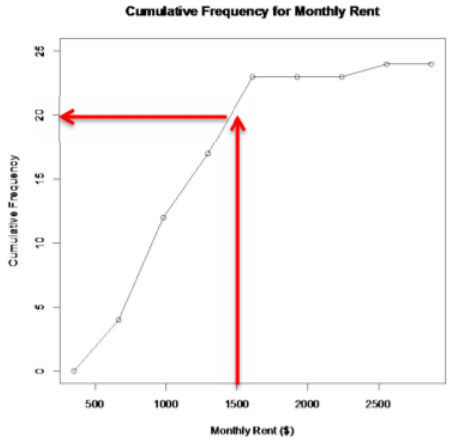


Figure 2.2.4: Ogive for Monthly Rent with Example

Also, if you want to know the amount that 15 students pay less than, then you start at 15 on the vertical axis and then go over to the graph and down to the horizontal axis where the line intersects the graph. You can see that 15 students pay less than about $1200 a month. (See *Graph 2.2.5.*)

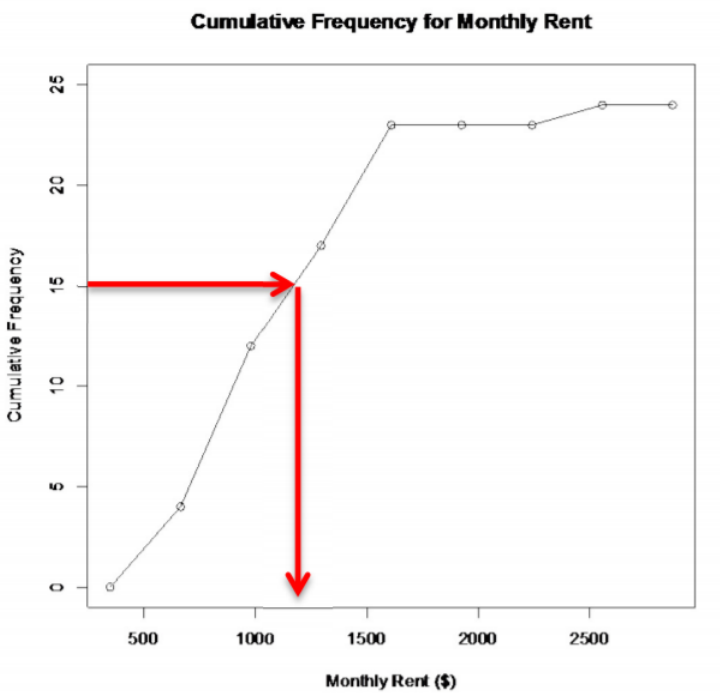


Figure 2.2.5: Ogive for Monthly Rent with Example

If you graph the cumulative relative frequency then you can find out what percentage is below a certain number instead of just the number of people below a certain value.

Shapes of the distribution:

When you look at a distribution, look at the basic shape. There are some basic shapes that 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.

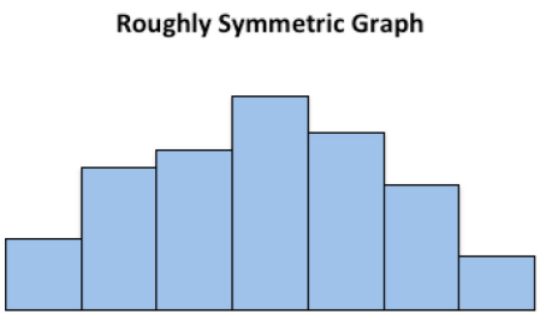
Symmetric means that you can fold the graph in half down the middle and the two sides will line up. You can think of the two sides as being mirror images of each other. Skewed means one “tail” of the graph is longer than the other. The graph is skewed in the direction of the longer tail (backwards from what you would expect). A uniform graph has all the bars 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 not longer of interest.

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

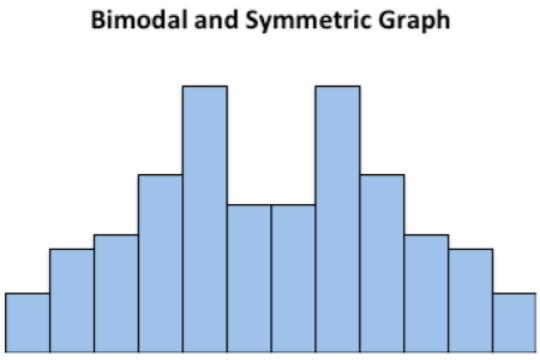
Examples of Graphs:

This graph is roughly symmetric and unimodal:



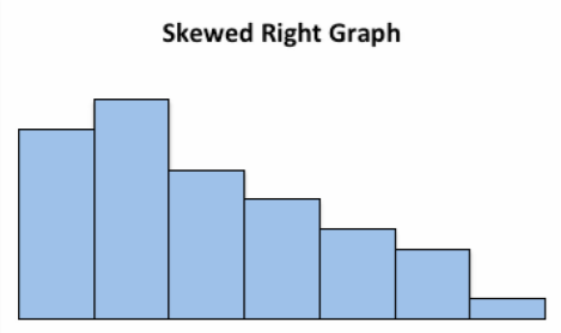
Figure

This graph is symmetric and bimodal:



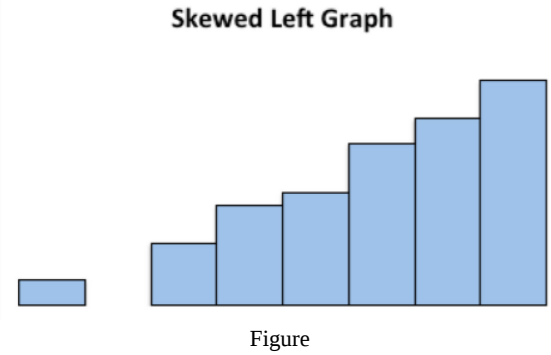
Figure

This graph is skewed to the right:



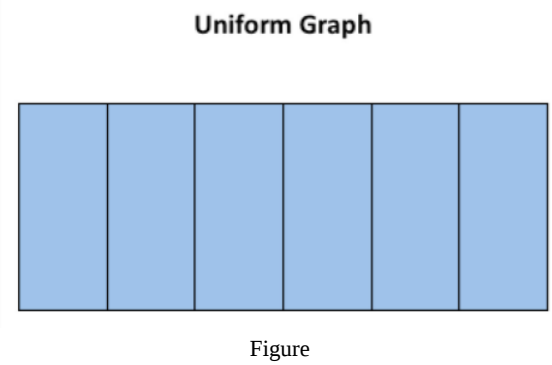
Figure

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



Figure

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



Figure

Example 2.2.7 creating a frequency distribution, histogram, and ogive

The following data represents the percent change in tuition levels at public, fouryear colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). Create a frequency distribution, histogram, and ogive for the data.

Table 2.2.5: Data of Tuition Levels at Public, Four-Year Colleges

19.5%	40.8%	57.0%	15.1%	17.4%	5.2%	13.0%
15.6%	51.5%	15.6%	14.5%	22.4%	19.5%	31.3%
21.7%	27.0%	13.1%	26.8%	24.3%	38.0%	21.1%
9.3%	46.7%	14.5%	78.4%	67.3%	21.1%	22.4%
5.3%	17.3%	17.5%	36.6%	72.0%	63.2%	15.1%
2.2%	17.5%	36.7%	2.8%	16.2%	20.5%	17.8%
30.1%	63.6%	17.8%	23.2%	25.3%	21.4%	28.5%
9.4%						

Solution

1. Find the range:
   largest value - smallest value = $78.4\% - 2.2\% = 76.2\%$
2. Pick the number of classes:
   Since there are 50 data points, then around 6 to 8 classes should be used. Let's use 8.
3. Find the class width:
   width $= \frac{\text{range}}{8} = \frac{76.2\%}{8} \approx 9.525\%$
   Since the data has one decimal place, then the class width should round to one decimal place. Make sure you round up.
   width $= 9.6\%$
4. Find the class limits:
   $2.2\% + 9.6\% = 11.8\%, 11.8\% + 9.6\% = 21.4\%, 21.4\% + 9.6\% = 31.0\%, \leftrightharpoons$
5. Find the class boundaries:
   Since the data has one decimal place, the class boundaries should have two decimal places, so subtract 0.05 from the lower class limit to get the class boundaries. Add 0.05 to the upper class limit for the last class's boundary.
   $2.2 - 0.05 = 2.15\%, 11.8 - 0.05 = 11.75\%, 21.4 - 0.05 = 21.35\% \leftrightharpoons$
   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:
   midpoint $= \frac{\text{lower limt } + \text{ upper limit}}{2}$
   $\frac{2.2 + 11.7}{2} = 6.95\%, \frac{11.8 + 21.3}{2} = 16.55\%, \leftrightharpoons$
7. Tally and find the frequency of the data:

Table 2.2.6: Frequency Distribution for Tuition Levels at Public, Four-Year Colleges

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency	Relative Frequency	Cumulative Frequency
2.2-11.7	2.15-11.75	6.95	𝍸 |	6	0.12	6
11.8-21.3	11.75-21.35	16.55	𝍸 𝍸 𝍸 𝍸	20	0.40	26
21.4-30.9	21.35-30.95	26.15	𝍸 𝍸 |	11	0.22	37
31.0-45.0	30.95-40.55	35.75	||||	4	0.08	41
40.6-50.1	40.55-50.15	45.35	||	2	0.04	43

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency	Relative Frequency	Cumulative Frequency
50.2-59.7	50.15-59.75	54.95	||	2	0.04	45
59.8-69.3	59.75-69.35	64.55	|||	3	0.06	48
69.4-78.9	69.35-78.95	74.15	||	2	0.04	50

Make sure the total of the frequencies is the same as the number of data points.

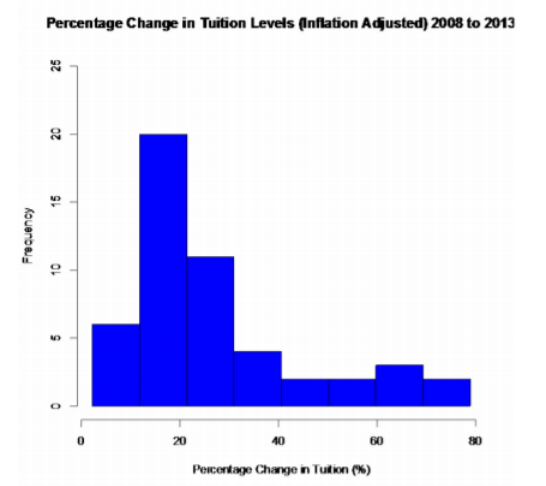


Figure 2.2.11: Histogram for Tuition Levels at Public, Four-Year Colleges

This graph is skewed right, with no gaps. This says that most percent increases in tuition were around 16.55%, with very few states having a percent increase greater than 45.35%.

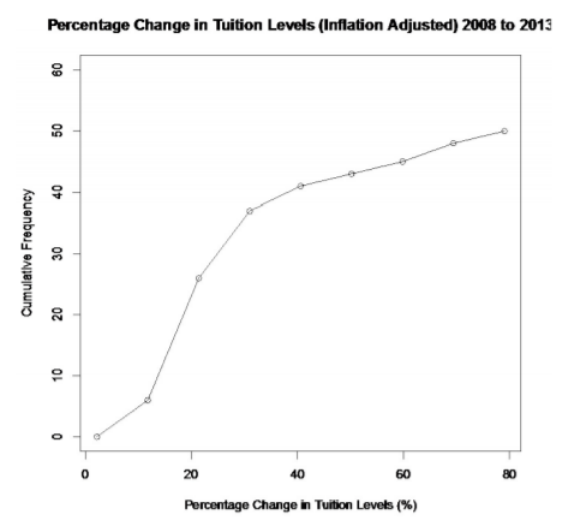


Figure 2.2.12: *Ogive for Tuition Levels at Public, Four-Year Colleges*

Looking at the ogive, you can see that 30 states had a percent change in tuition levels of about 25% or less.

There are occasions where the class limits in the frequency distribution are predetermined. Example 2.2.8 demonstrates this situation.

Example 2.2.8 creating a frequency distribution and histogram

The following are the percentage grades of 25 students from a statistics course. Make a frequency distribution and histogram.

Table 2.2.7: 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

Since this data is percent grades, it makes more sense to make the classes in multiples of 10, since grades are usually 90 to 100%, 80 to 90%, and so forth. It is easier to not use the class boundaries, but instead use the class limits and think of the upper class limit being up to but not including the next classes lower limit. As an example the class 80 – 90 means a grade of 80% up to but not including a 90%. A student with an 89.9% would be in the 80-90 class.

Table 2.2.8: Frequency Distribution for Test Grades

Class Limit	Class Midpoint	Tally	Freqeuncy
40-50	45	||	2
50-60	55	|	1
60-70	65	𝍸 ||	7
70-80	75	𝍸 |	6
80-90	85	𝍸 ||	7
90-100	95	||	2

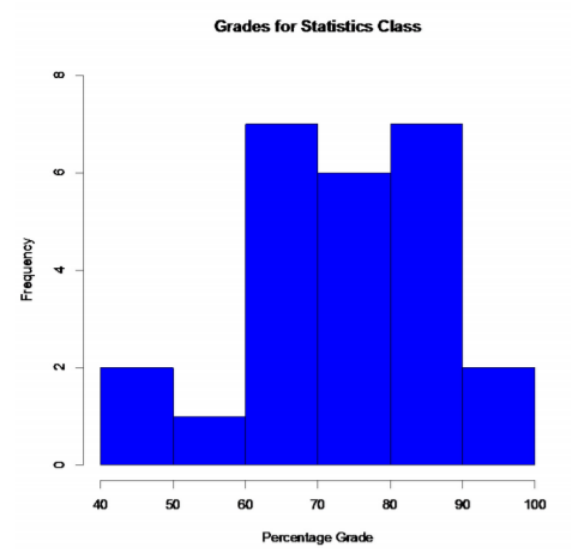


Figure 2.2.13: Histogram for Test Grades

It appears that most of the students had between 60 to 90%. This graph looks somewhat symmetric and also bimodal. The same number of students earned between 60 to 70% and 80 to 90%.

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

1. The median incomes of males in each state of the United States, including the District of Columbia and Puerto Rico, are given in Example 2.2.9 ("Median income of," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 7 classes.

Table 2.2.9: Data of Median Income for Males

$42,951	$52,379	$42,544	$37,488	$49,281	$50,987
$60,705	$50,411	$66,760	$40,951	$43,902	$45,494
$41,528	$50,746	$45,183	$43,624	$43,993	$41,612
$46,313	$43,944	$56,708	$60,264	$50,053	$50,580
$40,202	$43,146	$41,635	$42,182	$41,803	$53,033
$60,568	$41,037	$50,388	$41,950	$44,660	$46,176
$41,420	$45,976	$47,956	$22,529	$48,842	$41,464
$40,285	$41,309	$43,160	$47,573	$44,057	$52,805
$53,046	$42,125	$46,214	$51,630		

2. The median incomes of females in each state of the United States, including the District of Columbia and Puerto Rico, are given in Example 2.2.10 ("Median income of," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 7 classes.

Table 2.2.10: Data of Median Income for Females

$31,862	$40,550	$36,048	$30,752	$41,817	$40,236
$47,476	$40,500	$60,332	$33,823	$35,438	$37,242
$31,238	$39,150	$34,023	$33,745	$33,269	$32,684
$31,844	$34,599	$48,748	$46,185	$36,931	$40,416
$29,548	$33,865	$31,067	$33,424	$35,484	$41,021
$47,155	$32,316	$42,113	$33,459	$32,462	$35,746
$31,274	$36,027	$37,089	$22,117	$41,412	$31,330
$31,329	$33,184	$35,301	$32,843	$38,177	$40,969
$40,993	$29,688	$35,890	$34,381		

3. The density of people per square kilometer for African countries is in Example 2.2.11 ("Density of people," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 8 classes.

Table 2.2.11: Data of Density of People per Square Kilometer

15	16	81	3	62	367	42	123
8	9	337	12	29	70	39	83
26	51	79	6	157	105	42	45
72	72	37	4	36	134	12	3
630	563	72	29	3	13	176	341
415	187	65	194	75	16	41	18
69	49	103	65	143	2	18	31

4. The Affordable Care Act created a market place for individuals to purchase health care plans. In 2014, the premiums for a 27 year old for the bronze level health insurance are given in Example 2.2.12 ("Health insurance marketplace," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 5 classes.

Table 2.2.12: Data of Health Insurance Premiums

$114	$119	$121	$125	$132	$139
$139	$141	$143	$145	$151	$153
$156	$159	$162	$163	$165	$166
$170	$170	$176	$177	$181	$185
$185	$186	$186	$189	$190	$192
$196	$203	$204	$219	$254	$286

5. Create a histogram and relative frequency histogram for the data in Example 2.2.9. Describe the shape and any findings you can from the graph.

6. Create a histogram and relative frequency histogram for the data in Example 2.2.10. Describe the shape and any findings you can from the graph.

7. Create a histogram and relative frequency histogram for the data in Example 2.2.11. Describe the shape and any findings you can from the graph.

8. Create a histogram and relative frequency histogram for the data in Example 2.2.12. Describe the shape and any findings you can from the graph.

9. Create an ogive for the data in Example 2.2.9. Describe any findings you can from the graph.

10. Create an ogive for the data in Example 2.2.10. Describe any findings you can from the graph.

11. Create an ogive for the data in Example 2.2.11. Describe any findings you can from the graph.

12. Create an ogive for the data in Example 2.2.12. Describe any findings you can from the graph.

13. Students in a statistics class took their first test. The following are the scores they earned. Create a frequency distribution and histogram for the data using class limits that make sense for grade data. Describe the shape of the distribution.

Table 2.2.13: Data of Test 1 Grades

80	79	89	74	73	67	79
93	70	70	76	88	83	73
81	79	80	85	79	80	79

58	93	94	74			

14. Students in a statistics class took their first test. The following are the scores they earned. Create a frequency distribution and histogram for the data using class limits that make sense for grade data. Describe the shape of the distribution. Compare to the graph in question 13.

Table 2.2.14: Data of Test 1 Grades

67	67	76	47	85	70
87	76	80	72	84	98
84	64	65	82	81	81
88	74	87	83		

Answer

See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

2.3: Other Graphical Representations of Data

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

Stem Plots

Stem 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 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 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). The *Graph 2.3.1* shows the stems for this example.

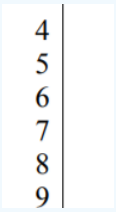


Figure 2.3.1: Stem Plot for Test Grades Step 1

Now go through the list of data and add the leaves. Put each leaf next to its corresponding stem. Don't worry about 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*.

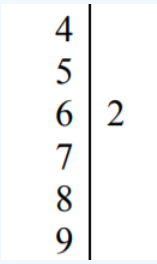


Figure 2.3.2: Stem Plot for Test Grades Step 2

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

4	
5	
6	2
7	
8	7
9	

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 = 40%
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.

You can create a stem and leaf plot on R. the command is:

stem(variable) – creates a stem and leaf plot, if you do not get a stem plot that shows all of the stems then use scale = a number. Adjust the number until you see all of the stems. So you would have stem(variable, scale = a number)

For Example 2.3.1, the command would be

grades<-c(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)
stem(grades, scale = 2)

Output:

The decimal point is 1 digit(s) to the right of the |

4 | 05
5 | 8
6 | 2224579
7 | 123667
8 | 0147779
9 | 35

Now just put a title on the stem plot.

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 axes. Once that is done, then just plot the x and y values as an ordered pair. In R, the command is:

independent variable<-c(type in data with commas in between values)
dependent variable<-c(type in data with commas in between values)
plot(independent variable, dependent variable, main="type in a title you want", xlab="type in a label for the horizontal axis", ylab="type in a label for the vertical axis", ylim=c(0, number above maximum y value)

For this example, that would be:
elevation<-c(7000, 4000, 6000, 3000, 7000, 4500, 5000)
temperature<-c(50, 60, 48, 70, 55, 55, 60)
plot(elevation, temperature, main="Temperature versus Elevation", xlab="Elevation (in feet)", ylab="Temperature (in degrees F)", ylim=c(0, 80))

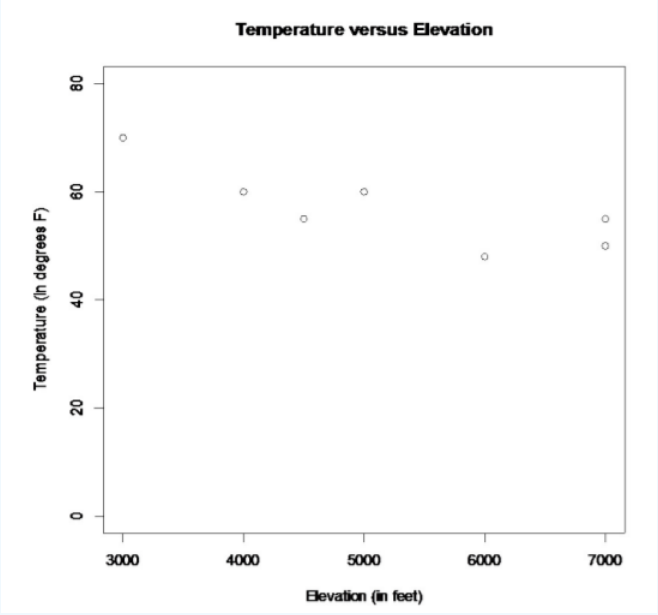


Figure 2.3.6: Scatter Plot of Temperature versus Elevation

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

A time-series plot is a graph showing the data measurements in chronological order, the data being quantitative data. 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, prior to 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

In R, the command would be:

variable1<-c(type in data with commas in between values, this should be the time variable)
variable2<-c(type in data with commas in between values)
plot(variable1, variable2, ylim=c(0,number over max), main="type in a title you want", xlab="type in a label for the horizontal axis", ylab="type in a label for the vertical axis")
lines(variable1, variable2) – connects the dots

For this example:
time<-c(0, 1, 2, 3, 4, 5)

```
weight<-c(200, 195, 192, 193, 190, 187)
plot(time, weight, ylim=c(0,250), main="Weight over Time", xlab="Time (Months) ", ylab="Weight (pounds)")
ines(time, weight)
```

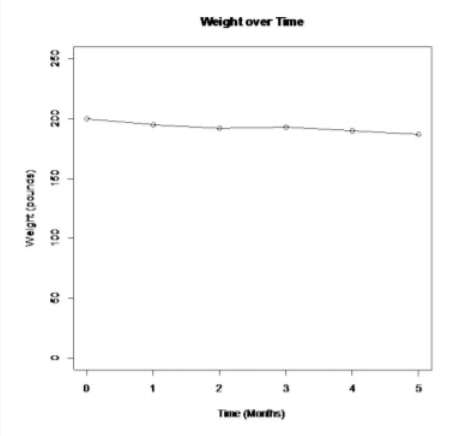
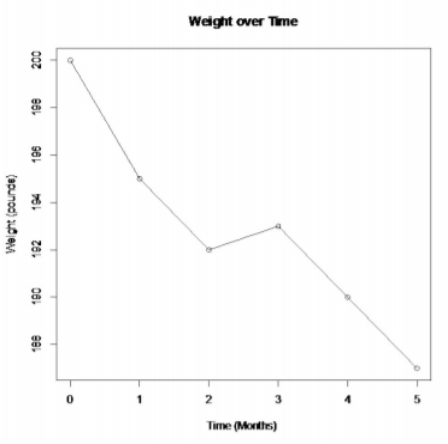


Figure of Weight versus Time

Notice, that over the 5 months, the weight appears to be decreasing. Though 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 the *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

Homework

1. Students in a statistics class took their first test. The data in Example 2.3.4 are the scores they earned. Create a stem plot.

Table 2.3.4: Data of Test 1 Grades

80	79	89	74	73	67	79
93	70	70	76	88	83	73
81	79	80	85	79	80	79
58	93	94	74			

2. Students in a statistics class took their first test. The data in Example 2.3.5 are the scores they earned. Create a stem plot. Compare to the graph in question 1.

Table 2.3.5: Data of Test 1 Grades

67	67	76	47	85	70
87	76	80	72	84	98
84	64	65	82	81	81
88	74	87	83		

3. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of one of their metacarpal bone (in cm) were collected and are in Example 2.3.6 ("Prediction of height," 2013). Create a scatter plot and state if there is a relationship between the height of a person and the length of their metacarpal.

Table 2.3.6: Data of Metacarpal versus Height

Length of Metacarpal	Height of Person
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

4. Example 2.3.7 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Create a scatter plot and state if there is a relationship between the value of the house and the annual rental income.

Table 2.3.7: Data of House Value versus Rental

Value	Rental	Value	Rental	Value	Rental	Value	Rental
81000	6656	77000	4576	75000	7280	67500	6864
95000	7904	94000	8736	90000	6240	85000	7072
121000	12064	115000	7904	110000	7072	104000	7904
135000	8320	130000	9776	126000	6240	125000	7904
145000	8320	140000	9568	140000	9152	135000	7488
165000	13312	165000	8528	155000	7488	148000	8320
178000	11856	174000	10400	170000	9568	170000	12688
200000	12272	200000	10608	194000	11232	190000	8320
214000	8528	280000	10400	200000	10400	200000	8320
240000	10192	240000	12064	240000	11648	225000	12480
289000	11648	270000	12896	262000	10192	244500	11232
325000	12480	310000	12480	303000	12272	300000	12480

5. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example 2.3.8. Create a scatter plot of the data and state if there appears to be a relationship between life expectancy and the number of births per woman.

Table 2.3.8: Data of Life Expectancy versus Fertility Rate

Life Expectancy	Fertility Rate	Life Expectancy	Fertility rate
77.2	1.7	72.3	3.9
55.4	5.8	76.0	1.5
69.9	2.2	66.0	4.2
76.4	2.1	5.9	5.2
75.0	1.8	54.4	6.8
78.2	2.0	62.9	4.7
73.0	2.6	78.3	2.1
70.8	2.8	72.1	2.9
82.6	1.4	80.7	1.4
68.9	2.6	74.2	2.5
81.0	1.5	73.3	1.5
54.2	6.9	67.1	2.4

6. The World Bank collected data on the percentage of gross domestic product (GDP) that a country spends on health expenditures ("Health expenditure," 2013) and the percentage of woman receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 is in Example 2.3.9. Create a scatter plot of the data and state if there appears to be a relationship between percentage spent on health expenditure and the percentage of woman receiving prenatal care.

Table 2.3.9: Data of Prenatal Care versus Health Expenditure

Prenatal Care (%)	Health Expenditure (% of GDP)
47.9	9.6
54.6	3.7
93.7	5.2
84.7	5.2
100.0	10.0
42.5	4.7
96.4	4.8
77.1	6.0
58.3	5.4
95.4	4.8
78.0	4.1
93.3	6.0
93.3	9.5

Prenatal Care (%)	Health Expenditure (% of GDP)
93.7	6.8
89.8	6.1

7. The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997 ("Deaths from firearms," 2013). The data is in Example 2.3.10. Create a time-series plot of the data and state any findings you can from the graph.

Table 2.3.10: Data of Year versus Number of Deaths due to Firearms

Year	1983	1984	1985	1986	1987	1988	1989	1990
Rate	4.31	4.42	4.52	4.35	4.39	4.21	3.40	3.61
Year	1991	1992	1993	1994	1995	1996	1997	
Rate	3.67	3.61	2.98	2.95	2.72	2.95	2.3	

8. The economic crisis of 2008 affected many countries, though some more than others. Some people in Australia have claimed that Australia wasn't hurt that badly from the crisis. The bank assets (in billions of Australia dollars (AUD)) of the Reserve Bank of Australia (RBA) for the time period of March 2007 through March 2013 are contained in Example 2.3.11 ("B1 assets of," 2013). Create a time-series plot and interpret any findings.

Table 2.3.11: Data of Date versus RBA Assets

Date	Assets in Billions of AUD
Mar-2006	96.9
Jun-2006	107.4
Sep-2006	107.2
Dec-2006	116.2
Mar-2007	123.7
Jun-2007	134.0
Sep-2007	123.0
Dec-2007	93.2
Mar-2008	93.7
Jun-2008	105.6
Sep-2008	101.5
Dec-2008	158.8
Mar-2009	118.7
Jun-2009	111.9
Sep-2009	87.0
Dec-2009	86.1
Mar-2010	83.4
Jun-2010	85.7
Sep-2010	74.8
Dec-2010	76.0

Date	Assets in Billions of AUD
Mar-2011	75.7
Jun-2011	75.9
Sep-2011	75.2
Dec-2011	87.9
Mar-2012	91.0
Jun-2012	90.1
Sep-2012	83.9
Dec-2012	95.8
Mar-2013	90.5

9. The consumer price index (CPI) is a measure used by the U.S. government to describe the cost of living. Example 2.3.12 gives the cost of living for the U.S. from the years 1947 through 2011, with the year 1977 being used as the year that all others are compared (DeNavas-Walt, Proctor & Smith, 2012). Create a time-series plot and interpret.

Table 2.3.12: Data of Time versus CPI

Year	CPI-U-RS1 index (December 1977=100)	Year	CPI-U-RS1 index (December 1977=100)
1947	37.5	1980	127.1
1948	40.5	1981	139.2
1949	40.0	1982	147.6
1950	40.5	1983	153.9
1951	43.7	1984	160.2
1952	44.5	1985	165.7
1953	44.8	1986	168.7
1954	45.2	1987	174.4
1955	45.0	1988	180.8
1956	45.7	1989	188.6
1957	47.2	1990	198.0
1958	48.5	1991	205.1
1959	48.9	1992	210.3
1960	49.7	1993	215.5
1961	50.2	1994	220.1
1962	50.7	1995	225.4
1963	51.4	1996	231.4
1964	52.1	1997	236.4
1965	52.9	1998	239.7
1966	54.4	1999	244.7

Year	CPI-U-RS1 index (December 1977=100)	Year	CPI-U-RS1 index (December 1977=100)
1967	56.1	2000	252.9
1968	58.3	2001	260.0
1969	60.9	2002	264.2
1970	63.9	2003	270.1
1971	66.7	2004	277.4
1972	68.7	2005	286.7
1973	73.0	2006	296.1
1974	80.3	2007	304.5
1975	86.9	2008	316.2
1976	91.9	2009	315.0
1977	97.7	2010	320.2
1978	104.4	2011	330.3
1979	114.4		

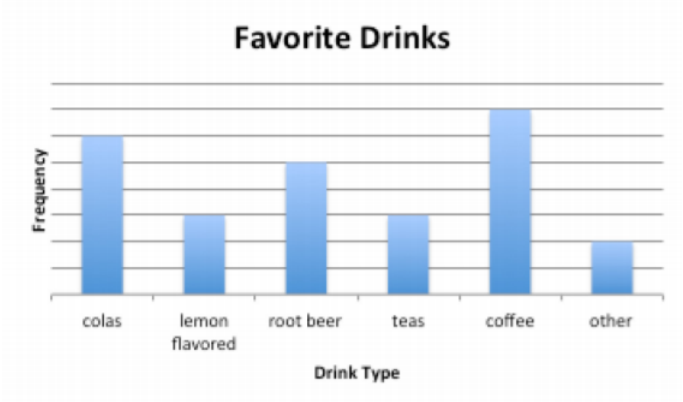
10. The median incomes for all households in the U.S. for the years 1967 to 2011 are given in Example 2.3.13 (DeNavas-Walt, Proctor & Smith, 2012). Create a time-series plot and interpret.

Table 2.3.13: Data of Time versus Median Income

Year	Median Income	Year	Median Income
1967	42,056	1990	49,950
1968	43,868	1991	48,516
1969	45,499	1992	48,117
1970	45,146	1993	47,884
1971	44,707	1994	48,418
1972	46,622	1995	49,935
1973	47,563	1996	50,661
1974	46,057	1997	51,704
1975	44,851	1998	53,582
1976	45,595	1999	54,932
1977	45,884	2000	54,841
1978	47,659	2001	53,646
1979	47,527	2002	53,019
1980	46,024	2003	52,973
1981	45,260	2004	52,788
1982	45,139	2005	53,371
1983	44,823	2006	53,768

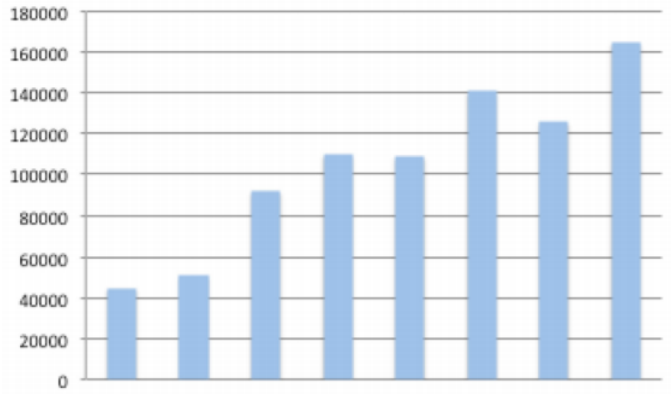
Year	Median Income	Year	Median Income
1984	46,215	2007	54,489
1985	47,079	2008	52,546
1986	48,746	2009	52,195
1987	49,358	2010	50,831
1988	49,737	2011	50,054
1989	50,624		

11. State everything that makes *Graph 2.3.9* a misleading or poor graph.



***Graph 2.3.9**: Example of a Poor Graph*

12. State everything that makes *Graph 2.3.10* a misleading or poor graph (Benen, 2011).



***Graph 2.3.10**: Example of a Poor Graph*

13. State everything that makes *Graph 2.3.11* a misleading or poor graph ("United States unemployment," 2013).



***Graph 2.3.11**: Example of a Poor Graph*

14. State everything that makes *Graph 2.3.12* a misleading or poor graph.



***Graph 2.3.12**: Example of a Poor Graph*

Answer

See solutions

Data Sources:

B1 assets of financial institutions. (2013, June 27). Retrieved from www.rba.gov.au/statistics/tables/xls/b01hist.xls

Benen, S. (2011, September 02). [Web log message]. Retrieved from http://www.washingtonmonthly.com/pol...edit031960.php

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/rentcap.html

Contraceptive use. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...gs.aspx?ind=35

Deaths from firearms. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/firearms.html

DeNavas-Walt, C., Proctor, B., & Smith, J. U.S. Department of Commerce, U.S. Census Bureau. (2012). *Income, poverty, and health insurance coverage in the United States: 2011* (P60-243). Retrieved from website: www.census.gov/prod/2012pubs/p60-243.pdf

Density of people in Africa. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...249,250,251,252,253,254,34227,255,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,274,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,294,295,296,297,298,299,300,301,302,304,305,306,307,308

Department of Health and Human Services, ASPE. (2013). *Health insurance marketplace premiums for 2014*. Retrieved from website: aspe.hhs.gov/health/reports/2...b_premiumsland scape.pdf

Electricity usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=162

Fertility rate. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Fuel oil usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=164

Gas usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=165

Health expenditure. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS Hinatov, M. U.S. Consumer Product Safety Commission, Directorate of Epidemiology. (2012). *Incidents, deaths, and in-depth investigations associated with non-fire carbon monoxide from engine-driven generators and other engine-driven tools, 1999-2011*. Retrieved from website: www.cpsc.gov/PageFiles/129857/cogenerators.pdf

Life expectancy at birth. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=137

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=136

Prediction of height from metacarpal bone length. (2013, September 26). Retrieved from http://www.statsci.org/data/general/stature.html

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

United States unemployment. (2013, October 14). Retrieved from http://www.tradingeconomics.com/unit...mployment-rate

Weissmann, J. (2013, March 20). A truly devastating graph on state higher education spending. *The Atlantic*. Retrieved from http://www.theatlantic.com/business/...ending/274199/

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

3: Examining the Evidence Using Graphs and Statistics

Chapter 1 discussed what a population, sample, parameter, and statistic are, and how to take different types of samples. Chapter 2 discussed ways to graphically display data. There was also a discussion of important characteristics: center, variations, distribution, outliers, and changing characteristics of the data over time. Distributions and outliers can be answered using graphical means. Finding the center and variation can be done using numerical methods that will be discussed in this chapter. Both graphical and numerical methods are part of a branch of statistics known as **descriptive statistics**. Later descriptive statistics will be used to make decisions and/or estimate population parameters using methods that are part of the branch called **inferential statistics**.

3.1: Measures of Center

3.2: Measures of Spread

3.3: Ranking

This page titled 3: Examining the Evidence Using Graphs and Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

3.1: Measures of Center

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 **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 **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 **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.

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 actually 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, and mode

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, and Mode

6.8	8.2	7.5	9.4	8.2

Find the mean, median, and mode 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 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.

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 number. 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.

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

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 ($\bar{x}$) 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).

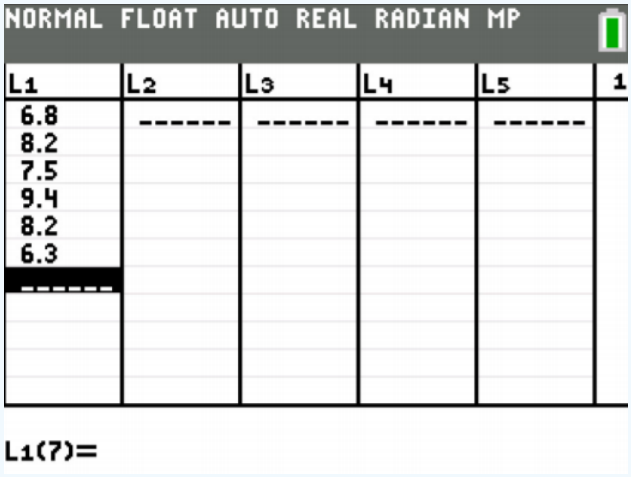


Figure 3.1.1: TI-83/84 Calculator Edit Setup

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.

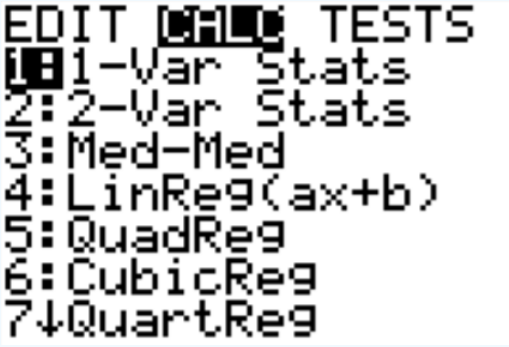


Figure 3.1.2: TI-83/84 Calculator CALC Menu

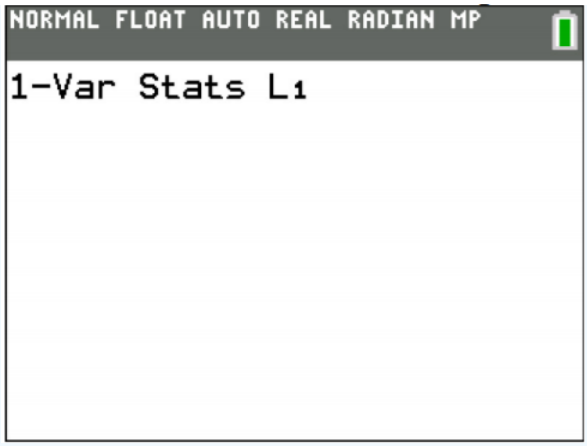


Figure 3.1.3: TI-83/84 Calculator Input for Example 3.1.3 Variable

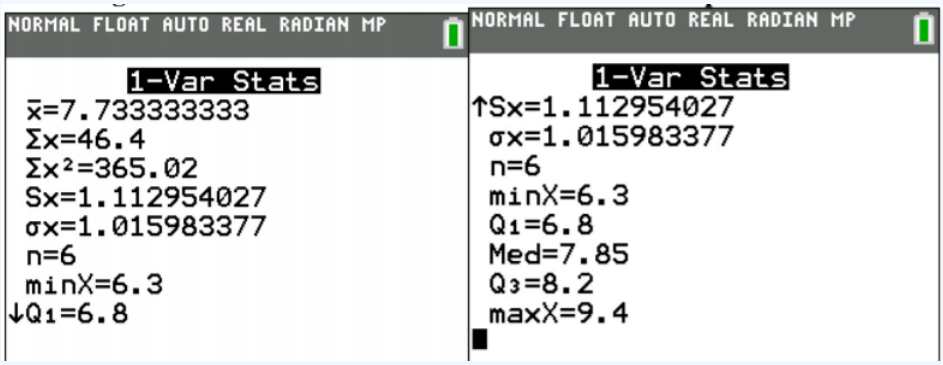


Figure 3.1.4: TI-83/84 Calculator Results for Example 3.1.3 Variable

The commands for finding the mean and median using R are as follows:

variable<-c(type in your data with commas in between)
To find the mean, use mean(variable)
To find the median, use median(variable)

So for this example, the commands would be

weights<-c(6.8, 8.2, 7.5, 9.4, 8.2, 6.3)
mean(weights)
[1] 7.733333
median(weights)
[1] 7.85

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

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

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

$$\text{median} = \frac{8.2 + 8.2}{2} = 8.2 \text{ pounds}$$

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 very heavy 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 an outlier if the extreme value is far enough from the center. In Example 3.1.4, 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 the mode are resistant measures because they are not affected by extreme values.

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 and it talks about the "average" it usually means the mean 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 because 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 related to skewed or symmetric distributions, see *Figure* 3.1.5. 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 values.

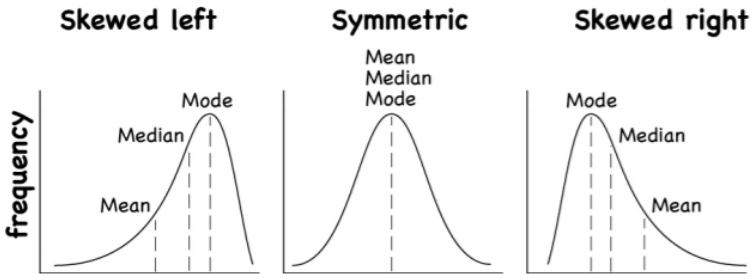


Figure 3.1.5: Mean, Median, Mode as Related to a Distribution

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 fulltime 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.

Definition 3.1.3

Weighted Average

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

Example 3.1.5 weighted average

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 average for the course.

Solution

Variable: $x =$ score

The weighted average is $\frac{\sum xw}{\sum w} = \frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$

$$\text{weighted average} = \frac{95(0.15) + 83(0.25) + 76(0.25) + 84(0.35)}{0.15 + 0.25 + 0.25 + 0.35} = \frac{83.4}{1.00} = 83.4\%$$

A weighted average 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).

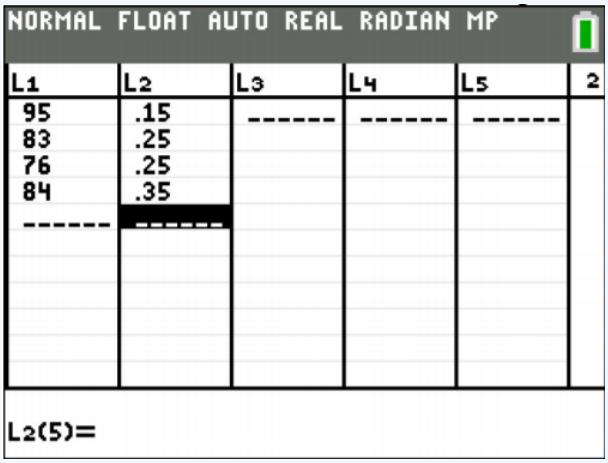


Figure 3.1.6: TI-3/84 Calculator Edit Setup

Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see *Figure* 3.1.7). 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 average.

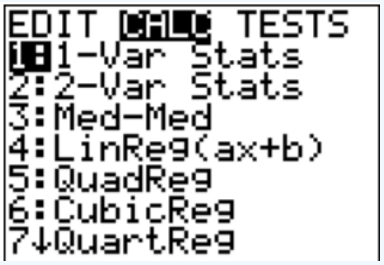


Figure 3.1.7: TI-83/84 Calculator CALC Menu

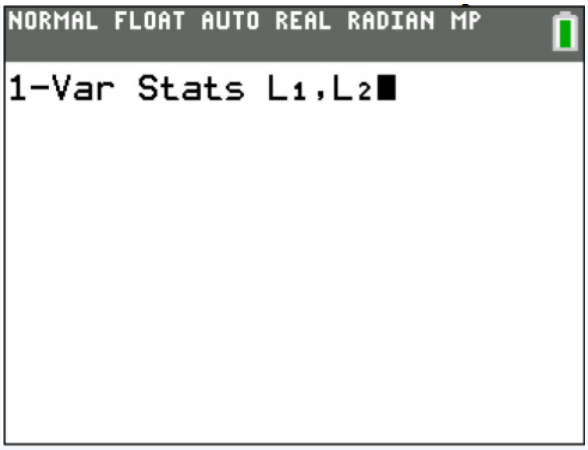


Figure 3.1.8: TI-83/84 Calculator Input for Weighted Average

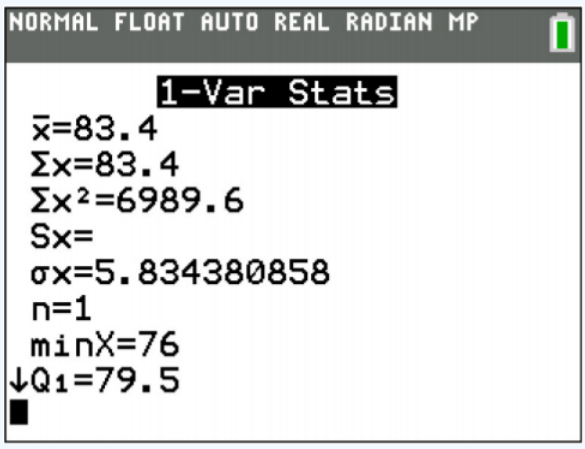


Figure 3.1.9: TI-83/84 Calculator Results for Weighted Average

The commands for finding the mean and median using R are as follows:

x<-c(type in your data with commas in between)
w<-c(type in your weights with commas in between
weighted.mean(x,w)

So for this example, the commands would be

x<-c(95, 83, 76, 84)
w<-c(.15, .25, .25, .35)
weighted.mean(x,w)
[1] 83.4

Example 3.1.6 weighted average

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 in a timely manner. The process involves reviewing student evaluations, peer evaluations, and supervisor evaluation for each teacher and awarding him/her 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 average of the evaluation.
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 average of the evaluation.
c. Which faculty member had the higher average evaluation?

Solution

a. Variable: $x =$ rating

The weighted average is $\frac{\sum xw}{\sum w} = \frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$

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

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

c. The second faculty member has a higher average evaluation.

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

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.

Homework

Exercise 3.1.1

1. Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in Example 3.1.7. Find the mean, median, and mode.

Table 3.1.7: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

2. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in Example 3.1.8 (Lee, 1994). Find the mean, median, and mode.

Table 3.1.8: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	138	Kakanui	64
Waipara	64	Rangitata	121
Ashley	97	Ophi	80
Waimakariri	161	Pareora	56
Selwyn	95	Waihao	64
Rakaia	145	Waitaki	209
Ashburton	90		

3. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed in Example 3.1.9 (Lee, 1994). Find the mean, median, and mode.

Table 3.1.9: Lengths of Rivers (km) Flowing to Tasman Sea

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

4. Eyeglassmatic manufactures eyeglasses for their retailers. They research to see how many defective lenses they made during the time period of January 1 to March 31. Example 3.1.10 contains the defect and the number of defects. Find the mean, median, and mode.

Table 3.1.10: Number of Defective Lenses

Defect Type	Number of Defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

5. Print-O-Matic printing company's employees have salaries that are contained in Example 3.1.11.

Employee	Salary ($)
CEO	272,500
Driver	58,456
CD74	100,702

Employee	Salary ($)
CD65	57,380
Embellisher	73,877
Folder	65,270
GTO	74,235
Handwork	52,718
Horizon	76,029
ITEK	64,553
Mgmt	108,448
Platens	69,573
Polar	75,526
Pre Press Manager	108,448
Pre Press Manager/ IT	98,837
Pre Press/ Graphic Artist	75,311
Designer	90,090
Sales	109,739
Administration	66,346

Table 3.1.11*: Salaries of Print-O-Matic Printing Company Employees*

a. Find the mean and median.

b. Find the mean and median with the CEO's salary removed.

c. What happened to the mean and median when the CEO's salary was removed? Why?

d. If you were the CEO, who is answering concerns from the union that employees are underpaid, which average of the complete data set would you prefer? Why?

e. If you were a platen worker, who believes that the employees need a raise, which average would you prefer? Why?

6. Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in Example 3.1.12.

Monthly charges	Monthly cost ($)
Bank charges	482
Cleaning	2208
Computer expensive	2471
Lease payments	2656
Postage	2117
Uniforms	2600

Table 3.1.12*: Fixed Costs for Print-O-Matic Printing Company*

a. Find the mean and median.

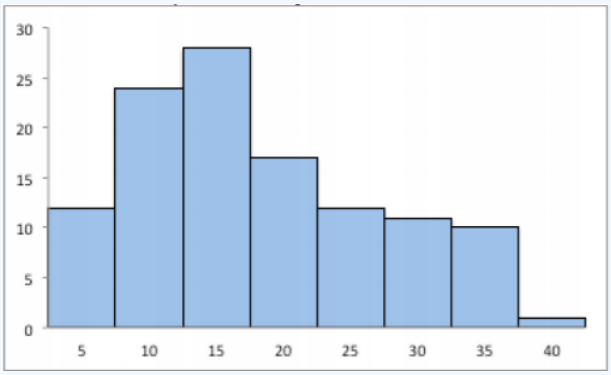
b. Find the mean and median with the bank charger removed.

c. What happened to the mean and median when the bank charger was removed? Why?

d. If it is your job to oversee the fixed costs, which average using te complete data set would you prefer to use when submitting a report to administration to show that costs are low? Why?

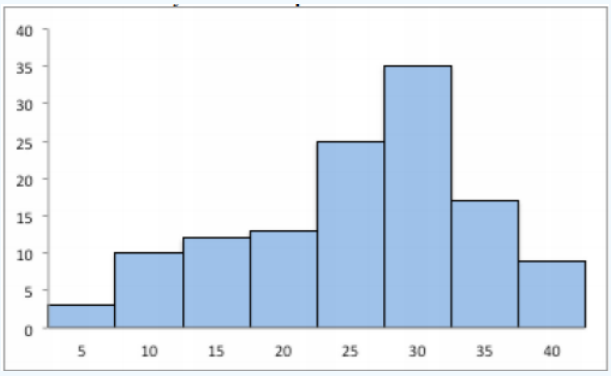
e. If it is your job to find places in the budget to reduce costs, which average using the complete data set would you prefer to use when submitting a report to administration to show that fixed costs need to be reduced? Why?

7. State which type of measurement scale each represents, and then which center measures can be use for the variable?
   a. You collect data on people's likelihood (very likely, likely, neutral, unlikely, very unlikely) to vote for a candidate.
   b. You collect data on the diameter at breast height of trees in the Coconino National Forest.
   c. You collect data on the year wineries were started.
   d. You collect the drink types that people in Sydney, Australia drink.
8. State which type of measurement scale each represents, and then which center measures can be use for the variable?
   a. You collect data on the height of plants using a new fertilizer.
   b. You collect data on the cars that people drive in Campbelltown, Australia.
   c. You collect data on the temperature at different locations in Antarctica.
   d. You collect data on the first, second, and third winner in a beer competition.
9. Looking at *Graph 3.1.1*, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?



***Graph 3.1.1**: Skewed or Symmetric Graph*

10. Looking at *Graph 3.1.2*, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?



***Graph 3.1.2**: Skewed or Symmetric Graph*

11. An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward the goals, job effectiveness, competencies, and CCC core values. Suppose for a specific employee, goal 1 has a weight of 30%, goal 2 has a weight of 20%, job effectiveness has a weight of 25%, competency 1 has a goal of 4%, competency 2 has a goal has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 3%, competency 5 has a weight of 2%, and core values has a weight of 10%. Suppose the employee has scores of 3.0 for goal 1, 3.0 for goal 2, 2.0 for job effectiveness, 3.0 for competency 1, 2.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, 4.0 for competency 5, and 3.0 for core values. Find the weighted average score for this employee. If an employee has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
12. An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward goals, job effectiveness, competencies, CCC core values. Suppose for a specific employee, goal 1 has a weight of 20%, goal 2 has a weight of 20%, goal 3 has a weight of 10%, job effectiveness has a weight of 25%, competency 1 has a goal of

4%, competency 2 has a goal has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 5%, and core values has a weight of 10%. Suppose the employee has scores of 2.0 for goal 1, 2.0 for goal 2, 4.0 for goal 3, 3.0 for job effectiveness, 2.0 for competency 1, 3.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, and 4.0 for core values. Find the weighted average score for this employee. If an employee that has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?

13. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives an 85 on test 1, a 76 on test 2, an 83 on test 3, a 74 on the homework, a 65 on the project, and a 79 on the final, what grade did the student earn in the course?
14. A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives a 92 on test 1, an 85 on test 2, a 95 on test 3, a 92 on the homework, a 55 on the project, and an 83 on the final, what grade did the student earn in the course?

Answer

1. mean = 253.93, median = 268, mode = none

3. mean = 67.68 km, median = 64 km, mode = 56 and 64 km

5. a. mean = $89,370.42, median = $75,311, b. mean = $79,196.56, median = $74,773, c. See solutions, d. See solutions, e. See solutions

7. a. ordinal- median and mode, b. ratio – all three, c. interval – all three, d. nominal – mode

9. Skewed right, mean higher

11. 2.71

13. 76.75

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

3.2: Measures of Spread

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 gives 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 3.2.1, 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 really spread out? What are the highest weight and the lowest weight? All you know is that the center of the weights is 8.02 pounds. You need more information.

Definition 3.2.1

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

$$\begin{aligned}\text{Range} &= \text{highest value} - \text{lowest value} \\ &= \text{maximum value} - \text{minimum value}\end{aligned}$$

Example 3.2.1: 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.

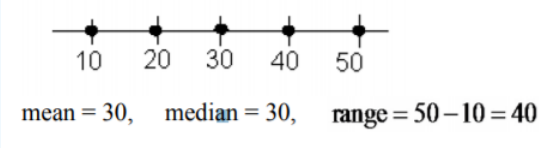


Figure 3.2.1: Dot Plot for Example 3.2.1a

b.

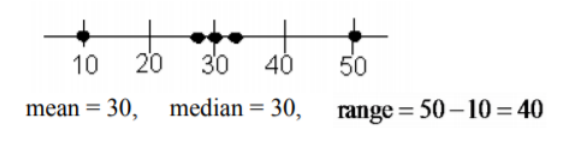


Figure 3.2.2: Dot Plot for Example 3.2.1b

c.

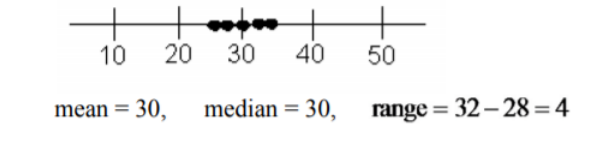


Figure 3.2.3: Dot Plot for Example 3.2.1

Based on the mean, median, and range in Example 3.2.1, the first two distributions are the same, but you can see from the graphs that they are different. In Example 3.2.1a the data are spread out equally. In Example 3.2.1b the data has a clump in the middle and a single value at each end. The mean and median are the same for Example 3.2.1c but the range is very different. All the data is clumped together in the middle.

The range doesn't really 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**.

Example 3.2.2: 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 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 3.2.1: 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
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 3.2.2: 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 because 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 3.2.3: 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 values you divide by the number of data values minus 1. In this case 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 3.2.2: 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 3.2.3: 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. Basically, it makes the sample standard deviation a better approximation of the population standard deviation.

Definition 3.2.4: 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 3.2.5: 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 3.2.3: Finding the Standard Deviation

Suppose that a manager wants to test two new training programs. He randomly selects 5 people for each training type and measures the time it takes to complete a task after the training. The times for both trainings are in Example 3.2.4. Which training method is better?

Table 3.2.4: 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 better, first you need some descriptive statistics. Start with the mean for each sample.

$$\overline{x}_1 = \frac{56 + 75 + 48 + 63 + 59}{5} = 60.2 \text{ minutes}$$

$$\overline{x}_2 = \frac{60 + 58 + 66 + 59 + 58}{5} = 60.2 \text{ minutes}$$

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

Table 3.2.5: Squared Deviations for Training 1

x_1	$x_1 - \overline{x}_1$	$(x_1 - \overline{x}_1)^2$
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 3.2.6: Squared Deviations for Training 2

x_2	$x_2 - \overline{x}_2$	$(x_2 - \overline{x}_2)^2$
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 sample is:

$$s_1^2 = \frac{394.8}{5 - 1} = 98.7 \text{ minutes}^2$$

$$s_2^2 = \frac{44.8}{5 - 1} = 11.2 \text{ minutes}^2$$

The standard deviations are:

$$s_1 = \sqrt{98.7} \approx 9.93 \text{ minutes}$$

$$s_2 = \sqrt{11.2} \approx 3.35 \text{ minutes}$$

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 3.2.3 on the TI-83/84 is in *Figures 3.2.1* through *3.2.4* (the procedure is the same for X_1). Note 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 sample. In almost all cases you have a sample and will be using s_x. Also, the calculator uses the notation s_x of 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* 3.2.1).

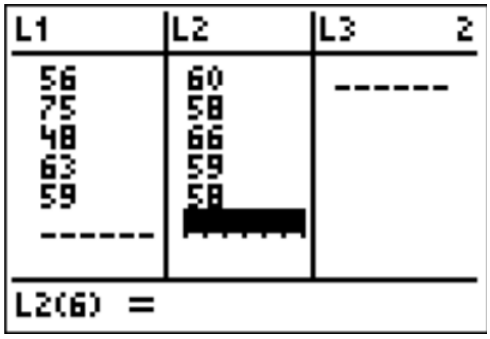


Figure 3.2.1: TI-83/84 Calculator Edit Setup

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* 3.2.2). 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* 3.2.4.

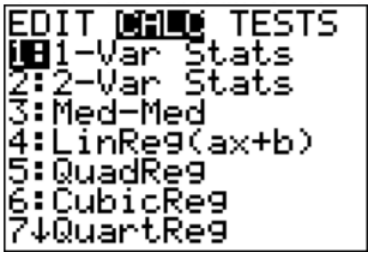


Figure 3.2.2: TI-83/84 Calculator CALC Menu

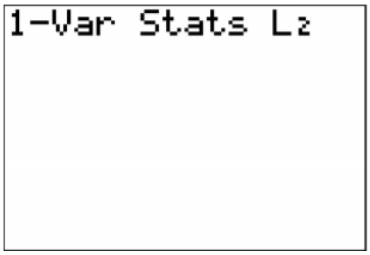


Figure 3.2.3: TI-83/84 Calculator Input for Example 3.2.3 Variable X_2

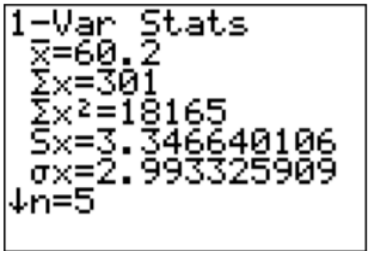


Figure 3.2.4: TI-83/84 Calculator Results for Example 3.2.3 Variable X_2

The processes for finding the mean, median, range, standard deviation, and variance on R are as follows:

variable<-c(type in your data)
To find the mean, use mean(variable)

To find the median, use median(variable)
To find the range, use range(variable). Then find maximum – minimum.
To find the standard deviation, use sd(variable)
To find the variance, use var(variable)

For the second data set in Example 3.2.3, the commands and results would be

```
productivity_2<-c(60, 58, 66, 59, 58)
mean(productivity_2)
[1] 60.2
median(productivity_2)
[1] 59
range(productivity_2)
[1] 58 66
sd(productivity_2)
[1] 3.34664
var(productivity_2)
[1] 11.2
```

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 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 are the good pilots and who are 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 3.2.6: 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 3.2.4: Using Chebyshev's Theorem

The U.S. Weather Bureau has provided the information in Example 3.2.7 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 3.2.7: Annual Number of Violent Tornados 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 about the mean annual number of strong to violent (F3+) tornados in which you would expect at least 75% of the years to fall.
b. Use Chebyshev's theorem to find an interval centered about the mean annual number of strong to violent (F3+) tornados 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 fractional number of tornados, round to the nearest whole number.

At least 75% of the years have between 2 and 91 strong to violent (F3+) tornados. (Actually, all but three years' values fall in this interval, that means that $\frac{56}{59} \approx 94.9\%$ actually 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 negative number of tornados, the lower limit is actually 0. Since you can't have fractional number of tornados, round to the nearest whole number.

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

(Actually, all but one year falls in this interval, that means that $\frac{58}{59} \approx 98.3\%$ actually 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, $\bar{x}$, and the sample standard deviation, s, to estimate the population parameter values. However, realize that using the sample standard deviation may not actually be very accurate.

Example 3.2.5 determining if a value is unusual

a. In 1974, there were 131 strong or violent (F3+) tornados in the United States. Is this value unusual? Why or why not?
b. In 1987, there were 15 strong or violent (F3+) tornados 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 3.2.4, 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, then it is unusual to have 131 strong or violent (F3+) tornados 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, then it is not unusual to have only 15 strong or violent (F3+) tornados in a year.

Homework

Exercise 3.2.1

1. Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in Example 3.2.8. Find the mean, median, range, variance, and standard deviation using technology.

Table 3.2.8: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

2. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in Example 3.2.9 (Lee, 1994).

Table 3.2.9: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	138	Kakanui	64
Waipara	64	Waitaki	209
Ashley	97	Waihao	64

River	Length (km)	River	Length (km)
Waimakariri	161	Pareora	56
Selwyn	95	Rangitata	121
Rakaia	145	Ophi	80
Ashburton	90		

a. Find the mean and median.
b. Find the range.
c. Find the variance and standard deviation.

3. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in Example 3.2.9 (Lee, 1994).

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Table 3.2.10*: Lengths of Rivers (km) Flowing to Tasman Sea*

a. Find the mean and median.
b. Find the range.
c. Find the variance and standard deviation.

4. Eyeglassmatic manufactures eyeglasses for their retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Example 3.2.11 gives the defect and the number of defects.

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398

Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

Table 3.2.11: *Number of Defective Lenses*

a. Find the mean and median.
b. Find the range.
c. Find the variance and standard deviation.

5. Print-O-Matic printing company's employees have salaries that are contained in Example 3.2.12. Find the mean, median, range, variance, and standard deviation using technology.

Table 3.2.12: Salaries of Print-O-Matic Printing Company Employees

Employee	Salary ($)	Employee	Salary ($)
CEO	272,500	Administration	66,346
Driver	58,456	Sales	109,739
CD74	100,702	Designer	90,090
CD65	57,380	Platens	69,573
Embellisher	73,877	Polar	75,526
Folder	65,270	ITEK	64,553
GTO	74,235	Mgmt	108,448
Pre Press Manager	108,448	Handwork	52,718
Pre Press Manager/IT	98,837	Horizon	76,029
Pre Press/ Graphic Artist	75,311		

6. Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in Example 3.2.13.

Table 3.2.13: Fixed Costs for Print-O-Matic Printing Company

Monthly charges	Monthly cost ($)
Bank charges	482
Cleaning	2208
Computer expensive	2471
Lease payments	2656
Postage	2117
Uniforms	2600

a. Find the mean and median.
b. Find the range.
c. Find the variance and standard deviation.

7. Compare the two data sets in problems 2 and 3 using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.
8. Example 3.2.14 contains pulse rates collected from males, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute. Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

Table 3.2.14: Pulse Rates of Males Before and After Exercise

Pulse before	Pulse after	Pulse before	Pulse after
76	88	59	92
56	110	60	104
64	126	65	82
50	90	76	150
49	83	145	155
68	136	84	140
68	125	78	141
88	150	85	131
80	146	78	132
78	168		

9. Example 3.2.15 contains pulse rates collected from females, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute. Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

Table 3.2.15: Pulse Rates of Females Before and After Exercise

Pulse before	Pulse after	Pulse before	Pulse after
96	176	92	120
82	150	70	96
86	150	75	130
72	115	70	119
78	129	70	95
90	160	68	84
88	120	47	136
71	125	64	120
66	89	70	98
76	132	74	168
70	120	85	130

10. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997) and the data is in Example 3.2.16. Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

Table 3.2.16: Pain Measurements Before and After Reiki Treatment

VAS before	VAS after	VAS before	VAS after
6	3	5	1
2	1	1	0
2	0	6	4
9	1	6	1
3	0	4	4
3	2	4	1
4	1	7	6
5	2	2	1
2	2	4	3
3	0	8	8

11. Example 3.2.17 contains data collected on the time it takes in seconds of each passage of play in a game of rugby. ("Time of passages," 2013)

Table 3.2.17: Times (in seconds) of rugby plays

39.2	2.7	9.2	14.6	1.9	17.8	15.5	53.8	17.5	27.5
4.8	8.6	22.1	29.8	10.4	9.8	27.7	32.7	32	34.3
29.1	6.5	2.8	10.8	9.2	12.9	7.1	23.8	7.6	36.4
35.6	28.4	37.2	16.8	21.2	14.7	44.5	24.7	36.2	20.9
19.9	24.4	7.9	2.8	2.7	3.9	14.1	28.4	45.5	38
18.5	8.3	56.2	10.2	5.5	2.5	46.8	23.1	9.2	10.3
10.2	22	28.5	24	17.3	12.7	15.5	4	5.6	3.8
21.6	49.3	52.4	50.1	30.5	37.2	15	38.7	3.1	11
10	5	48.8	3.6	12.6	9.9	58.6	37.9	19.4	29.2
12.3	39.2	22.2	39.7	6.4	2.5	34			

a. Using technology, find the mean and standard deviation.
b. Use Chebyshev's theorem to find an interval centered about the mean times of each passage of play in the game of rugby in which you would expect at least 75% of the times to fall.
c. Use Chebyshev's theorem to find an interval centered about the mean times of each passage of play in the game of rugby in which you would expect at least 88.9% of the times to fall.

12. Yearly rainfall amounts (in millimeters) in Sydney, Australia, are in table #3.2.18 ("Annual maximums of," 2013).

Table 3.2.18: Yearly Rainfall Amounts in Sydney, Australia

146.8	383	90.9	178.1	267.5	95.5	156.5	180
90.9	139.7	200.2	171.7	187.2	184.9	70.1	58
84.1	55.6	133.1	271.8	135.9	71.9	99.4	110.6
47.5	97.8	122.7	58.4	154.4	173.7	118.8	88
84.6	171.5	254.3	185.9	137.2	138.9	96.2	85

45.2	74.7	264.9	113.8	133.4	68.1	156.4

a. Using technology, find the mean and standard deviation.
b. Use Chebyshev's theorem to find an interval centered about the mean yearly rainfall amounts in Sydney, Australia, in which you would expect at least 75% of the amounts to fall.
c. Use Chebyshev's theorem to find an interval centered about the mean yearly rainfall amounts in Sydney, Australia, in which you would expect at least 88.9% of the amounts to fall.

13. The number of deaths attributed to UV radiation in African countries in the year 2002 is given in Example 3.2.19 ("UV radiation: Burden," 2013).

Table 3.2.19: Number of Deaths from UV Radiation

50	84	31	338	6	504	40	7	58
204	15	27	39	1	45	174	98	94
199	9	27	58	356	5	45	5	94
26	171	13	57	138	39	3	171	41
1177	102	123	433	35	40	456	125	

a. Using technology, find the mean and standard deviation.
b. Use Chebyshev's theorem to find an interval centered about the mean number of deaths from UV radiation in which you would expect at least 75% of the numbers to fall.
c. Use Chebyshev's theorem to find an interval centered about the mean number of deaths from UV radiation in which you would expect at least 88.9% of the numbers to fall.

14. The time (in 1/50 seconds) between successive pulses along a nerve fiber ("Time between nerve," 2013) are given in Example 3.2.20.

Table 3.2.20: Time (in 1/50 seconds) Between Successive Pulses

10.5	1.5	2.5	5.5	29.5	3	9	27.5	18.5	4.5
7	9.5	1	7	4.5	2.5	7.5	11.5	7.5	4
12	8	3	5.5	7.5	4.5	1.5	10.5	1	7
12	14.5	8	3.5	3.5	2	1	7.5	6	13
7.5	16.5	3	25.5	5.5	14	18	7	27.5	14

a. Using technology, find the mean and standard deviation.
b. Use Chebyshev's theorem to find an interval centered about the mean time between successive pulses along a nerve fiber in which you would expect at least 75% of the times to fall.
c. Use Chebyshev's theorem to find an interval centered about the mean time between successive pulses along a nerve fiber in which you would expect at least 88.9% of the times to fall.

15. Suppose a passage of play in a rugby game takes 75.1 seconds. Would it be unusual for this to happen? Use the mean and standard deviation that you calculated in problem 11.
16. Suppose Sydney, Australia received 300 mm of rainfall in a year. Would this be unusual? Use the mean and standard deviation that you calculated in problem 12.
17. Suppose in a given year there were 2257 deaths attributed to UV radiation in an African country. Is this value unusual? Use the mean and standard deviation that you calculated in problem 13.
18. Suppose it only takes 2 (1/50 seconds) for successive pulses along a nerve fiber. Is this value unusual? Use the mean and standard deviation that you calculated in problem 14.

Answer

1. mean = 253.93, median = 268, range = 218, variance = 2276.29, st dev = 47.71

3. a. mean = 67.68 km, median = 64 km, b. range = 145 km, c. variance = 1107.9416 km^2, st dev = 33.29 km

5. mean = $89,370.42, median = $75,311, range = $219,782, variance =2298639399, st dev = $47,944.13

7. See solutions

9. $\overline{x}_1 \approx 75.45, s_1 \approx 11.10, \overline{x}_2 \approx 125.55, s_2 \approx 24.72$

11. a. $\overline{x} \approx 21.24\text{sec}, s \approx 14.95\text{sec}$ b. $(-8.66\text{sec}, 51.14\text{sec})$ c. $(-23.61\text{sec}, 66.09\text{sec})$

13. a. $\overline{x} \approx 130.98, s \approx 205.44$ b. $(-279.90, 541.86)$ c. $(-485.34, 747.3)$

15. 3.61

17. 10.35

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

3.3: Ranking

Along with the center and the variability, another useful numerical measure is the ranking of a number. 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 3.3.1 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 3.3.2 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 really 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.

There are special percentiles called quartiles. Quartiles are numbers that divide the data into fourths. One fourth (or a quarter) of the data falls between consecutive quartiles.

Definition 3.3.1

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 *Q1*.
4. Find the median of the data above the median. This value is Q3.
   Ignore the median in both calculations for *Q1* and *Q3*

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 (*Q1*), the median, the third quartile (*Q3*), and the maximum (in that order).

The interquartile range, *IQR*, is the difference between the first and third quartiles, *Q1* and *Q3*. 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 3.3.2

Interquartile Range (*IQR*)

$IQR = Q3 - Q1$

Determining probable outliers from IQR: *fences*

A value that is less than $Q1\text{-}1.5*IQR$ (this value is often referred to as a ***low fence***) is considered an outlier.

Similarly, a value that is more than $Q3+1.5*IQR$ (the ***high 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.

An alternate form of a Box-and-Whiskers Plot, known as a modified box plot, only extends the left line to the smallest value greater than the *low fence*, and extends the left line to the largest value less than the *high fence*, and displays markers (dots, circles or asterisks) for each outlier.

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*.

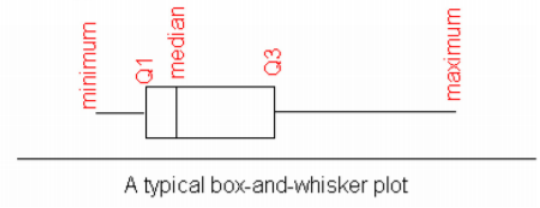


Figure 3.3.1: Typical Box Plot

Example 3.3.3 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 3.3.1 ("Reserve bank of," 2013). Find the five-number summary and the interquartile range (IQR), and draw a box-and-whiskers plot.

Table 3.3.1: Total Assets (in billions of AUD) of Australian Banks

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

Solution

Variable: $x =$ total assets of Australian banks

First sort the data.

Table 3.3.2: 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

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

Table 3.3.3: *Sorted Data for Total Assets with Median*

To find *QI*, find the median of the first half of the list.

2855	2861	2862	2884	2950	2964

Q1

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

Table 3.3.4: *Finding QI*

To find *Q3*, find the median of the second half of the list.

2965	2967	2971	3002	3014	3032

Q3

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

Table 3.3.5: *Finding Q3*

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

Minimum: 2855

Q1: 2873

Median: 2964.5

Q3: 2986.5

Maximum: 3032

To find the interquartile range, *IQR*, find *Q3-Q1*

IQR = 2986.5 - 2873 = 113.5 billion AUD

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

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

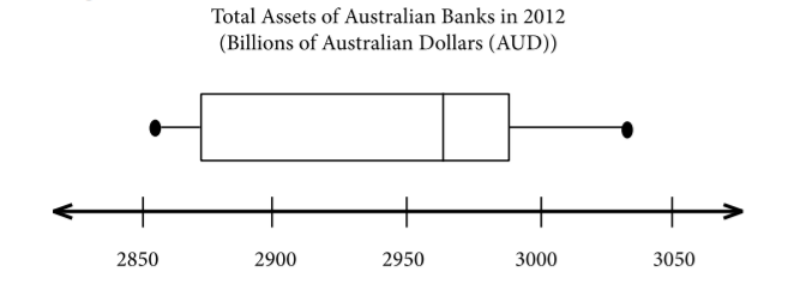


Figure 3.3.1: Box Plot of Total Assets of Australian Banks

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

Example 3.3.4 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). Find the five-number summary for the data and the *IQR*, then draw a box-and-whiskers plot.

Table 3.3.6: Life Expectancy of a Person Living in South-East Asia

70	67	69	65	69	77
65	68	75	74	64	

Solution

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

Sort the data first.

Table 3.3.7: 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.

64	65	65	67	68	69	69	70	74	75	77

Median = 69 years

Table 3.3.8*: Finding the Median of Life Expectancies*

Finding the *Q1* and *Q3* you need to find the median of the numbers below the median and above the median. The median is not included in either calculation.

64	65	65	67	68

Q1

Table 3.3.9*: Finding Q1*

69	70	74	75	77

Q3

Table 3.3.10*: Finding Q3*

Q1=65 years and *Q3*=74 years

The five-number summary is (in years)

Minimum: 64

Q1: 65

Median: 69

Q3: 74

Maximum: 77

To find the interquartile range (*IQR*)

$$IQR=Q3-Q1=74-65=9 \text{ years}$$

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

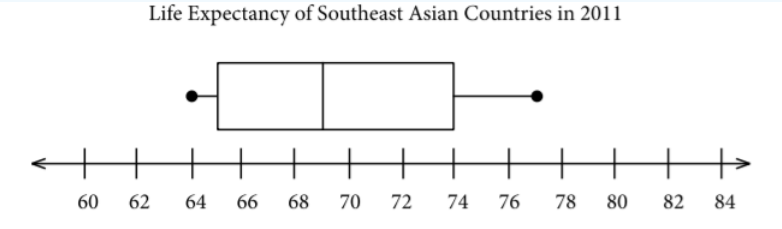


Figure 3.3.2: Box Plot of Life Expectancy

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

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.

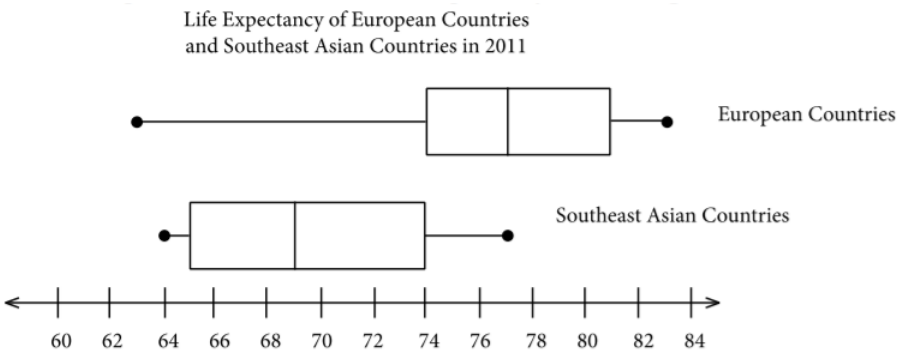


Figure 3.3.3: Box Plot of Life Expectancy of Two Regions

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.

To find the five-number summary using R, the command is:

variable<-c(type in data with commas)
summary(variable)

This command will give you the five number summary and the mean.

For Example 3.3.4, the commands would be

expectancy<-c(70, 67, 69, 65, 69, 77, 65, 68, 75, 74, 64)
summary(expectancy)

The output would be:

Min.	Ist Qu.	Median	Mean	3rd Qu.	Max.
64.00	66.00	69.00	69.36	72.00	77.00

To draw the box plot the command is boxplot(variable, main="title you want", xlab="label you want", horizontal = TRUE). The horizontal = TRUE orients the box plot to be horizontal. If you leave that part off, the box plot will be vertical by default.

For Example 3.3.4, the command is
boxplot(expectancy, main="Life Expectancy of Southeast Asian Countries in 2011",horizontal=TRUE, xlab="Life Expectancy")

You should get the box plot in *Graph 3.3.4*.

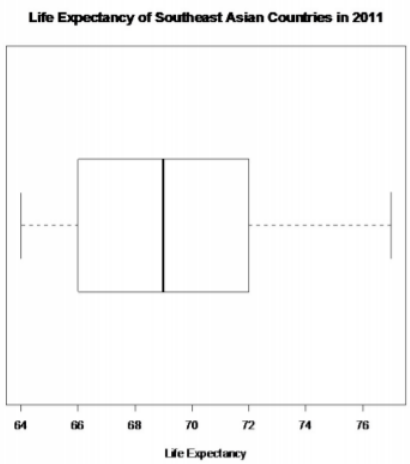


Figure 3.3.4: Box plot for Life Expectance in Southeast Asian Countries

This is known as a modified box plot. Instead of plotting the maximum and minimum, the box plot has as a lower line *Q1*-1.5**IQR* , and as an upper line, *Q3*+1.5**IQR*. Any values below the lower line or above the upper line are considered outliers. Outliers are

plotted as dots on the modified box plot. This data set does not have any outliers.

Example 3.3.5 putting it all together

A random sample was collected on the health expenditures (as a % of GDP) of countries around the world. The data is in Example 3.3.11. Using graphical and numerical descriptive statistics, analyze the data and use it to predict the health expenditures of all countries in the world.

Table 3.3.11: Health Expenditures as a Percentage of GDP

3.35	5.94	10.64	5.24	3.79	5.65	7.66	7.38	5.87	11.15
5.96	4.78	7.75	2.72	9.50	7.69	10.05	11.96	8.18	6.74
5.89	6.20	5.98	8.83	6.78	6.66	9.45	5.41	5.16	8.55

Solution

First, it might be useful to look at a visualization of the data, so create a histogram.

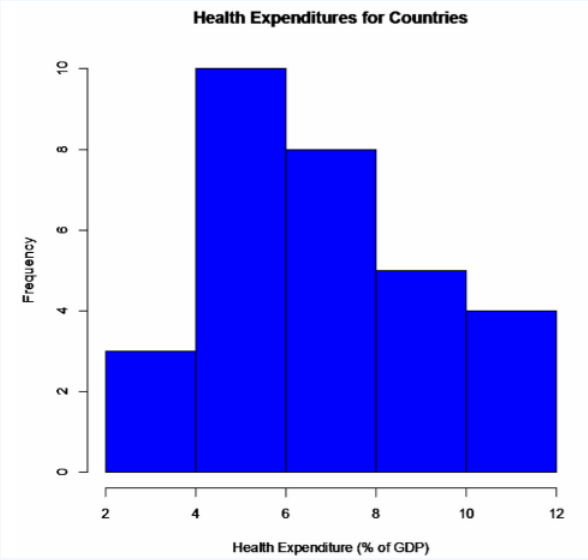


Figure 3.3.5: Histogram of Health Expenditure

From the graph, the data appears to be somewhat skewed right. So there are some countries that spend more on health based on a percentage of GDP than other countries, but the majority of countries appear to spend around 4 to 8% of their GDP on health.

Numerical descriptions might also be useful. Using technology, the mean is 7.03%, the standard deviation is 2.27%, and the five-number summary is minimum = 2.72%, Q1 = 5.71%, median = 6.70%, Q3 = 8.46%, and maximum = 11.96%. To visualize the five-number summary, create a box plot.

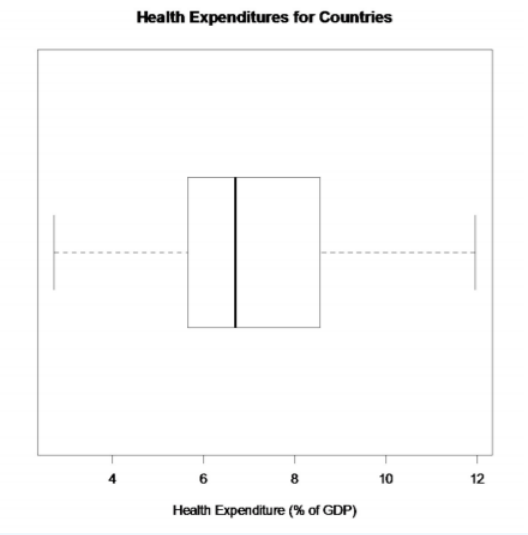


Figure 3.3.6: Box Plot of Health Expenditure

So it appears that countries spend on average about 7% of their GPD on health. The spread is somewhat low, since the standard deviation is fairly small, which means that the data is fairly consistent. The five-number summary confirms that the data is slightly skewed right. The box plot shows that there are no outliers. So from all of this information, one could say that countries spend a small percentage of their GDP on health and that most countries spend around the same amount. There doesn't appear to be any country that spends much more than other countries or much less than other countries.

Homework

Exercise 3.3.1

1. Suppose you take a standardized test and you are in the 10th percentile. What does this percentile mean? Can you say that you failed the test? Explain.
2. Suppose your child takes a standardized test in mathematics and scores in the 96th percentile. What does this percentile mean? Can you say your child passed the test? Explain.
3. Suppose your child is in the 83rd percentile in height and 24th percentile in weight. Describe what this tells you about your child's stature.
4. Suppose your work evaluates the employees and places them on a percentile ranking. If your evaluation is in the 65th percentile, do you think you are working hard enough? Explain.
5. Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in Example 3.3.12. Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot.

Table 3.3.12: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

6. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in Example 3.3.13 (Lee, 1994). Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot.

Table 3.3.13: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322

River	Length (km)	River	Length (km)
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	169	Kakanui	64
Waipara	64	Waitaki	209
Ashley	97	Waihao	64
Waimakariri	161	Pareora	56
Selwyn	95	Rangitata	121
Rakaia	145	Ophi	80
Ashburton	90		

7. The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed in Example 3.3.14 (Lee, 1994). Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot.

Table 3.3.14: Lengths of Rivers (km) Flowing to Tasman Sea

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

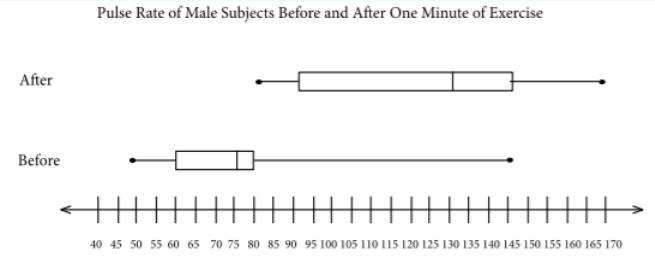
8. Eyeglassmatic manufactures eyeglasses for their retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Example 3.3.15 gives the defect and the number of defects. Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot.

Table 3.3.15: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546

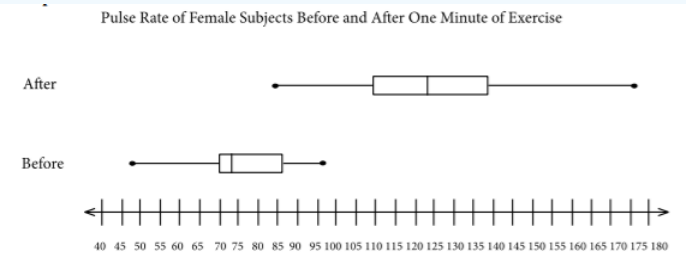
Defect type	Number of defects
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

9. A study was conducted to see the effect of exercise on pulse rate. Male subjects were taken who do not smoke, but do drink. Their pulse rates were measured ("Pulse rates before," 2013). Then they ran in place for one minute and then measured their pulse rate again. *Graph 3.3.7* is of box-and-whiskers plots that were created of the before and after pulse rates. Discuss any conclusions you can make from the graphs.



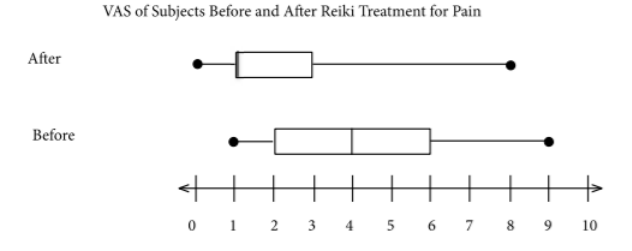
***Graph 3.3.7**: Box-and-Whiskers Plot of Pulse Rates for Males*

10. A study was conducted to see the effect of exercise on pulse rate. Female subjects were taken who do not smoke, but do drink. Their pulse rates were measured ("Pulse rates before," 2013). Then they ran in place for one minute, and after measured their pulse rate again. *Graph 3.3.8* is of box-and-whiskers plots that were created of the before and after pulse rates. Discuss any conclusions you can make from the graphs.



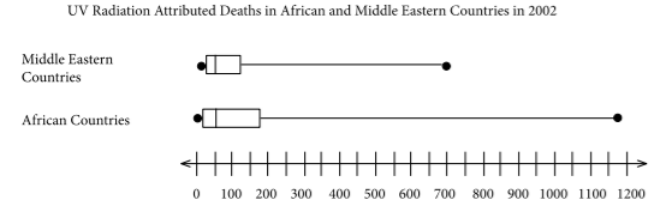
***Graph 3.3.8**: Box-and-Whiskers Plot of Pulse Rates for Females*

11. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). *Graph 3.3.9* is of box-and-whiskers plots that were created of the before and after VAS ratings. Discuss any conclusions you can make from the graphs.



***Graph 3.3.9**: Box-and-Whiskers Plot of Pain Using Reiki*

12. The number of deaths attributed to UV radiation in African countries and Middle Eastern countries in the year 2002 were collected by the World Health Organization ("UV radiation: Burden," 2013). *Graph 3.3.10* is of box-and-whiskers plots that were created of the deaths in African countries and deaths in Middle Eastern countries. Discuss any conclusions you can make from the graphs.



***Graph 3.3.10**: Box-and-Whiskers Plot of UV Radiation Deaths in Different Regions*

Answer

Note: Q1, Q3, and IQR may differ slightly due to how technology finds them.

1. See solutions

3. See solutions

5. min = 142, Q1 = 225, med = 268, Q3 = 282, max = 360, IQR = 57, see solutions

7. min = 32 km, Q1 = 46 km, med = 64 km, Q3 = 77 km, max = 177 km, IQR = 31 km, see solutions

9. See solutions

11. See solutions

Data Sources:

Annual maximums of daily rainfall in Sydney. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/sydrain.html

Lee, A. (1994). *Data analysis: An introduction based on r. Auckland.* Retrieved from http://www.statsci.org/data/oz/nzrivers.html

Life expectancy in southeast Asia. (2013, September 23). Retrieved from http://apps.who.int/gho/data/node.main.688

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, 1(2), 108-13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9765732

Pulse rates before and after exercise. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/ms212.html

Reserve bank of Australia. (2013, September 23). Retrieved from http://data.gov.au/dataset/banks-assets

Ryan, B. F., Joiner, B. L., & Ryan, Jr, T. A. (1985). *Cholesterol levels after heart attack*. Retrieved from http://www.statsci.org/data/general/cholest.html

Time between nerve pulses. (2013, September 25). Retrieved from http://www.statsci.org/data/general/nerve.html

Time of passages of play in rugby. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/rugby.html

U.S. tornado climatology. (17, May 2013). Retrieved from www.ncdc.noaa.gov/oa/climate/...tornadoes.html

UV radiation: Burden of disease by country. (2013, September 4). Retrieved from http://apps.who.int/gho/data/node.main.165?lang=en

This page titled 3.3: Ranking is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

4: Probability

4.1: Empirical Probability
4.2: Theoretical Probability
4.3: Conditional Probability
4.4: Counting Techniques

This page titled 4: Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Empirical Probability

One story about how probability theory was developed is that a gambler wanted to know when to bet more and when to bet less. He talked to a couple of friends of his that happened to be mathematicians. Their names were Pierre de Fermat and Blaise Pascal. Since then many other mathematicians have worked to develop probability theory.

Understanding probabilities are important in life. Examples of mundane questions that probability can answer for you are if you need to carry an umbrella or wear a heavy coat on a given day. More important questions that probability can help with are your chances that the car you are buying will need more maintenance, your chances of passing a class, your chances of winning the lottery, your chances of being in a car accident, and the chances that the U.S. will be attacked by terrorists. Most people do not have a very good understanding of probability, so they worry about being attacked by a terrorist but not about being in a car accident. The probability of being in a terrorist attack is much smaller than the probability of being in a car accident, thus it actually would make more sense to worry about driving. Also, the chance of you winning the lottery is very small, yet many people will spend the money on lottery tickets. Yet, if instead they saved the money that they spend on the lottery, they would have more money. In general, events that have a low probability (under 5%) are unlikely to occur. Whereas if an event has a high probability of happening (over 80%), then there is a good chance that the event will happen. This chapter will present some of the theory that you need to help make a determination of whether an event is likely to happen or not.

First you need some definitions.

Definition 4.1.1

Experiment: an activity that has specific result that can occur, but it is unknown which results will occur.

Definition 4.1.2

Outcomes: the result of an experiment.

Definition 4.1.3

Event: a set of certain outcomes of an experiment that you want to have happen.

Definition 4.1.4

Sample Space: collection of all possible outcomes of the experiment. Usually denoted as SS.

Definition 4.1.5

Event Space: the set of outcomes that make up an event. The symbol is usually a capital letter.

Start with an experiment. Suppose that the experiment is rolling a die. The sample space is {1, 2, 3, 4, 5, 6}. The event that you want is to get a 6, and the event space is {6}. To do this, roll a die 10 times. When you do that, you get a 6 two times. Based on this experiment, the probability of getting a 6 is 2 out of 10 or 1/5. To get more accuracy, repeat the experiment more times. It is easiest to put this in a table, where *n* represents the number of times the experiment is repeated. When you put the number of 6s found over the number of times you repeat the experiment, this is 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
1000	163	0.163

Notice that as *n* increased, the relative frequency seems to approach 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 probabilities** since they are found by actually doing the experiment. 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 by $\frac{163}{1000} = 0.163$.

You must do experimental probabilities whenever it is not possible to calculate probabilities using other means. An example is if you want to find the probability that a family has 5 children, you would have to actually 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 have to roll the die many times and count how often each side comes up. Make sure you repeat an experiment many times, because otherwise you will not 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.

Homework

Exercise 4.1.1

1. Example 4.1.2 contains the number of M&M's of each color that were found in a case (Madison, 2013). Find the probability of choosing each color based on this experiment.

Table 4.1.2: M&M Distribution

Blue	Brown	Green	Orange	Red	Yellow	Total
481	371	483	544	372	369	2620

2. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Example 4.1.3 gives the defect and the number of defects. Find the probability of each defect type based on this data.

Table 4.1.3: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485

Defect type	Number of defects
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

3. In Australia in 1995, of the 2907 indigenous people in prison 17 of them died. In that same year, of the 14501 non-indigenous people in prison 42 of them died ("Aboriginal deaths in," 2013). Find the probability that an indigenous person dies in prison and the probability that a non-indigenous person dies in prison. Compare these numbers and discuss what the numbers may mean.
4. A project conducted by the Australian Federal Office of Road Safety asked people many questions about their cars. One question was the reason that a person chooses a given car, and that data is in Example 4.1.4 ("Car preferences," 2013). Find the probability a person chooses a car for each of the given reasons.

Table 4.1.4: Reason for Choosing a Car

Safety	Reliability	Cost	Performance	Comfort	Looks
84	62	46	34	47	27

Answer

1. P(blue) = 0.184, P(brow) = 0.142, P(green) = 0.184, P(orange) = 0.208, P(red) = 0.142, P(yellow) = 0.141

3. P(indigenous person dies) = 0.0058, P(non-indigenous person dies) = 0.0029, see solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

4.2: Theoretical Probability

It is not always feasible to conduct an experiment over and over again, so it would be better to be able to find the probabilities without conducting the experiment. These probabilities are called **Theoretical Probabilities**.

To be able to do theoretical probabilities, there is an assumption that you need to consider. It is that all of the outcomes in the sample space need to be **equally likely outcomes**. This means that every outcome of the experiment needs to have the same chance of happening.

Example 4.2.1 Equally likely outcomes

Which of the following experiments have equally likely outcomes?

a. Rolling a fair die.
b. Flip a coin that is weighted so one side comes up more often than the other.
c. Pull a ball out of a can containing 6 red balls and 8 green balls. All balls are the same size.
d. Picking a card from a deck.
e. Rolling a die to see if it is fair.

Solution

a. Since the die is fair, every side of the die has the same chance of coming up. The outcomes are the different sides, so each outcome is equally likely.
b. Since the coin is weighted, one side is more likely to come up than the other side. The outcomes are the different sides, so each outcome is not equally likely.
c. Since each ball is the same size, then each ball has the same chance of being chosen. The outcomes of this experiment are the individual balls, so each outcome is equally likely. Don't assume that because the chances of pulling a red ball are less than pulling a green ball that the outcomes are not equally likely. The outcomes are the individual balls and they are equally likely.
d. If you assume that the deck is fair, then each card has the same chance of being chosen. Thus the outcomes are equally likely outcomes. You do have to make this assumption. For many of the experiments you will do, you do have to make this kind of assumption.
e. In this case you are not sure the die is fair. The only way to determine if it is fair is to actually conduct the experiment, since you don't know if the outcomes are equally likely. If the experimental probabilities are fairly close to the theoretical probabilities, then the die is fair.

If the outcomes are not equally likely, then you must do experimental probabilities. If the outcomes are equally likely, then you can do theoretical probabilities.

Definition 4.2.1: Theoretical Probabilities

If the outcomes of an experiment are equally likely, then the probability of event A happening is

$$P(A) = \frac{\#\text{ of outcomes in event space}}{\#\text{ of outcomes in sample space}}$$

Example 4.2.2 calculating theoretical probabilities

Suppose you conduct an experiment where you flip a fair coin twice.

a. What is the sample space?
b. What is the probability of getting exactly one head?
c. What is the probability of getting at least one head?
d. What is the probability of getting a head and a tail?
e. What is the probability of getting a head or a tail?
f. What is the probability of getting a foot?
g. What is the probability of each outcome? What is the sum of these probabilities?

Solution

a. There are several different sample spaces you can do. One is SS={0, 1, 2} where you are counting the number of heads. However, the outcomes are not equally likely since you can get one head by getting a head on the first flip and a tail on the second or a tail on the first flip and a head on the second. There are 2 ways to get that outcome and only one way to get the other outcomes. Instead it might be better to give the sample space as listing what can happen on each flip. Let H = head and T = tail, and list which can happen on each flip.

SS={HH, HT, TH, TT}

b. Let A = getting exactly one head. The event space is A = {HT, TH}. So

$$P(A) = \frac{2}{4} \text{ or } \frac{1}{2}$$

It may not be advantageous to reduce the fractions to lowest terms, since it is easier to compare fractions if they have the same denominator.

c. Let B = getting at least one head. At least one head means get one or more. The event space is B = {HT, TH, HH} and

$$P(B) = \frac{3}{4}$$

Since P(B) is greater than the P(A), then event B is more likely to happen than event A.

d. Let C = getting a head and a tail = {HT, TH} and

$$P(C) = \frac{2}{4}$$

This is the same event space as event A, but it is a different event. Sometimes two different events can give the same event space.

e. Let D = getting a head or a tail. Since or means one or the other or both and it doesn't specify the number of heads or tails, then D = {HH, HT, TH, TT} and

$$P(D) = \frac{4}{4} = 1$$

f. Let E = getting a foot. Since you can't get a foot, E = {} or the empty set and

$$P(E) = \frac{0}{4} = 0$$

g. $P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$. If you add all of these probabilities together you get 1.

This example had some results in it that are important concepts. They are summarized below:

Probability Properties

1. $0 \leq P(\text{ event }) \leq 1$
2. If the P(event)=1, then it will happen and is called the certain event.
3. If the P(event)=0, then it cannot happen and is called the impossible event.
4. $\sum P(\text{ outcome }) = 1$

Example 4.2.3 calculating theoretical probabilities

Suppose you conduct an experiment where you pull a card from a standard deck.

a. What is the sample space?
b. What is the probability of getting a Spade?
c. What is the probability of getting a Jack?
d. What is the probability of getting an Ace?
e. What is the probability of not getting an Ace?
f. What is the probability of getting a Spade and an Ace?
g. What is the probability of getting a Spade or an Ace?
h. What is the probability of getting a Jack and an Ace?
i. What is the probability of getting a Jack or an Ace?

Solution

a. SS = {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS, 2C, 3C, 4C, 5C, 6C, 7C, 8C, 9C, 10C, JC, QC, KC, AC, 2D, 3D, 4D, 5D, 6D, 7D, 8D, 9D, 10D, JD, QD, KD, AD, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH, AH}

b. Let A = getting a spade = {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS} so

$$P(A) = \frac{13}{52}$$

c. Let B = getting a Jack = {JS, JC, JH, JD} so

$$P(B) = \frac{4}{52}$$

d. Let C = getting an Ace = {AS, AC, AH, AD} so

$$P(C) = \frac{4}{52}$$

e. Let D = not getting an Ace = {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, 2C, 3C, 4C, 5C, 6C, 7C, 8C, 9C, 10C, JC, QC, KC, 2D, 3D, 4D, 5D, 6D, 7D, 8D, 9D, 10D, JD, QD, KD, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH} so

$$P(D) = \frac{48}{52}$$

Notice, $P(D) + P(C) = \frac{48}{52} + \frac{4}{52} = 1$, so you could have found the probability of D by doing 1 minus the probability of C $P(D) = 1 - P(C) = 1 - \frac{4}{52} = \frac{48}{52}$.

f. Let E = getting a Spade and an Ace = {AS} so

$$P(E) = \frac{1}{52}$$

g. Let F = getting a Spade and an Ace ={2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS, AC, AD, AH} so

$$P(F) = \frac{16}{52}$$

h. Let G = getting a Jack and an Ace = { } since you can't do that with one card. So

$$P(G) = 0$$

i. Let H = getting a Jack or an Ace = {JS, JC, JD, JH, AS, AC, AD, AH} so

$$P(H) = \frac{8}{52}$$

Example 4.2.4 calculating theoretical probabilities

Suppose you have an iPod Shuffle with the following songs on it: 5 Rolling Stones songs, 7 Beatles songs, 9 Bob Dylan songs, 4 Faith Hill songs, 2 Taylor Swift songs, 7 U2 songs, 4 Mariah Carey songs, 7 Bob Marley songs, 6 Bunny Wailer songs, 7 Elton John songs, 5 Led Zeppelin songs, and 4 Dave Mathews Band songs. The different genre that you have are rock from the 60s which includes Rolling Stones, Beatles, and Bob Dylan; country includes Faith Hill and Taylor Swift; rock of the 90s includes U2 and Mariah Carey; Reggae includes Bob Marley and Bunny Wailer; rock of the 70s includes Elton John and Led Zeppelin; and bluegrasss/rock includes Dave Mathews Band.

The way an iPod Shuffle works, is it randomly picks the next song so you have no idea what the next song will be. Now you would like to calculate the probability that you will hear the type of music or the artist that you are interested in. The sample set is too difficult to write out, but you can figure it from looking at the number in each set and the total number. The total number of songs you have is 67.

a. What is the probability that you will hear a Faith Hill song?
b. What is the probability that you will hear a Bunny Wailer song?
c. What is the probability that you will hear a song from the 60s?
d. What is the probability that you will hear a Reggae song?
e. What is the probability that you will hear a song from the 90s or a bluegrass/rock song?
f. What is the probability that you will hear an Elton John or a Taylor Swift song?
g. What is the probability that you will hear a country song or a U2 song?

Solution

a. There are 4 Faith Hill songs out of the 67 songs, so

$$P(\text{ Faith Hill song }) = \frac{4}{67}$$

b. There are 6 Bunny Wailer songs, so

$$P(\text{ Bunny Wailer }) = \frac{6}{67}$$

c. There are 5, 7, and 9 songs that are classified as rock from the 60s, which is 21 total, so

$$P(\text{ rock from the 60s}) = \frac{21}{67}$$

d. There are 6 and 7 songs that are classified as Reggae, which is 13 total, so

$$P(\text{ Reggae }) = \frac{13}{67}$$

e. There are 7 and 4 songs that are songs from the 90s and 4 songs that are bluegrass/rock, for a total of 15, so

$$P(\text{ rock from the 90 s or bluegrass/rock }) = \frac{15}{67}$$

f. There are 7 Elton John songs and 2 Taylor Swift songs, for a total of 9, so

$$P(\text{ Elton John or Taylor Swift song }) = \frac{9}{67}$$

g. There are 6 country songs and 7 U2 songs, for a total of 13, so

$$P(\text{ country or U2 song }) = \frac{13}{67}$$

Of course you can do any other combinations you would like.

Notice in Example 4.2.3 part e, it was mentioned that the probability of event D plus the probability of event C was 1. This is because these two events have no outcomes in common, and together they make up the entire sample space. Events that have this property are called **complementary events**.

Definition 4.2.2: complementary events

If two events are **complementary events** then to find the probability of one just subtract the probability of the other from one. Notation used for complement of A is not A or A^c.

$$P(A) + P(A^c) = 1, \text{ or } P(A) = 1 - P(A^c)$$

Example 4.2.5 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?

Solution

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 do 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 = {QS, QC, QD, QH} so

$$P(\text{ Queen }) = \frac{4}{52} \text{ and}$$

$$P(\text{ not Queen }) = 1 - P(\text{ Queen }) = 1 - \frac{4}{52} = \frac{48}{52}$$

The complement is useful when you are trying to find the probability of an event that involves the words at least or an event that involves the words at most. As an example of an at least event is suppose you want to find the probability of making at least $50,000 when you graduate from college. That means you want the probability of your salary being greater than or equal to $50,000. An example of an at most event is suppose you want to find the probability of rolling a die and getting at most a 4. That means that you want to get less than or equal to a 4 on the die. The reason to use the complement is that sometimes it is easier to find the probability of the complement and then subtract from 1. Example 4.2.6 demonstrates how to do this.

Example 4.2.6 using the complement to find probabilities

a. In an experiment of rolling a fair die one time, find the probability of rolling at most a 4 on the die.
b. In an experiment of pulling a card from a fair deck, find the probability of pulling at least a 5 (ace is a high card in this example).

Solution

a. The sample space for this experiment is {1, 2, 3, 4, 5, 6}. You want the event of getting at most a 4, which is the same as thinking of getting 4 or less. The event space is {1, 2, 3, 4}. The probability is

$$P(\text{ at most } 4) = \frac{4}{6}$$

Or you could have used the complement. The complement of rolling at most a 4 would be rolling number bigger than 4. The event space for the complement is {5, 6}. The probability of the complement is $\frac{2}{6}$. The probability of at most 4 would be

$$P(\text{ at most } 4) = 1 - P(\text{ more than } 4) = 1 - \frac{2}{6} = \frac{4}{6}$$

Notice you have the same answer, but the event space was easier to write out. On this example it probability wasn't that useful, but in the future there will be events where it is much easier to use the complement.

b. The sample space for this experiment is

SS = {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS, 2C, 3C, 4C, 5C, 6C, 7C, 8C, 9C, 10C, JC, QC, KC, AC, 2D, 3D, 4D, 5D, 6D, 7D, 8D, 9D, 10D, JD, QD, KD, AD, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH, AH}

Pulling a card that is at least a 5 would involve listing all of the cards that are a 5 or more. It would be much easier to list the outcomes that make up the complement. The complement of at least a 5 is less than a 5. That would be the event of 4 or less. The event space for the complement would be {2S, 3S, 4S, 2C, 3C, 4C, 2D, 3D, 4D, 2H, 3H, 4H}. The probability of the complement would be $\frac{12}{52}$. The probability of at least a 5 would be

$$P(\text{ at least a5}) = 1 - P(4 \text{ or less }) = 1 - \frac{12}{52} = \frac{40}{52}$$

Another concept was show in Example 4.2.3 parts g and i. The problems were looking for the probability of one event or another. In part g, it was looking for the probability of getting a Spade or an Ace. That was equal to $\frac{16}{52}$. In part i, it was looking for the probability of getting a Jack or an Ace. That was equal to $\frac{8}{52}$. If you look back at the parts b, c, and d, you might notice the following result:

$$P(\text{ Jack }) + P(\text{ Ace }) = P(\text{ Jack or Ace }) \text{ but } P(\text{ Spade }) + P(\text{ Ace }) \neq P(\text{ Spade or Ace })$$

Why does adding two individual probabilities together work in one situation to give the probability of one or another event and not give the correct probability in the other?

The reason this is true in the case of the Jack and the Ace is that these two events cannot happen together. There is no overlap between the two events, and in fact the $P(\text{ Jack and Acc}) = 0$. However, in the case of the Spade and Ace, they can happen together. There is overlap, mainly the ace of spades. The $P(\text{ Spade and Ace}) \neq 0$.

When two events cannot happen at the same time, they are called **mutually exclusive**. In the above situation, the events Jack and Ace are mutually exclusive, while the events Spade and Ace are not mutually exclusive.

Addition Rules:

If two events A and B are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) \text{ and } P(A \text{ and } B) = 0$$

If two events A and B are not mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example 4.2.7 using addition rules

Suppose your experiment is to roll two fair dice.

a. What is the sample space?
b. What is the probability of getting a sum of 5?
c. What is the probability of getting the first die a 2?
d. What is the probability of getting a sum of 7?
e. What is the probability of getting a sum of 5 and the first die a 2?
f. What is the probability of getting a sum of 5 or the first die a 2?
g. What is the probability of getting a sum of 5 and sum of 7?
h. What is the probability of getting a sum of 5 or sum of 7?

Solution

a. As with the other examples you need to come up with a sample space that has equally likely outcomes. One sample space is to list the sums possible on each roll. That sample space would look like: SS = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. However, there are more ways to get a sum of 7 then there are to get a sum of 2, so these outcomes are not equally likely. Another thought is to list the possibilities on each roll. As an example you could roll the dice and on the first die you could get a 1. The other die could be any number between 1 and 6, but say it is a 1 also. Then this outcome would look like (1,1). Similarly, you could get (1, 2), (1, 3), (1,4), (1, 5), or (1, 6). Also, you could get a 2, 3, 4, 5, or 6 on the first die instead. Putting this all together, you get the sample space:

$$\begin{aligned} \text{SS} = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

Notice that a (2,3) is different from a (3,2), since the order that you roll the die is important and you can tell the difference between these two outcomes. You don't need any of the doubles twice, since these are not distinguishable from each other in either order. This will always be the sample space for rolling two dice.

b. Let A = getting a sum of 5 = {(4,1), (3,2), (2,3), (1,4)} so

$$P(A) = \frac{4}{36}$$

c. Let B = getting first die a 2 = {(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)} so

$$P(B) = \frac{6}{36}$$

d. Let C = getting a sum of 7 = {(6,1), (5,2), (4,3), (3,4), (2,5), (1,6)} so

$$P(C) = \frac{6}{36}$$

e. This is events A and B which contains the outcome {(2,3)} so

$$P(A \text{ and } B) = \frac{1}{36}$$

f. Notice from part e, that these two events are not mutually exclusive, so

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B)\\ &= \frac{4}{36} + \frac{6}{36} - \frac{1}{36} \end{aligned}$$

$$= \frac{9}{36}$$

g. These are the events A and C, which have no outcomes in common. Thus A and C = { } so

$$P(A \text{ and } C) = 0$$

h. From part g, these two events are mutually exclusive, so

$$P(A \text{ or } C) = P(A) + P(C)$$
$$= \frac{4}{36} + \frac{6}{36}$$
$$= \frac{10}{36}$$

Odds

Many people like to talk about the odds of something happening or not happening. Mathematicians, statisticians, and scientists prefer to deal with probabilities since odds are difficult to work with, but gamblers prefer to work in odds for figuring out how much they are paid if they win.

Definition 4.2.3

The **actual odds against** event *A* occurring are the ratio $P(A^c)/P(A)$, usually expressed in the form *a:b* or *a* to *b*, where *a* and *b* are integers with no common factors.

Definition 4.2.4

The **actual odds in favor** event *A* occurring are the ratio $P(A)/P(A^c)$, which is the reciprocal of the odds against. If the odds against event *A* are *a:b*, then the odds in favor event *A* are *b:a*.

Definition 4.2.5

The **payoff odds** against event *A* occurring are the ratio of the net profit (if you win) to the amount bet.

payoff odds against event *A* = (net profit) : (amount bet)

Example 4.2.8 odds against and payoff odds

In the game of Craps, if a shooter has a come-out roll of a 7 or an 11, it is called a natural and the pass line wins. The payoff odds are given by a casino as 1:1.

a. Find the probability of a natural.
b. Find the actual odds for a natural.
c. Find the actual odds against a natural.
d. If the casino pays 1:1, how much profit does the casino make on a $10 bet?

Solution

a. A natural is a 7 or 11. The sample space is

$$\begin{aligned} \text{SS} = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

The event space is {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1), (5,6), (6,5)}

$$\text{So } P(7 \text{ or } 11) = \frac{8}{36}$$

b.

$$\text{odd for a natural} = \frac{P(7 \text{ or } 11)}{P(\text{not} 7 \text{ or } 11)}$$
$$= \frac{8/36}{1 - 8/36}$$
$$= \frac{8/36}{28/36}$$
$$= \frac{8}{28}$$
$$= \frac{2}{7}$$

c.

$$\text{odds against a natural} = \frac{P(\text{ not } 7 \text{ or } 11)}{P(7 \text{ or } 11)} = \frac{28}{8} = \frac{7}{2} = \frac{3.5}{1}$$

d. The actual odds are 3.5 to 1 while the payoff odds are 1 to 1. The casino pays you \$10 for your \$10 bet. If the casino paid you the actual odds, they would pay \$3.50 on every \$1 bet, and on \$10, they pay $3.5 * \$10 = \$35$. Their profit is $\$35 - \$10 = \$25$.

Homework

Exercise 4.2.1

1. Example 4.2.1 contains the number of M&M's of each color that were found in a case (Madison, 2013).

Blue	Brown	Green	Orange	Red	Yellow	Total
481	371	483	544	372	369	2620

Table 4.2.1: *M&M Distribution*

a. Find the probability of choosing a green or red M&M.
b. Find the probability of choosing a blue, red, or yellow M&M.
c. Find the probability of not choosing a brown M&M.
d. Find the probability of not choosing a green M&M.

2. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made in a time period. Example 4.2.2 gives the defect and the number of defects.

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble	976

Table 4.2.2: *Number of Defective Lenses*

a. Find the probability of picking a lens that is scratched or flaked.
b. Find the probability of picking a lens that is the wrong PD or was lost in lab.
c. Find the probability of picking a lens that is not scratched.
d. Find the probability of picking a lens that is not the wrong shape.

3. An experiment is to flip a fair coin three times.
   a. State the sample space.
   b. Find the probability of getting exactly two heads. Make sure you state the event space.
   c. Find the probability of getting at least two heads. Make sure you state the event space.
   d. Find the probability of getting an odd number of heads. Make sure you state the event space.
   e. Find the probability of getting all heads or all tails. Make sure you state the event space.
   f. Find the probability of getting exactly two heads or exactly two tails.
   g. Find the probability of not getting an odd number of heads.
4. An experiment is rolling a fair die and then flipping a fair coin.
   a. State the sample space.
   b. Find the probability of getting a head. Make sure you state the event space.
   c. Find the probability of getting a 6. Make sure you state the event space.
   d. Find the probability of getting a 6 or a head.
   e. Find the probability of getting a 3 and a tail.
5. An experiment is rolling two fair dice.
   a. State the sample space.
   b. Find the probability of getting a sum of 3. Make sure you state the event space.
   c. Find the probability of getting the first die is a 4. Make sure you state the event space.
   d. Find the probability of getting a sum of 8. Make sure you state the event space.
   e. Find the probability of getting a sum of 3 or sum of 8.
   f. Find the probability of getting a sum of 3 or the first die is a 4.
   g. Find the probability of getting a sum of 8 or the first die is a 4.
   h. Find the probability of not getting a sum of 8.
6. An experiment is pulling one card from a fair deck.
   a. State the sample space.
   b. Find the probability of getting a Ten. Make sure you state the event space.
   c. Find the probability of getting a Diamond. Make sure you state the event space.
   d. Find the probability of getting a Club. Make sure you state the event space.
   e. Find the probability of getting a Diamond or a Club.
   f. Find the probability of getting a Ten or a Diamond.
7. An experiment is pulling a ball from an urn that contains 3 blue balls and 5 red balls.
   a. Find the probability of getting a red ball.
   b. Find the probability of getting a blue ball.
   c. Find the odds for getting a red ball.
   d. Find the odds for getting a blue ball.
8. In the game of roulette, there is a wheel with spaces marked 0 through 36 and a space marked 00.
   a. Find the probability of winning if you pick the number 7 and it comes up on the wheel.
   b. Find the odds against winning if you pick the number 7.
   c. The casino will pay you $20 for every dollar you bet if your number comes up. How much profit is the casino making on the bet?

Answer

1. a. P(green or red) = 0.326, b. P(blue, red, or yellow) = 0.466, c. P(not brown) = 0.858, d. P(not green) = 0.816

3. a. See solutions, b. P(2 heads) = 0.375, c. P(at least 2 heads) = 0.50, d. P(odd number of heads) = 0.50, e. P(all heads or all tails) = 0.25, f. P(two heads or two tails) = 0.75, g. P(no an odd number of heads) = 0.50

5. a. See solutions, b. P(sum of 3) = 0.056, c. P(1st die a 4) = 0.167, d. P(sum of 8) = 0.139, e. P(sum of 3 or sum of 8) = 0.194, f. P(sum of 3 or 1st die a 4) = 0.222, g. P(sum of 8 or 1st die a 4) = 0.278, h. P(not getting a sum of 8) = 0.861

7. a. P(red ball) = 0.625, b. P(blue ball) = 0.375, c. 5 to 3 d. 3 to 5

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

4.3: Conditional Probability

Suppose you want to figure out if you should buy a new car. When you first go and look, you find two cars that you like the most. In your mind they are equal, and so each has a 50% chance that you will pick it. Then you start to look at the reviews of the cars and realize that the first car has had 40% of them needing to be repaired in the first year, while the second car only has 10% of the cars needing to be repaired in the first year. You could use this information to help you decide which car you want to actually purchase. Both cars no longer have a 50% chance of being the car you choose. You could actually calculate the probability you will buy each car, which is a conditional probability. You probably wouldn't do this, but it gives you an example of what a conditional probability is.

Conditional probabilities are probabilities calculated after information is given. This is where you want to find the probability of event A happening after you know that event B has happened. If you know that B has happened, then you don't need to consider the rest of the sample space. You only need the outcomes that make up event B. Event B becomes the new sample space, which is called the **restricted sample space**, R. If you always write a restricted sample space when doing conditional probabilities and use this as your sample space, you will have no trouble with conditional probabilities. The notation for conditional probabilities is $P(A,\text{ given } B) = P(A|B)$. The event following the vertical line is always the restricted sample space.

Example 4.3.1 conditional probabilities

a. Suppose you roll two dice. What is the probability of getting a sum of 5, given that the first die is a 2?
b. Suppose you roll two dice. What is the probability of getting a sum of 7, given the first die is a 4?
c. Suppose you roll two dice. What is the probability of getting the second die a 2, given the sum is a 9?
d. Suppose you pick a card from a deck. What is the probability of getting a Spade, given that the card is a Jack?
e. Suppose you pick a card from a deck. What is the probability of getting an Ace, given the card is a Queen?

Solution

a. Since you know that the first die is a 2, then this is your restricted sample space, so

R = {(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)}

Out of this restricted sample space, the way to get a sum of 5 is {(2,3)}. Thus

$$P(\text{ sum of 5}|\text{ the first die is a 2}) = \frac{1}{6}$$

b. Since you know that the first die is a 4, this is your restricted sample space, so

R = {(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)}

Out of this restricted sample space, the way to get a sum of 7 is {(4,3)}. Thus

$$P(\text{ sum of 7}|\text{ the first die is a 4}) = \frac{1}{6}$$

c. Since you know the sum is a 9, this is your restricted sample space, so

R = {(3,6), (4,5), (5,4), (6,3)}

Out of this restricted sample space there is no way to get the second die a 2. Thus

$$P(\text{ second die is a 2}|\text{ sum is 9}) = 0$$

d. Since you know that the card is a Jack, this is your restricted sample space, so

R = {JS, JC, JD, JH}

Out of this restricted sample space, the way to get a Spade is {JS}. Thus

$$P(\text{ Spade }|\text{Jack}) = \frac{1}{4}$$

e. on: Since you know that the card is a Queen, then this is your restricted sample space, so

R = {QS, QC, QD, QH}

Out of this restricted sample space, there is no way to get an Ace, thus

$$P(\text{ Ace }|\text{ Queen }) = 0$$

If you look at the results of Example 4.3.7 part d and Example 4.3.1 part b, you will notice that you get the same answer. This means that knowing that the first die is a 4 did not change the probability that the sum is a 7. This added knowledge did not help you in any way. It is as if that information was not given at all. However, if you compare Example 4.3.7 part b and Example 4.3.1 part a, you will notice that they are not the same answer. In this case, knowing that the first die is a 2 did change the probability of getting a sum of 5. In the first case, the events sum of 7 and first die is a 4 are called **independent events**. In the second case, the events sum of 5 and first die is a 2 are called **dependent events**.

Events A and B are considered **independent events** if the fact that one event happens does not change the probability of the other event happening. In other words, events A and B are independent if the fact that B has happened does not affect the probability of event A happening and the fact that A has happened does not affect the probability of event B happening. Otherwise, the two events are dependent. In symbols, A and B are independent if

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

Example 4.3.2 independent events

a. Suppose you roll two dice. Are the events "sum of 7" and "first die is a 3" independent?
b. Suppose you roll two dice. Are the events "sum of 6" and "first die is a 4" independent?
c. Suppose you pick a card from a deck. Are the events "Jack" and "Spade" independent?
d. Suppose you pick a card from a deck. Are the events "Heart" and "Red" card independent?
e. Suppose you have two children via separate births. Are the events "the first is a boy" and "the second is a girl" independent?
f. Suppose you flip a coin 50 times and get a head every time, what is the probability of getting a head on the next flip?

Solution

a. To determine if they are independent, you need to see if $P(A|B) = P(A)$. It doesn't matter which event is A or B, so just assign one as A and one as B.

Let A = sum of 7 = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)} and B = first die is a 3 = {(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)} $P(A|B)$ means that you assume that B has happened. The restricted sample space is B,

R = {(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)}

In this restricted sample space, the way for A to happen is {(3,4)}, so

$$P(A|B) = \frac{1}{6}$$

$$\text{The } P(A) = \frac{6}{36} = \frac{1}{6}$$

$P(A|B) = P(A)$ Thus "sum of 7" and "first die is a 3" are independent events.

b. To determine if they are independent, you need to see if $P(A|B) = P(A)$. It doesn't matter which event is A or B, so just assign one as A and one as B.

Let A = sum of 6 = {(1,5), (2,4), (3,3), (4,2), (5,1)} and B = first die is a 4 = {(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)}, so

$$P(A) = \frac{5}{36}$$

For $P(A|B)$, the restricted sample space is B,

R = {(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)}

In this restricted sample space, the way for A to happen is {(4,2)}, so

$$P(A|B) = \frac{1}{6}.$$

In this case, "sum of 6" and "first die is a 4" are dependent since $P(A|B) \neq P(A)$.

c. To determine if they are independent, you need to see if $P(A|B) = P(A)$. It doesn't matter which event is A or B, so just assign one as A and one as B.

Let A = Jack = {JS, JC, JD, JH} and B = Spade {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS}

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

For $P(A|B)$, the restricted sample space is B,

R = {2S, 3S, 4S, 5S, 6S, 7S, 8S, 9S, 10S, JS, QS, KS, AS}

In this restricted sample space, the way A happens is {JS}, so

$$P(A|B) = \frac{1}{13}$$

In this case, “Jack” and “Spade” are independent since $P(A|B) = P(A)$.

d. To determine if they are independent, you need to see if $P(A|B) = P(A)$. It doesn’t matter which event is A or B, so just assign one as A and one as B.

Let A = Heart = {2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH, AH} and B = Red card = {2D, 3D, 4D, 5D, 6D, 7D, 8D, 9D, 10D, JD, QD, KD, AD, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH, AH}, so

$$P(A) = \frac{13}{52} = \frac{1}{4}$$

For $P(A|B)$, the restricted sample space is B,

R = {2D, 3D, 4D, 5D, 6D, 7D, 8D, 9D, 10D, JD, QD, KD, AD, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH, AH}

In this restricted sample space, the way A can happen is 13,

$$P(A|B) = \frac{13}{26} = \frac{1}{2}.$$

In this case, “Heart” and “Red” card are dependent, since $P(A|B) \neq P(A)$.

e. In this case, you actually don’t need to do any calculations. The gender of one child does not affect the gender of the second child, the events are independent.

f. Since one flip of the coin does not affect the next flip (the coin does not remember what it did the time before), the probability of getting a head on the next flip is still one-half.

Multiplication Rule:

Two more useful formulas: If two events are dependent, then $P(A \text{ and } B) = P(A) * P(B|A)$

If two events are independent, then $P(A \text{ and } B) = P(A)^*P(B)$

If you solve the first equation for $P(B|A)$, you obtain $P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$, which is a formula to calculate a conditional probability. However, it is easier to find a conditional probability by using the restricted sample space and counting unless the sample space is large.

Example 4.3.3 Multiplication rule

a. Suppose you pick three cards from a deck, what is the probability that they are all Queens if the cards are not replaced after they are picked?
b. Suppose you pick three cards from a deck, what is the probability that they are all Queens if the cards are replaced after they are picked and before the next card is picked?

Solution

a. This sample space is too large to write out, so using the multiplication rule makes sense. Since the cards are not replaced, then the probability will change for the second and third cards. They are dependent events. This means that on the second draw there is one less Queen and one less card, and on the third draw there are two less Queens and 2 less cards.

$$P(\text{3 Queens}) = P(\text{Q on 1st and Q on 2nd and Q on 3rd})$$

$$= P(\text{Q on 1st}) * P(\text{Q on 2nd} | \text{Q on 1st}) * P(\text{Q on 3rd} | \text{1st and 2nd Q})$$

$$= \frac{4}{52} * \frac{3}{51} * \frac{2}{50}$$

$$= \frac{24}{132600}$$

b. Again, the sample space is too large to write out, so using the multiplication rule makes sense. Since the cards are put back, one draw has no affect on the next draw and they are all independent.

$$P(\text{3 Queens}) = P(\text{Queen on 1st and Queen on 2nd and Queen on 3rd})$$

$$= P(\text{Queen on 1st}) * P(\text{Queen on 2nd}) * P(\text{Queen on 3rd})$$

$$= \frac{4}{52} * \frac{4}{52} * \frac{4}{52}$$

$$= \left(\frac{4}{52}\right)^3$$

$$= \frac{64}{140608}$$

Example 4.3.4 application problem

The World Health Organization (WHO) keeps track of how many incidents of leprosy there are in the world. Using the WHO regions and the World Banks income groups, one can ask if an income level and a WHO region are dependent on each other in terms of predicting where the disease is. Data on leprosy cases in different countries was collected for the year 2011 and a summary is presented in Example 4.3.1 ("Leprosy: Number of," 2013).

Table 4.3.1: Number of Leprosy Cases

WHO Region	World Bank Income Group				Row Total
	High Income	Upper Middle Income	Lower Middle Income	Low Income	
Americas	174	36028	615	0	36817
Eastern Mediterranean	54	6	1883	604	2547
Europe	10	0	0	0	10
Western Pacific	26	216	3689	1155	5086
Africa	0	39	1986	15928	17953
South-East Asia	0	0	149896	10236	160132
Column Total	264	36289	158069	27923	222545

a. Find the probability that a person with leprosy is from the Americas.
b. Find the probability that a person with leprosy is from a high-income country.
c. Find the probability that a person with leprosy is from the Americas and a high-income country.
d. Find the probability that a person with leprosy is from a high-income country, given they are from the Americas.
e. Find the probability that a person with leprosy is from a low-income country.
f. Find the probability that a person with leprosy is from Africa.
g. Find the probability that a person with leprosy is from Africa and a low-income country.
h. Find the probability that a person with leprosy is from Africa, given they are from a low-income country.
i. Are the events that a person with leprosy is from “Africa” and “low-income country” independent events? Why or why not?
j. Are the events that a person with leprosy is from “Americas” and “high-income country” independent events? Why or why not?

Solution

a. There are 36817 cases of leprosy in the Americas out of 222,545 cases worldwide. So,

$$P(\text{ Americas }) = \frac{36817}{222545} \approx 0.165$$

There is about a 16.5% chance that a person with leprosy lives in a country in the Americas.

b. There are 264 cases of leprosy in high-income countries out of 222,545 cases worldwide. So,

$$P(\text{ high-income }) = \frac{264}{222545} \approx 0.0001$$

There is about a 0.1% chance that a person with leprosy lives in a high-income country.

c. There are 174 cases of leprosy in countries in a high-income country in the Americas out the 222,545 cases worldwide. So,

$$P(\text{ Americas and high-income }) = \frac{174}{222545} 0.0008$$

There is about a 0.08% chance that a person with leprosy lives in a high-income country in the Americas.

d. In this case you know that the person is in the Americas. You don't need to consider people from Easter Mediterranean, Europe, Western Pacific, Africa, and South-east Asia. You only need to look at the row with Americas at the start. In that row, look to see how many leprosy cases there are from a high-income country. There are 174 countries out of the 36,817 leprosy cases in the Americas. So,

$$P(\text{ high-income } | \text{ Americas }) = \frac{174}{36817} \approx 0.0047$$

There is 0.47% chance that a person with leprosy is from a high-income country given that they are from the Americas.

e. There are 27,923 cases of leprosy in low-income countries out of the 222,545 leprosy cases worldwide. So,

$$P(\text{ low-income }) = \frac{27923}{222545} \approx 0.125$$

There is a 12.5% chance that a person with leprosy is from a low-income country.

f. There are 17,953 cases of leprosy in Africa out of 222,545 leprosy cases worldwide. So,

$$P(\text{ Africa }) = \frac{17953}{222545} \approx 0.081$$

There is an 8.1% chance that a person with leprosy is from Africa.

g. There are 15,928 cases of leprosy in low-income countries in Africa out of all the 222,545 leprosy cases worldwide. So,

$$P(\text{ Africa and low-income }) = \frac{15928}{222545} \approx 0.072$$

There is a 7.2% chance that a person with leprosy is from a low-income country in Africa.

h. In this case you know that the person with leprosy is from low-income country. You don't need to include the high income, upper-middle income, and lowermiddle income country. You only need to consider the column headed by lowincome. In that column, there are 15,928 cases of leprosy in Africa out of the 27,923 cases of leprosy in low-income countries. So,

$$P(\text{ Africa } | \text{ low-income }) = \frac{15928}{27923} \approx 0.570$$

There is a 57.0% chance that a person with leprosy is from Africa, given that they are from a low-income country.

i. In order for these events to be independent, either $P(\text{ Africa } | \text{ low-income }) = P(\text{ Africa })$ or $P(\text{ low-income } | \text{ Africa }) = P(\text{ low-income })$ have to be true. Part (h) showed $P(\text{ Africa } | \text{ low-income }) \approx 0.570$ and part (f) showed $P(\text{ Africa }) \approx 0.081$. Since these are not equal, then these two events are dependent.

j. In order for these events to be independent, either $P(\text{ Americas } | \text{ high-income }) = P(\text{ Americas })$ or $P(\text{ high-income } | \text{ Americas }) = P(\text{ high-income })$ have to be true. Part (d) showed $P(\text{ high-income } | \text{ Americas }) \approx 0.0047$ and part (b) showed $P(\text{ high-income }) \approx 0.001$. Since these are not equal, then these two events are dependent.

A big deal has been made about the difference between dependent and independent events while calculating the probability of *and* compound events. You must multiply the probability of the first event with the conditional probability of the second event.

Why do you care? You need to calculate probabilities when you are performing sampling, as you will learn later. But here is a simplification that can make the calculations a lot easier: when the sample size is very small compared to the population size, you can assume that the conditional probabilities just don't change very much over the sample.

For example, consider acceptance sampling. Suppose there is a big population of parts delivered to you factory, say 12,000 parts. Suppose there are 85 defective parts in the population. You decide to randomly select ten parts, and reject the shipment. What is the probability of rejecting the shipment?

There are many different ways you could reject the shipment. For example, maybe the first three parts are good, one is bad, and the rest are good. Or all ten parts could be bad, or maybe the first five. So many ways to reject! But there is only ***one*** way that you'd accept the shipment: if ***all ten*** parts are good. That would happen if the first part is good, **and** the second part is good, **and** the third part is good, and so on. Since the probability of the second part being good is (slightly) dependent on whether the first part was good, technically you should take this into consideration when you calculate the probability that all ten are good.

The probability of getting the first sampled part good is $\frac{12000-85}{12000}=\frac{11915}{12000}$. So the probability that all ten being good is $\frac{11915}{12000} * \frac{11914}{11999} * \frac{11913}{11998} * ... * \frac{11906}{11991} \approx 93.1357\%$. If instead you assume that the probability doesn't change much, you get $\left(\frac{11915}{12000}\right)^{10} \approx 93.1382\%$. So as you can see, there is not much difference. So here is the rule: if the sample is very small compared to the size of the population, then you can assume that the probabilities are independent, even though they aren't technically. By the way, the probability of rejecting the shipment is $1-0.9314=0.0686=6.86\%$.

Homework

Exercise 4.3.1

1. Are owning a refrigerator and owning a car independent events? Why or why not?
2. Are owning a computer or tablet and paying for Internet service independent events? Why or why not?
3. Are passing your statistics class and passing your biology class independent events? Why or why not?
4. Are owning a bike and owning a car independent events? Why or why not?
5. An experiment is picking a card from a fair deck.
   a. What is the probability of picking a Jack given that the card is a face card?
   b. What is the probability of picking a heart given that the card is a three?
   c. What is the probability of picking a red card given that the card is an ace?
   d. Are the events Jack and face card independent events? Why or why not?
   e. Are the events red card and ace independent events? Why or why not?
6. An experiment is rolling two dice.
   a. What is the probability that the sum is 6 given that the first die is a 5?
   b. What is the probability that the first die is a 3 given that the sum is 11?
   c. What is the probability that the sum is 7 given that the fist die is a 2?
   d. Are the two events sum of 6 and first die is a 5 independent events? Why or why not?
   e. Are the two events sum of 7 and first die is a 2 independent events? Why or why not?
7. You flip a coin four times. What is the probability that all four of them are heads?
8. You flip a coin six times. What is the probability that all six of them are heads?
9. You pick three cards from a deck with replacing the card each time before picking the next card. What is the probability that all three cards are kings?
10. You pick three cards from a deck without replacing a card before picking the next card. What is the probability that all three cards are kings?
11. The number of people who survived the Titanic based on class and sex is in Example 4.3.2 ("Encyclopedia Titanica," 2013). Suppose a person is picked at random from the survivors.

Class	Sex		Total
	Female	Male	
1st	134	59	193
2nd	94	25	119
3rd	80	58	138
Total	308	142	450

Table 4.3.2: *Surviving the Titanic*

a. What is the probability that a survivor was female?
b. What is the probability that a survivor was in the 1st class?
c. What is the probability that a survivor was a female given that the person was in 1st class?
d. What is the probability that a survivor was a female and in the 1st class?
e. What is the probability that a survivor was a female or in the 1st class?
f. Are the events survivor is a female and survivor is in 1st class mutually exclusive? Why or why not?
g. Are the events survivor is a female and survivor is in 1st class independent? Why or why not?

12. Researchers watched groups of dolphins off the coast of Ireland in 1998 to determine what activities the dolphins partake in at certain times of the day ("Activities of dolphin," 2013). The numbers in Example 4.3.3 represent the number of groups of dolphins that were partaking in an activity at certain times of days.

Activity	Period				Total
	Morning	Noon	Afternoon	Evening	
Travel	6	6	14	13	39
Feed	28	4	0	56	88
Social	38	5	9	10	62
Total	72	15	23	79	189

Table 4.3.3: *Dolphin Activity*

a. What is the probability that a dolphin group is partaking in travel?
b. What is the probability that a dolphin group is around in the morning?
c. What is the probability that a dolphin group is partaking in travel given that it is morning?
d. What is the probability that a dolphin group is around in the morning given that it is partaking in socializing?
e. What is the probability that a dolphin group is around in the afternoon given that it is partaking in feeding?
f. What is the probability that a dolphin group is around in the afternoon and is partaking in feeding?
g. What is the probability that a dolphin group is around in the afternoon or is partaking in feeding?
h. Are the events dolphin group around in the afternoon and dolphin group feeding mutually exclusive events? Why or why not?
i. Are the events dolphin group around in the morning and dolphin group partaking in travel independent events? Why or why not?

Answer

1. Independent, see solutions

3. Dependent, see solutions

5. a. P(Jack/face card) = 0.333, b. P(heart/card a 3) = 0.25, c. P(red card/ace) = 0.50, d. not independent, see solutions, e. independent, see solutions

7. 0.0625

9. 4.55×10^{-4}

11. a. P(female) = 0.684, b. P(1st class) = 0.429, c. P(female/1st class) = 0.694, d. P(female and 1st class) = 0.298, e. P(female or 1st class) = 0.816, f. No, see solutions, g. Dependent, see solutions

This page titled 4.3: Conditional Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

4.4: Counting Techniques

There are times when the sample space or event space are very large, that it isn't feasible to write it out. In that case, it helps to have mathematical tools for counting the size of the sample space and event space. These tools are known as counting techniques.

Definition 4.4.1

Multiplication Rule in Counting Techniques

If task 1 can be done m_1 ways, task 2 can be done m_2 ways, and so forth to task *n* being done m_n ways. Then the number of ways to do task 1, 2,…, *n* together would be $m_1^* m_2^* \overset{\Delta^*}{r} m_n$.

Example 4.4.1 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?

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

$$\underset{\text{salad}}{3}\ \underset{\text{main}}{8}\ \underset{\text{dessert}}{5} = 120 \text{ different meals}$$

Example 4.4.2 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

There are three tasks that 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 ways, since the first and second task took two of the letters. There are

$$\underset{\text{first letter}}{3} * \underset{\text{second letter}}{2} * \underset{\text{third letter}}{1}$$

Which is

$$3^*2^*1 = 6$$

You can also look at this in a tree diagram:

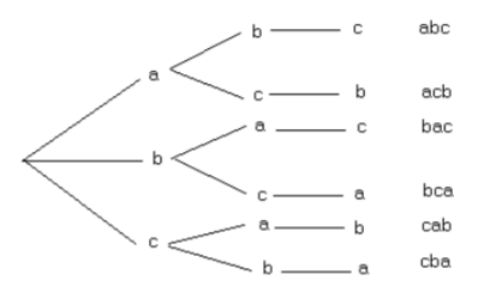


Figure 4.4.1: Tree diagram

So, there are 6 different "words."

In Example 4.4.2, 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 4.4.2

Factorial

$$n! = n(n-1)(n-2)\cdots(3)(2)(1)$$

As an example:

$$5! = 5*4*3*2*1 = 120$$

$$8! = 8*7*6*5*4*3*2*1 = 40320$$

0 factorial is defined to be 0!=1 and 1 factorial is defined to be 1!=1.

Sometimes you are trying to select *r* objects from *n* total objects. The number of ways to do this depends on if the order you choose the *r* objects matters or if it doesn't. As an example if you are trying to call a person on the phone, you have to have their number in the right order. Otherwise, you call someone you didn't mean to. In this case, the order of the numbers matters. If however you were picking random numbers for the lottery, it doesn't matter which number you pick first. As long as you have the same numbers that the lottery people pick, you win. In this case the order doesn't matter. A **permutation** is an arrangement of items with a specific order. You use permutations to count items when the order matters. When the order doesn't matter you use combinations. A **combination** is an arrangement of items when order is not important. When you do a counting problem, the first thing you should ask yourself is "does order matter?"

Definition 4.4.3

Permutation Formula

Picking *r* objects from *n* total objects when order matters

$${}_nP_r = \frac{n!}{(n-r)!}$$

Definition 4.4.4

Combination Formula

Picking *r* objects from *n* total objects when order doesn't matter

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Example 4.4.3 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.

$${}_{15}P_3 = \frac{15!}{(15-3)!} = \frac{15!}{12!} = 2730$$

Example 4.4.4 calculating the number of ways

Suppose you want to pick 7 people out of 20 people to take part in a survey. How many ways can you do this?

Solution

In this case the order doesn't matter, since you just want 7 people. This is a combination with n=20 and r=7.

$${}_{20}C_7 = \frac{20!}{7!(20-7)!} = \frac{20!}{7!13!} = 77520$$

Most calculators have a factorial button on them, and many have the combination and permutation functions also. R has a combination command.

Homework

Exercise 4.4.1

1. You are going to a benefit dinner, and need to decide before the dinner what you want for salad, main dish, and dessert. You have 2 different salads to choose from, 3 main dishes, and 5 desserts. How many different meals are available?
2. How many different phone numbers are possible in the area code 928?
3. You are opening a T-shirt store. You can have long sleeves or short sleeves, three different colors, five different designs, and four different sizes. How many different shirts can you make?
4. The California license plate has one number followed by three letters followed by three numbers. How many different license plates are there?
5. Find ${}_9P_4$
6. Find ${}_{10}P_6$
7. Find ${}_{10}P_5$
8. Find ${}_{20}P_4$
9. You have a group of twelve people. You need to pick a president, treasurer, and secretary from the twelve. How many different ways can you do this?
10. A baseball team has a 25-man roster. A batting order has nine people. How many different batting orders are there?
11. An urn contains five red balls, seven yellow balls, and eight white balls. How many different ways can you pick two red balls?
12. How many ways can you choose seven people from a group of twenty?

Answer

1. 30 meals

3. 120 shirts

5. 3024

7. 252

9. 1320

11. 10

Data sources

Aboriginal deaths in custody. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/custody.html

Activities of dolphin groups. (2013, September 26). Retrieved from http://www.statsci.org/data/general/dolpacti.html

Car preferences. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/carprefs.html

Encyclopedia Titanica. (2013, November 09). Retrieved from www.encyclopediatitanica.org/

Leprosy: Number of reported cases by country. (2013, September 04). Retrieved from http://apps.who.int/gho/data/node.main.A1639

Madison, J. (2013, October 15). *M&M's color distribution analysis.* Retrieved from http://joshmadison.com/2007/12/02/mm...tion-analysis/

This page titled 4.4: Counting Techniques is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

5: Discrete Probability Distributions

5.1: Basics of Probability Distributions

5.2: Binomial Probability Distribution

5.3: Mean and Standard Deviation of Binomial Distribution

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Basics of Probability Distributions

As a reminder, a variable or what will be called the random variable from now on, is represented by the letter *x* and it represents a quantitative (numerical) variable that is measured or observed in an experiment.

Also remember 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. **Continuous** data can take on any value in a range. Discrete data usually arises from counting while continuous data usually arises from measuring.

Examples of each

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. This is something you count.

If you have a variable, and can find a probability associated with that variable, it is called a **random variable**. In many cases the random variable is what you are measuring, but when it comes to discrete random variables, it is usually what you are counting. So for the example of how tall is a plant given a new fertilizer, the random variable is the height of the plant given a new fertilizer. For the example of how many fleas are on prairie dogs in a colony, the random variable is the number of fleas on a prairie dog in a colony.

Now suppose you put all the values of the random variable together with the probability that that random variable would occur. You could then have a distribution like before, 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. 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. 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 = number of people in a household. This is a discrete random variable, since you are counting the number of people in a household.

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 of the probabilities is one.

You can give a probability distribution in table form (as in Example 5.1.1) or as a graph. The graph looks like a histogram. A probability distribution is basically 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, where the x values are on the horizontal axis and are the x values of the classes (for the 7 or more category, just call it 7). The probabilities are on the vertical axis.

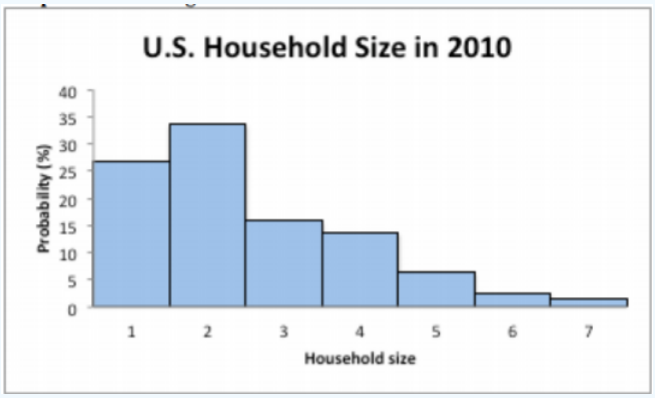


Figure 5.1.1: Histogram of Household Size from US Census of 2010

Notice this graph is skewed right.

Just as with any data set, you can calculate the mean and standard deviation. In problems involving a probability distribution function (pdf), you consider the probability distribution the population even though the pdf in most cases come 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 basically a population, the mean and standard deviation that are calculated are actually 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.

Note

The mean can be thought of as the **expected value**. It is the value you expect to get if the trials were repeated 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 $\sigma^2 = \sum (x-\mu)^2 P(x)$

The standard deviation is $\sigma = \sqrt{\sum (x-\mu)^2 P(x)}$

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
e. Using R to calculate the mean

Solution

State random variable:

x= number of people in a household

a. To find the mean it is easier to just use a table as shown below. Consider the category 7 or more to just be 7. The formula for the mean says to multiply the *x* value by the *P(x)* value, so add a row into the table for this calculation. Also convert all *P(x)* to decimal form.

Table 5.1.4: Calculating the Mean for a Discrete PDF

x	1	2	3	4	5	6	7
P(x)	0.267	0.336	0.158	0.137	0.063	0.024	0.015
xP(x)	0.267	0.672	0.474	0.548	0.315	0.144	0.098

Now add up the new row and you get the answer 2.525. This is the mean or the expected value, 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 try to 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.

Table 5.1.5: Calculating the Variance for a Discrete PDF

x	1	2	3	4	5	6	7
P(x)	0.267	0.336	0.158	0.137	0.063	0.024	0.015
$x-\mu$	-1.525	-0.525	0.475	1.475	2.475	3.475	4.475
$(x-\mu)^2$	2.3256	0.2756	0.2256	2.1756	6.1256	12.0756	20.0256
$(x-\mu)^2 P(x)$	0.6209	0.0926	0.0356	0.2981	0.3859	0.2898	0.3004

Now add up the last row to find the variance, $\sigma^2 = 2.02375 \text{ people}^2$. (Note: try not to round your numbers too much so you aren't creating rounding error 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 *Figure* 5.1.1.

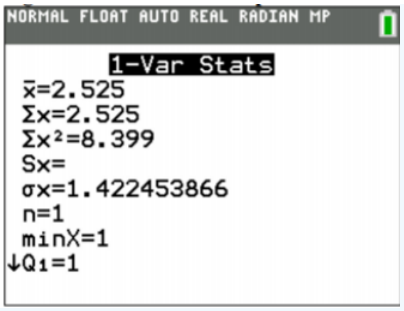


Figure 5.1.1: TI-83/84 Output

The mean is 2.525 people and the standard deviation is 1.422 people.

e. The command would be weighted.mean(x, p). So for this example, the process would look like:

x<-c(1, 2, 3, 4, 5, 6, 7)
p<-c(0.267, 0.336, 0.158, 0.137, 0.063, 0.024, 0.015)
weighted.mean(x, p)

Output:
[1] 2.525

So the mean is 2.525.

To find the standard deviation, you would need to program the process into R. So it is easier to just do it using the formula.

Example 5.1.4 Calculating the expected value

In the Arizona lottery called Pick 3, a player pays $1 and then picks a three-digit number. If those three numbers are picked in that specific order the person wins $500. What is the expected value in this game?

Solution

To find the expected value, you need to first create the probability distribution. In this case, the random variable x = winnings. If you pick the right numbers in the right order, then you win $500, but you paid $1 to play, so you actually win $499. If you didn't pick the right numbers, you lose the $1, the x value is -$1. You also need the probability of winning and losing. Since you are picking a three-digit number, and for each digit there are 10 numbers you can pick with each independent of the others, you can use the multiplication rule. To win, you have to pick the right numbers in the right order. The first digit, you pick 1 number out of 10, the second digit you pick 1 number out of 10, and the third digit you pick 1 number out of 10. The probability of picking the right number in the right order is $\frac{1}{10} * \frac{1}{10} * \frac{1}{10} = \frac{1}{1000} = 0.001$. The probability of losing (not winning) would be $1 - \frac{1}{1000} = \frac{999}{1000} = 0.999$. Putting this information into a table will help to calculate the expected value.

Table 5.1.6: Finding Expected Value

Win or lose	x	$P(x)$	$xP(x)$
Win	$499	0.001	$0.499
Lose	-$1	0.999	-$0.999

Now add the two values together and you have the expected value. It is $\$0.499 + (-\$0.999) = -\$0.50$. In the long run, you will expect to lose $0.50. Since the expected value is not 0, then this game is not fair. Since you lose money, Arizona makes money, which is why they have the lottery.

The reason probability is studied in statistics is to help in making decisions in inferential statistics. To understand how that is done the concept of a rare event is needed.

Definition 5.1.1: Rare Event Rule for Inferential Statistics

If, under a given assumption, the probability of a particular observed event is extremely small, then you can conclude that the assumption is probably not correct.

An example of this is suppose you roll an assumed fair die 1000 times and get a six 600 times, when you should have only rolled a six around 160 times, then you should believe that your assumption about it being a fair die is untrue.

Determining if an event is unusual

If you are looking at a value of x for a discrete variable, and the *P*(the variable has a value of x or more) < 0.05, then you can consider the x an unusually high value. Another way to think of this is if the probability of getting such a high value is less than 0.05, then the event of getting the value x is unusual.

Similarly, if the *P*(the variable has a value of x or less) < 0.05, then you can consider this an unusually low value. Another way to think of this is if the probability of getting a value as small as x is less than 0.05, then the event x is considered unusual.

Why is it "x or more" or "x or less" instead of just "x" when you are determining if an event is unusual? Consider this example: you and your friend go out to lunch every day. Instead of Going Dutch (each paying for their own lunch), you decide to flip a coin, and the loser pays for both. Your friend seems to be winning more often than you'd expect, so you want to determine if this is unusual before you decide to change how you pay for lunch (or accuse your friend of cheating). The process for how to calculate these probabilities will be presented in the next section on the binomial distribution. If your friend won 6 out of 10 lunches, the probability of that happening turns out to be about 20.5%, not unusual. The probability of winning 6 or more is about 37.7%. But what happens if your friend won 501 out of 1,000 lunches? That doesn't seem so unlikely! The probability of winning 501 or more lunches is about 47.8%, and that is consistent with your hunch that this isn't so unusual. But the probability of winning exactly 501 lunches is much less, only about 2.5%. That is why the probability of getting exactly that value is not the right question to ask: you should ask the probability of getting that value or more (or that value or less on the other side).

The value 0.05 will be explained later, and it is not the only value you can use.

Example 5.1.5 is the event unusual

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 5.1.7: 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. Is it unusual for a household to have six people in the family?
b. If you did come upon many families that had six people in the family, what would you think?
c. Is it unusual for a household to have four people in the family?
d. If you did come upon a family that has four people in it, what would you think?

Solution

State random variable:

x= number of people in a household

a. To determine this, you need to look at probabilities. However, you cannot just look at the probability of six people. You need to look at the probability of *x* being six or more people or the probability of x being six or less people. The

$$\begin{aligned} P(x \leq 6) &= P(x=1)+P(x=2)+P(x=3)+P(x=4)+P(x=5)+P(x=6) \\ &= 26.7\%+33.6\%+15.8\%+13.7\%+6.3\%+2.4\% \\ &= 98.5\% \end{aligned}$$

Since this probability is more than 5%, then six is not an unusually low value. The

$$\begin{aligned} P(x \geq 6) &= P(x = 6) + P(x \geq 7) \\ &= 2.4\% + 1.5\% \\ &= 3.9\% \end{aligned}$$

Since this probability is less than 5%, then six is an unusually high value. It is unusual for a household to have six people in the family.

b. Since it is unusual for a family to have six people in it, then you may think that either the size of families is increasing from what it was or that you are in a location where families are larger than in other locations.

c. To determine this, you need to look at probabilities. Again, look at the probability of x being four or more or the probability of x being four or less. The

$$\begin{aligned} P(x \geq 4) &= P(x = 4) + P(x = 5) + P(x = 6) + P(x = 7) \\ &= 13.7\% + 6.3\% + 2.4\% + 1.5\% \\ &= 23.9\% \end{aligned}$$

Since this probability is more than 5%, four is not an unusually high value. The

$$\begin{aligned} P(x \leq 4) &= P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) \\ &= 26.7\% + 33.6\% + 15.8\% + 13.7\% \\ &= 89.8\% \end{aligned}$$

Since this probability is more than 5%, four is not an unusually low value. Thus, four is not an unusual size of a family.

d. Since it is not unusual for a family to have four members, then you would not think anything is amiss.

Homework

Exercise 5.1.1

1. Eyeglassomatic manufactures eyeglasses for different retailers. The number of days it takes to fix defects in an eyeglass and the probability that it will take that number of days are in the table.

Number of days	Probabilities
1	24.9%
2	10.8%
3	9.1%
4	12.3%
5	13.3%
6	11.4%
7	7.0%
8	4.6%
9	1.9%
10	1.3%
11	1.0%
12	0.8%
13	0.6%
14	0.4%
15	0.2%
16	0.2%
17	0.1%

18	0.1%

Table 5.1.8: *Number of Days to Fix Defects*

a. State the random variable.
b. Draw a histogram of the number of days to fix defects
c. Find the mean number of days to fix defects.
d. Find the variance for the number of days to fix defects.
e. Find the standard deviation for the number of days to fix defects.
f. Find probability that a lens will take at least 16 days to make a fix the defect.
g. Is it unusual for a lens to take 16 days to fix a defect?
h. If it does take 16 days for eyeglasses to be repaired, what would you think?

2. Suppose you have an experiment where you flip a coin three times. You then count the number of heads.
   a. State the random variable.
   b. Write the probability distribution for the number of heads.
   c. Draw a histogram for the number of heads.
   d. Find the mean number of heads.
   e. Find the variance for the number of heads.
   f. Find the standard deviation for the number of heads.
   g. Find the probability of having two or more number of heads.
   h. Is it unusual for to flip two heads?
3. The Ohio lottery has a game called Pick 4 where a player pays $1 and picks a four-digit number. If the four numbers come up in the order you picked, then you win $2,500. What is your expected value?
4. An LG Dishwasher, which costs $800, has a 20% chance of needing to be replaced in the first 2 years of purchase. A two-year extended warrantee costs $112.10 on a dishwasher. What is the expected value of the extended warranty assuming it is replaced in the first 2 years?

Answer

1. a. See solutions, b. See solutions, c. 4.175 days, d. 8.414375 days^2, e. 2.901 days, f. 0.004, g. See solutions, h. See solutions

3. -$0.75

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

5.2: Binomial Probability Distribution

Section 5.1 introduced the concept of a probability distribution. The focus of the section was on discrete **probability distributions** (pdf). To find the pdf for a situation, you usually needed to actually conduct the experiment and collect data. Then you can calculate the experimental probabilities. Normally you cannot calculate the theoretical probabilities instead. However, there are certain types of experiment that allow you to calculate 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 a success and a failure.
4. The probability of a success doesn't change from trial to trial, where p = probability of success and *q* = probability of failure, *q* = 1-*p*.

If you know you have a binomial experiment, then you can calculate binomial probabilities. 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.

Example 5.2.1: Deriving the Binomial Probability Formula

Suppose you are given a 3 question multiple-choice test. Each question has 4 responses and only one is correct. Suppose you want to find the probability that you can just guess at the answers and get 2 questions right. (Teachers do this all the time when they make up a multiple-choice test to see if students can still pass without studying. In most cases the students can't.) To help with the idea that you are going to guess, suppose the test is in Martian.

a. What is the random variable?
b. Is this a binomial experiment?
c. What is the probability of getting 2 questions right?
d. What is the probability of getting zero right, one right, and all three right?

Solution

a. x = number of correct answers

b.

1. There are 3 questions, and each question is a trial, so there are a fixed number of trials. In this case, *n* = 3.
2. Getting the first question right has no affect on getting the second or third question right, thus the trials are independent.
3. Either you get the question right or you get it wrong, so there are only two outcomes. In this case, the success is getting the question right.
4. The probability of getting a question right is one out of four. This is the same for every trial since each question has 4 responses. In this case, $p = \frac{1}{4}$ and $q = 1 - \frac{1}{4} = \frac{3}{4}$

This is a binomial experiment, since all of the properties are met.

c. To answer this question, start with the sample space. SS = {RRR, RRW, RWR, WRR, WWR, WRW, RWW, WWW}, where RRW means you get the first question right, the second question right, and the third question wrong. The same is similar for the other outcomes.

Now the event space for getting 2 right is {RRW, RWR, WRR}. What you did in chapter four was just to find three divided by eight. However, this would not be right in this case. That is because the probability of getting a question right is different from getting a question wrong. What else can you do?

Look at just P(RRW) for the moment. Again, that means P(RRW) = P(R on 1st, R on 2nd, and W on 3rd)

Since the trials are independent, then P(RRW) = P(R on 1st, R on 2nd, and W on 3rd) = P(R on 1st) * P(R on 2nd) * P(W on 3rd)

Just multiply p * p * q

$$P(\text{RRW}) = \frac{1}{4} * \frac{1}{4} * \frac{3}{4} = \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1$$

The same is true for P(RWR) and P(WRR). To find the probability of 2 correct answers, just add these three probabilities together. You get

$$\begin{aligned} P(2 \text{ correct answers }) &= P(\text{RRW}) + P(\text{RWR}) + P(\text{WRR}) \\ &= \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 + \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 + \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 \\ &= 3\left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1 \end{aligned}$$

d. You could go through the same argument that you did above and come up with the following:

Table 5.2.1: Binomial pattern

r right	P(r right)
0 right	$1^* \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^3$
1 right	$3^* \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^2$
2 right	$3 * \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^1$
3 right	$1^* \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^0$

Hopefully you see the pattern that results. You can now write the general formula for the probabilities for a Binomial experiment

First, the random variable in a binomial experiment is *x* = number of successes. Be careful, a success is not always a good thing. Sometimes a success is something that is bad, like finding a defect. A success just means you observed the outcome you wanted to see happen.

Definition 5.2.1

Binomial Formula for the probability of *r* successes in *n* trials is

$$P(x = r) = {}_n C_r p^r q^{n-r} \text{ where } {}_n C_r = \frac{n!}{r!(n-r)!}$$

The ${}_n C_r$ is the number of combinations of *n* things taking *r* at a time. It is read "*n* choose *r*". Some other common notations for *n* choose *r* are $C_{n,r}$, and $\binom{n}{r}$. *n*! means you are multiplying $n^*(n-1)^*(n-2)^* \ldots {}^*2^*1$. As an example, $5! = 5^*4^*3^*2^*1 = 120$.

When solving problems, make sure you define your random variable and state what *n*, *p*, *q*, and *r* are. Without doing this, the problems are a great deal harder.

Example 5.2.2: Calculating Binomial Probabilities

When looking at a person's eye color, it turns out that 1% of people in the world has green eyes ("What percentage of," 2013). Consider a group of 20 people.

a. State the random variable.
b. Argue that this is a binomial experiment.
c. Find the probability that none have green eyes.
d. Find the probability that nine have green eyes.
e. Find the probability that at most three have green eyes.
f. Find the probability that at most two have green eyes.

g. Find the probability that at least four have green eyes.
h. In Europe, four people out of twenty have green eyes. Is this unusual? What does that tell you?

Solution

a. *x* = number of people with green eyes

b.

1. There are 20 people, and each person is a trial, so there are a fixed number of trials. In this case, $n = 20$.
2. 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.
3. Either a person has green eyes or they do not have green eyes, so there are only two outcomes. In this case, the success is a person has green eyes.
4. 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 three means that three is the highest value you will have. Find the probability of *x* is less than or equal to three.

$$\begin{aligned} P(x \leq 3) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) \\ &=_{20} C_0(0.01)^0(0.99)^{20} +_{20} C_1(0.01)^1(0.99)^{19} \\ &+_{20} C_2(0.01)^2(0.99)^{18} +_{20} C_3(0.01)^3(0.99)^{17} \\ &\approx 0.818 + 0.165 + 0.016 + 0.001 > 0.999 \end{aligned}$$

The reason the answer is written as being greater than 0.999 is because the answer is actually 0.9999573791, and when that is rounded to three decimal places you get 1. But 1 means that the event will happen, when in reality there is a slight chance that it won't happen. It is best to write the answer as greater than 0.999 to represent that the number is very close to 1, but isn't 1.

f.

$$\begin{aligned} P(x \leq 2) &= P(x = 0) + P(x = 1) + P(x = 2) \\ &=_{20} C_0(0.01)^0(0.99)^{20} +_{20} C_1(0.01)^1(0.99)^{19} +_{20} C_2(0.01)^2(0.99)^{18} \\ &\approx 0.818 + 0.165 + 0.016 \approx 0.999 \end{aligned}$$

g. At least four means four or more. Find the probability of x being greater than or equal to four. That would mean adding up all the probabilities from four to twenty. This would take a long time, so it is better to use the idea of complement. The complement of being greater than or equal to four is being less than four. That would mean being less than or equal to three. Part (e) has the answer for the probability of being less than or equal to three. Just subtract that number from 1.

$$P(x \geq 4) = 1 - P(x \leq 3) = 1 - 0.999 = 0.001$$

Actually the answer is less than 0.001, but it is fine to write it this way.

h. Since the probability of finding four or more people with green eyes is much less than 0.05, it is unusual to find four people out of twenty with green eyes. That should make you wonder if the proportion of people in Europe with green eyes is more than the 1% for the general population. If this is true, then you may want to ask why Europeans have a higher proportion of green-eyed people. That of course could lead to more questions.

The binomial formula is cumbersome to use, so you can find the probabilities by 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 $x < r$ or $x \leq r - 1$ are complementary events. Both binompdf and binomcdf commands are found in the DISTR menu. Using R, the commands are $P(x = r) =$ dbinom (r, n, p) and $P(x \leq r) =$ pbinom (r, n, p).

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 has 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 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.
f. Find the probability that at least four 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 or dbinom command on R. On the TI83/84, you go to the DISTR menu, select the binompdf, and then type into the parenthesis your n, p, and r 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 given the answer. (If you have the new software on the TI-84, the screen looks a bit different.)

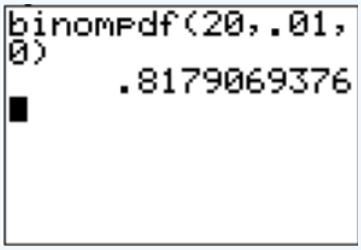


Figure 5.2.1: Calculator Results for binompdf

On R, the command would look like dbinom(0, 20, 0.01)

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 or dbinom(9,20,0.01) on R. Your answer is $P(x = 9) = 1.50 \times 10^{-13}$. (Remember when the calculator gives you $1.50E-13$ and R give you $1.50e-13$, this is how they display scientific notation.) The probability that out of twenty people, nine of them 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 and pbinom command in R. You use the command on the TI-83/84 of binomcdf(20, .01, 3) and the command on R of pbinom(3,20,0.01)

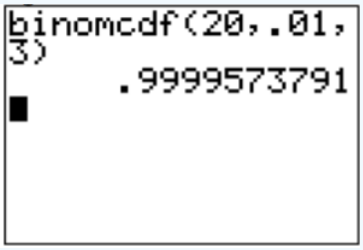


Figure 5.2.2: Calculator Results for binomcdf

Your answer is 0.99996. Thus there is a really 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 in reality 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 really good chance that at most two people in the room will have green eyes.

f. At least four means four or more. Find the probability of x being greater than or equal to four. That would mean adding up all the probabilities from four to twenty. This would take a long time, so it is better to use the idea of complement. The complement of being greater than or equal to four is being less than four. That would mean being less than or equal to three. Part (e) has the answer for the probability of being less than or equal to three. Just subtract that number from 1.

$P(x \geq 4) = 1 - P(x \leq 3) = 1 - 0.99996 = 0.00004$ You can also find this answer by doing the following on TI-83/84:

$P(x \geq 4) = 1 - P(x \leq 3) = 1 - \text{ binomcdf } (20, .01, 3) = 1 - 0.99996 = 0.00004$ on R:

$P(x \geq 4) = 1 - P(x \leq 3) = 1 - \text{ pbinom } (3, 20, .01) = 1 - 0.99996 = 0.0004$ Again, this is very unlikely to happen.

There are other technologies that will compute binomial probabilities.

Example 5.2.4 calculating binomial probabilities

According to the Center for Disease Control (CDC), about 1 in 88 children in the U.S. have been diagnosed with autism ("CDC-data and statistics,," 2013). Suppose you consider a group of 10 children.

a. State the random variable.
b. Argue that this is a binomial experiment.
c. Find the probability that none have autism.
d. Find the probability that 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 are 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 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:

$$P(x = 0) = \text{ binompdf } (10, 1 \div 88, 0) \approx 0.892$$

Using R:

$$P(x = 0) = \text{ pbinom } (0, 10, 1/88) \approx 0.892$$

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:

$$P(x = 7) = \text{ binompdf } (10, 1 \div 88, 7) \approx 2.84 \times 10^{-12}$$

Using R:

$$P(x = 7) = \text{dbinom}(7, 10, 1/88) \approx 2.84 \times 10^{-12}$$

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.

$$
\begin{aligned}
P(x \geq 5) &= 1 - P(x < 5) \\
&= 1 - P(x \leq 4) \\
&= 1 - \text{ binomcdf } (10, 1 \div 88, 4) \\
&\approx 1 - 0.9999999 = 0.000
\end{aligned}
$$

Using R:

To use R you need to use the complement.

$$
\begin{aligned}
P(x \geq 5) &= 1 - P(x < 5) \\
&= 1 - P(x \leq 4) \\
&= 1 - \text{ pbinom } (4, 10, 1/88) \\
&\approx 1 - 0.9999999 = 0.000
\end{aligned}
$$

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:

$$P(x \leq 2) = \text{ binomcdf } (10, 1 \div 88, 2) \approx 0.9998$$

Using R:

$$P(x \leq 2) = \text{ pbinom } (2, 10, 1/88) \approx 0.9998$$

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 actually more than 1/88.

Homework

Exercise 5.2.1

1. Suppose a random variable, *x*, arises from a binomial experiment. If *n* = 14, and *p* = 0.13, find the following probabilities using the binomial formula.
   a. *P* (x=5)
   b. *P* (x=8)
   c. *P* (x=12)
   d. $P(x \leq 4)$
   e. $P(x \geq 8)$
   f. $P(x \leq 12)$

2. Suppose a random variable, x, arises from a binomial experiment. If n = 22, and p = 0.85, find the following probabilities using the binomial formula.
   a. P (x=18)
   b. P (x=5)
   c. P (x=20)
   d. $P(x \leq 3)$
   e. $P(x \geq 18)$
   f. $P(x \leq 20)$
3. Suppose a random variable, x, arises from a binomial experiment. If n = 10, and p = 0.70, find the following probabilities using the binomial formula.
   a. P (x=2)
   b. P (x=8)
   c. P (x=7)
   d. $P(x \leq 3)$
   e. $P(x \geq 7)$
   f. $P(x \leq 4)$
4. Suppose a random variable, x, arises from a binomial experiment. If n = 6, and p = 0.30, find the following probabilities using the binomial formula.
   a. P (x=1)
   b. P (x=5)
   c. P (x=3)
   d. $P(x \leq 3)$
   e. $P(x \geq 5)$
   f. $P(x \leq 4)$
5. Suppose a random variable, x, arises from a binomial experiment. If n = 17, and p = 0.63, find the following probabilities using the binomial formula.
   a. P (x=8)
   b. P (x=15)
   c. P (x=14)
   d. $P(x \leq 12)$
   e. $P(x \geq 10)$
   f. $P(x \leq 7)$
6. Suppose a random variable, x, arises from a binomial experiment. If n = 23, and p = 0.22, find the following probabilities using the binomial formula.
   a. P (x=21)
   b. P (x=6)
   c. P (x=12)
   d. $P(x \leq 14)$
   e. $P(x \geq 17)$
   f. $P(x \leq 9)$
7. Approximately 10% of all people are left-handed ("11 little-known facts," 2013). Consider a grouping of fifteen people.
   a. State the random variable.
   b. Argue that this is a binomial experiment Find the probability that
   c. None are left-handed.
   d. Seven are left-handed.
   e. At least two are left-handed.
   f. At most three are left-handed.
   g. At least seven are left-handed.
   h. Seven of the last 15 U.S. Presidents were left-handed. Is this unusual? What does that tell you?

8. According to an article in the American Heart Association's publication Circulation, 24% of patients who had been hospitalized for an acute myocardial infarction did not fill their cardiac medication by the seventh day of being discharged (Ho, Bryson & Rumsfeld, 2009). Suppose there are twelve people who have been hospitalized for an acute myocardial infarction.
    a. State the random variable.
    b. Argue that this is a binomial experiment Find the probability that
    c. All filled their cardiac medication.
    d. Seven did not fill their cardiac medication.
    e. None filled their cardiac medication.
    f. At most two did not fill their cardiac medication.
    g. At least three did not fill their cardiac medication.
    h. At least ten did not fill their cardiac medication.
    i. Suppose of the next twelve patients discharged, ten did not fill their cardiac medication, would this be unusual? What does this tell you?
9. Eyeglassomatic manufactures eyeglasses for different retailers. In March 2010, they tested to see how many defective lenses they made, and there were 16.9% defective lenses due to scratches. Suppose Eyeglassomatic examined twenty eyeglasses.
    a. State the random variable.
    b. Argue that this is a binomial experiment Find the probability that
    c. None are scratched.
    d. All are scratched.
    e. At least three are scratched.
    f. At most five are scratched.
    g. At least ten are scratched.
    h. Is it unusual for ten lenses to be scratched? If it turns out that ten lenses out of twenty are scratched, what might that tell you about the manufacturing process?
10. The proportion of brown M&M's in a milk chocolate packet is approximately 14% (Madison, 2013). Suppose a package of M&M's typically contains 52 M&M's.
    a. State the random variable.
    b. Argue that this is a binomial experiment Find the probability that
    c. Six M&M's are brown.
    d. Twenty-five M&M's are brown.
    e. All of the M&M's are brown.
    f. Would it be unusual for a package to have only brown M&M's? If this were to happen, what would you think is the reason?

Answer

1. a. P(x=5) = 0.0212, b. P(x=8) = 1.062×10^{-4}, c. P(x=12) = 1.605×10^{-9}, d. $P(x \leq 4) = 0.973$, e. $P(x \geq 8) = 1.18 \times 10^{-4}$, f. $P(x \leq 12) = 0.99999$

3. a. $P(x = 2) = 0.0014$, b. $P(x = 8) = 0.2335$, c. $P(x = 7) = 0.2668$, d. $P(x \leq 3) = 0.0106$, e. $P(x \geq 7) = 0.6496$, f. $P(x \leq 4) = 0.0473$

5. a. $P(x = 8) = 0.0784$, b. $P(x = 15) = 0.0182$, c. $P(x = 14) = 0.0534$, d. $P(x \leq 12) = 0.8142$, e. $P(x \geq 10) = 0.7324$, f. $P(x \leq 7) = 0.0557$

7. a. See solutions, b. See solutions, c. P(x=0) = 0.2059, d. $P(x = 7) = 2.770 \times 10^{-4}$, e. $P(x \geq 2) = 0.4510$, f. $P(x \leq 3) = 0.944$, g.$P(x \geq 7) = 3.106 \times 10^{-4}$, h. See solutions

9. a. See solutions, b. See solutions, c. $P(x = 0) = 0.0247$, d. $P(x = 20) = 3.612 \times 10^{-16}$, e. $P(x \geq 3) = 0.6812$, f. $P(x \leq 5) = 0.8926$, g. $P(x \geq 10) = 6.711 \times 10^{-4}$, h. See solutions

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 Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Mean and Standard Deviation of Binomial Distribution

If you list all possible values of x in a Binomial distribution, you get the **Binomial Probability Distribution** (pdf). You can draw a histogram of the pdf and find the mean, variance, and standard deviation of it.

For a general discrete probability distribution, you can find the mean, the variance, and the standard deviation for a pdf using the general formulas

$$\mu = \sum xP(x), \sigma^2 = \sum (x-\mu)^2 P(x), \text{ and } \sigma = \sqrt{\sum (x-\mu)^2 P(x)}$$

These formulas are useful, but if you know the type of distribution, like Binomial, then you can find the mean and standard deviation using easier formulas. They are derived from the general formulas.

Note

For a Binomial distribution, μ, the expected number of successes, σ^2, the variance, and σ, the standard deviation for the number of success are given by the formulas:

$$\mu = np \quad \sigma^2 = npq \quad \sigma = \sqrt{npq}$$

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 has green eyes ("What percentage of," 2013). Consider a group of 20 people.

a. State the random variable.
b. Write the probability distribution.
c. Draw a histogram.
d. Find the mean.
e. Find the variance.
f. Find the standard deviation.

Solution

a. x = number of people who have green eyes

b. In this case you need to write each value of x and its corresponding probability. It is easiest to do this by using the binompdf command, but don't put in the r value. You may want to set your calculator to only three decimal places, so it is easier to see the values and you don't need much more precision than that. The command would look like $\text{binompdf}(20, .01)$. This produces the information in Example 5.3.1.

Table 5.3.1: Probability Distribution for Number of People with Green Eyes

x	P (x=r)
0	0.818
1	0.165
2	0.016
3	0.001
4	0.000
5	0.000
6	0.000
7	0.000
8	0.000

x	P (x=r)
9	0.000
10	0.000
⋮	⋮
20	0.000

Notice that after x = 4, the probability values are all 0.000. This just means they are really small numbers.

c. You can draw the histogram on the TI-83/84 or other technology. The graph would look like in *Figure* 5.3.1.

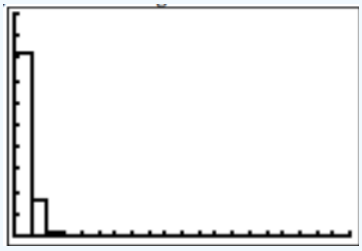


Figure 5.3.1: Histogram Created on TI-83/84

This graph is very skewed to the right.

d. 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.

e. Since this is a binomial, then you can use the formula $\sigma^2 = npq$.

$$q = 1 - 0.01 = 0.99$$

$$\sigma^2 = 20(0.01)(0.99) = 0.198 \text{ people}^2$$

f. Once you have the variance, you just take the square root of the variance to find the standard deviation.

$$\sigma = \sqrt{0.198} \approx 0.445$$

Homework

Exercise 5.3.1

1. Suppose a random variable, *x*, arises from a binomial experiment. Suppose *n* = 6, and *p* = 0.13.
   a. Write the probability distribution.
   b. Draw a histogram.
   c. Describe the shape of the histogram.
   d. Find the mean.
   e. Find the variance.
   f. Find the standard deviation.
2. Suppose a random variable, *x*, arises from a binomial experiment. Suppose *n* = 10, and *p* = 0.81.
   a. Write the probability distribution.
   b. Draw a histogram.
   c. Describe the shape of the histogram.
   d. Find the mean.
   e. Find the variance.
   f. Find the standard deviation.
3. Suppose a random variable, *x*, arises from a binomial experiment. Suppose *n* = 7, and *p* = 0.50.
   a. Write the probability distribution.
   b. Draw a histogram.

c. Describe the shape of the histogram.
d. Find the mean.
e. Find the variance.
f. Find the standard deviation.

4. Approximately 10% of all people are left-handed. Consider a grouping of fifteen people.
   a. State the random variable.
   b. Write the probability distribution.
   c. Draw a histogram.
   d. Describe the shape of the histogram.
   e. Find the mean.
   f. Find the variance.
   g. Find the standard deviation.
5. According to an article in the American Heart Association's publication Circulation, 24% of patients who had been hospitalized for an acute myocardial infarction did not fill their cardiac medication by the seventh day of being discharged (Ho, Bryson & Rumsfeld, 2009). Suppose there are twelve people who have been hospitalized for an acute myocardial infarction.
   a. State the random variable.
   b. Write the probability distribution.
   c. Draw a histogram.
   d. Describe the shape of the histogram.
   e. Find the mean.
   f. Find the variance.
   g. Find the standard deviation.
6. Eyeglassomatic manufactures eyeglasses for different retailers. In March 2010, they tested to see how many defective lenses they made, and there were 16.9% defective lenses due to scratches. Suppose Eyeglassomatic examined twenty eyeglasses.
   a. State the random variable.
   b. Write the probability distribution.
   c. Draw a histogram.
   d. Describe the shape of the histogram.
   e. Find the mean.
   f. Find the variance.
   g. Find the standard deviation.
7. The proportion of brown M&M's in a milk chocolate packet is approximately 14% (Madison, 2013). Suppose a package of M&M's typically contains 52 M&M's.
   a. State the random variable.
   b. Find the mean.
   c. Find the variance.
   d. Find the standard deviation.

Answer

1. a. See solutions, b. See solutions, c. Skewed right, d. 0.78, e. 0.6786, f. 0.8238

3. a. See solutions, b. See solutions, c. Symmetric, d. 3.5, e. 1.75, f. 1.3229

5. a. See solutions, b. See solutions, c. See solutions, d. Skewed right, e. 2.88, f. 2.1888, g. 1.479

7. a. See solutions, b. 7.28, c. 6.2608, d. 2.502

Data Sources: *11 little-known facts about left-handers*. (2013, October 21). Retrieved from www.huffingtonpost.com/2012/1...n_2005864.html

CDC-data and statistics, autism spectrum disorders - ncbdd. (2013, October 21). Retrieved from http://www.cdc.gov/ncbddd/autism/data.html

Ho, P. M., Bryson, C. L., & Rumsfeld, J. S. (2009). Medication adherence. *Circulation*, *119* (23), 3028-3035. Retrieved from http://circ.ahajournals.org/content/119/23/3028

Households by age of householder and size of household: 1990 to 2010. (2013, October 19). Retrieved from www.census.gov/compendia/stat...es/12s0062.pdf

Madison, J. (2013, October 15). *M&M's color distribution analysis*. Retrieved from http://joshmadison.com/2007/12/02/mm...tion-analysis/

What percentage of people have green eyes?. (2013, October 21). Retrieved from www.ask.com/question/what-per...ave-green-eyes

This page titled 5.3: Mean and Standard Deviation of Binomial Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

6: Continuous Probability Distributions

Chapter 5 dealt with probability distributions arising from **discrete** random variables. Mostly that chapter focused on the binomial experiment. There are many other experiments from discrete random variables that exist but are not covered in this book. This chapter deals with probability distributions that arise from **continuous** random variables. The focus of this chapter is a distribution known as the normal distribution, though realize that there are many other distributions that exist. A few others are examined in future chapters.

6.1: Uniform Distribution
6.2: Graphs of the Normal Distribution
6.3: Finding Probabilities for the Normal Distribution
6.4: Assessing Normality
6.5: Sampling Distribution and the Central Limit Theorem

This page titled 6: Continuous Probability Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

6.1: Uniform Distribution

If you have a situation where the probability is always the same, then this is known as a uniform distribution. An example would be waiting for a commuter train. The commuter trains on the Blue and Green Lines for the Regional Transit Authority (RTA) in Cleveland, OH, have a waiting time during peak hours of ten minutes ("2012 annual report," 2012). If you are waiting for a train, you have anywhere from zero minutes to ten minutes to wait. Your probability of having to wait any number of minutes in that interval is the same. This is a uniform distribution. The graph of this distribution is in *Figure* 6.1.1.



Figure

Suppose you want to know the probability that you will have to wait between five and ten minutes for the next train. You can look at the probability graphically such as in *Figure* 6.1.2.



Figure 6.1.2: Uniform Distribution with P (5<x<10)

How would you find this probability? Calculus says that the probability is the area under the curve. Notice that the shape of the shaded area is a rectangle, and the area of a rectangle is length times width. The length is $10 - 5 = 5$ and the width is 0.1. The probability is $P(5 < x < 10) = 0.1 * 5 = 0.5$, where and *x* is the waiting time during peak hours.

Example 6.1.1 finding probabilities in a uniform distribution

The commuter trains on the Blue and Green Lines for the Regional Transit Authority (RTA) in Cleveland, OH, have a waiting time during peak rush hour periods of ten minutes ("2012 annual report," 2012).

a. State the random variable.
b. Find the probability that you have to wait between four and six minutes for a train.
c. Find the probability that you have to wait between three and seven minutes for a train.
d. Find the probability that you have to wait between zero and ten minutes for a train.
e. Find the probability of waiting exactly five minutes.

Solution

a. *x* = waiting time during peak hours

b. $P(4 < x < 6) = (6 - 4)^{*}0.1 = 0.2$

c. $P(3 < x < 7) = (7 - 3) * 0.1 = 0.4$

d. $P(0 < x < 10) = (10 - 0) * 0.1 = 1.0$

e. Since this would be just one line, and the width of the line is 0, then the $P(x = 5) = 0 * 0.1 = 0$.

Notice that in Example 6.1.1d, the probability is equal to one. This is because the probability that was computed is the area under the entire curve. Just like in discrete probability distributions, where the total probability was one, the probability of the entire

curve is one. This is the reason that the height of the curve is 0.1. In general, the height of a uniform distribution that ranges between *a* and *b*, is $\frac{1}{b-a}$.

Homework

Exercise 6.1.1

1. The commuter trains on the Blue and Green Lines for the Regional Transit Authority (RTA) in Cleveland, OH, have a waiting time during peak rush hour periods of ten minutes ("2012 annual report," 2012).
   a. State the random variable.
   b. Find the probability of waiting between two and five minutes.
   c. Find the probability of waiting between seven and ten minutes.
   d. Find the probability of waiting eight minutes exactly.
2. The commuter trains on the Red Line for the Regional Transit Authority (RTA) in Cleveland, OH, have a waiting time during peak rush hour periods of eight minutes ("2012 annual report," 2012).
   a. State the random variable.
   b. Find the height of this uniform distribution.
   c. Find the probability of waiting between four and five minutes.
   d. Find the probability of waiting between three and eight minutes.
   e. Find the probability of waiting five minutes exactly.

Answer

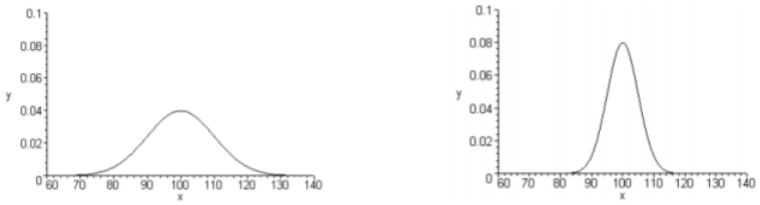
1. a. See solutions, b. $P(2 < x < 5) = 0.3$, c. $P(7 < x < 10) = 0.3$, d. $P(x = 8) = 0$

This page titled 6.1: Uniform Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

6.2: Graphs of the Normal Distribution

Many real life problems produce a histogram that is a symmetric, unimodal, and bellshaped 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* 6.2.1 shows two different normal curves drawn on the same scale. Both have $\mu = 100$ but the one on the left has a standard deviation of 10 and the one on the right has a standard deviation of 5. Notice that the larger standard deviation makes the graph wider (more spread out) and shorter.



Figures

Every normal curve has common features. These are detailed in *Figure* 6.2.2.

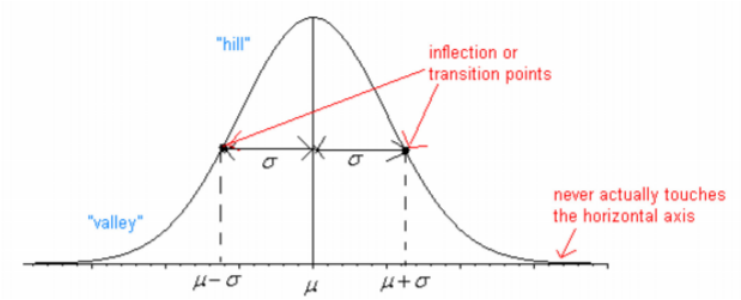


Figure of a Normal Curve

- The center, or the highest point, is at the population mean, μ.
- The transition points (inflection points) are the places where the curve changes from a “hill” to a “valley”. The distance from the mean to the transition point is one standard deviation, σ.
- The area under the whole curve is exactly 1. Therefore, the area under the half below or above the mean is 0.5.

The equation that creates this curve is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Just as in a discrete probability distribution, the object is to find the probability of an event occurring. However, unlike in a discrete probability distribution where the event can be a single value, in a continuous probability distribution the event must be a range. You are interested in finding the probability of *x* occurring in the range between *a* and *b*, or $P(a \leq x \leq b) = P(a < x < b)$. Calculus tells us that to find this you find the area under the curve above the interval from *a* to *b*.

$P(a \leq x \leq b) = P(a < x < b)$ is the area under the curve above the integral from *a* to *b*.

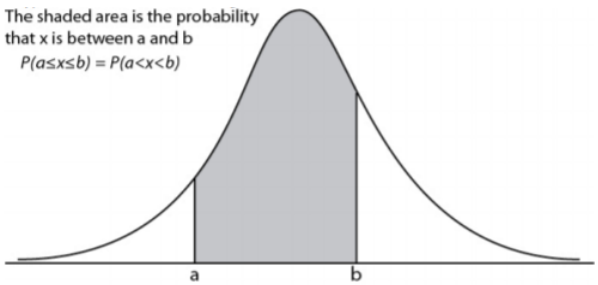


Figure 6.2.3: Probability of an Event

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 real rough estimates.

Definition 6.2.1: 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.

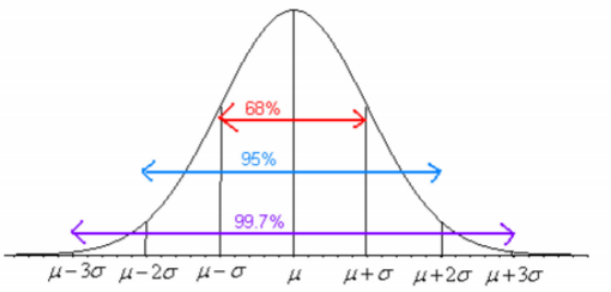


Figure 6.2.4: Empirical Rule

Be careful, there is still some area left over in each end. Remember, the maximum a probability can be is 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 $\mu \pm 3\sigma$.

This page titled 6.2: Graphs of the Normal Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

6.3: Finding Probabilities for the 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 6.3.1: z-score

$$z = \frac{x - \mu}{\sigma} \tag{6.3.1}$$

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 6.3.2: standard normal distribution

The **standard normal distribution**, z, has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$.

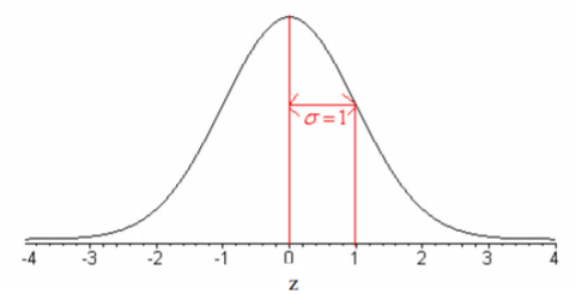


Figure 6.3.1: Standard Normal Curve

Luckily, these days technology can find probabilities for you without converting to the zscore 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 and with R. 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, standard deviation in that order and including the commas. The command on R to find the area to the left is pnorm(z-value or x-value, mean, standard deviation).

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).

a. State the random variable.
b. Find the probability of a pregnancy lasting more than 280 days.
c. Find the probability of a pregnancy lasting less than 250 days.
d. Find the probability that a pregnancy lasts between 265 and 280 days.
e. Find the length of pregnancy that 10% of all pregnancies last less than.
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 the center of this normal curve is 272.

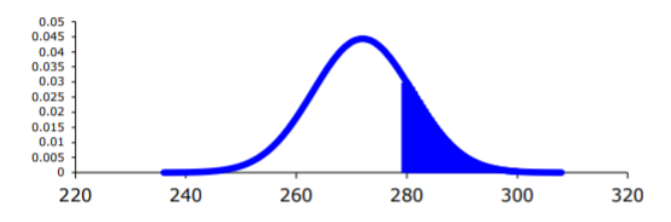


Figure for Example 6.3.1b

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:

$\text{normalcdf}(280, 1E99, 272, 9)$

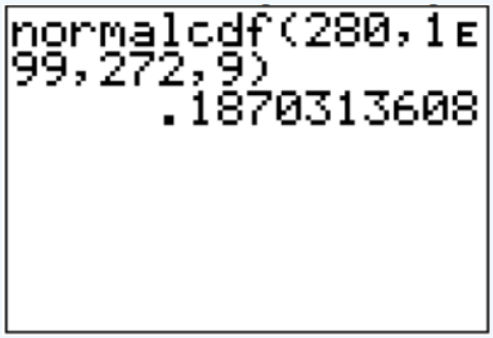


Figure 6.3.3: TI-83/84 Output for Example 6.3.1b

To find the probability on R, R always gives the probability to the left of the value. The total area under the curve is 1, so if you want the area to the right, then you find the area to the left and subtract from 1. The command looks like:

$1 - \text{ pnom } (280, 272, 9)$

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%.

c. First translate the statement into a mathematical statement.

$$P \text{ (x<250)}$$

Now, draw a picture. Remember the center of this normal curve is 272.

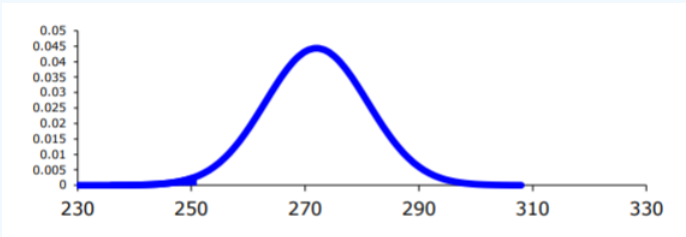


Figure for Example 6.3.1c

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.

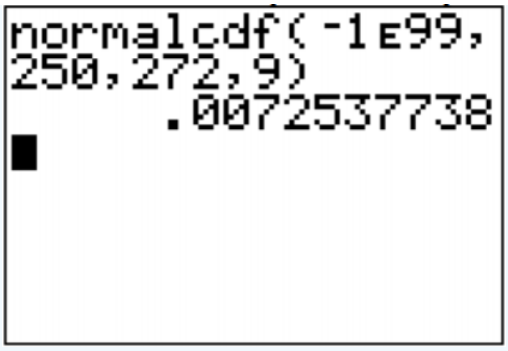


Figure 6.3.5: TI-83/84 Output for Example 6.3.1c

$$P(x < 250) = \text{ normalcdf } (-1E99, 250, 272, 9) = 0.0073$$

To find the probability on R, R always gives the probability to the left of the value. Looking at the figure, you can see the area you want is to the left. The command looks like:

$$P(x < 250) = \text{ pnorm } (250, 272, 9) = 0.0073$$

Thus 0.73% of all pregnancies last less than 250 days. This is unusual since the probability is less than 5%.

d. First translate the statement into a mathematical statement.

$$P\ (265<x<280)$$

Now, draw a picture. Remember the center of this normal curve is 272.

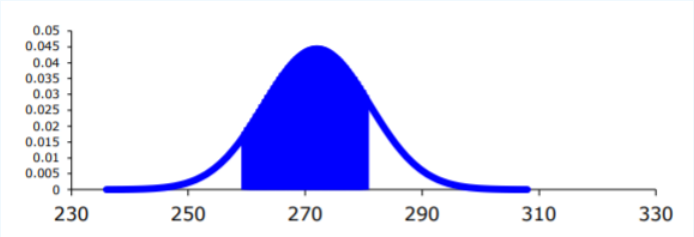


Figure for Example 6.3.1d

In this case, the lower limit is 265 and the upper limit is 280.

Using the calculator

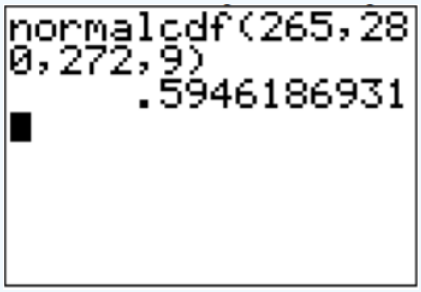


Figure 6.3.7: TI-83/84 Output for Example 6.3.1d

$$P(265 < x < 280) = \text{ normalcdf } (265, 280, 272, 9) = 0.595$$

To use R, you have to remember that R gives you the area to the left. So $P(x < 280) = \text{ pnom } (280, 272, 9)$ is the area to the left of 280 and $P(x < 265) = \text{ pnom } (265, 272, 9)$ is the area to the left of 265. So the area is between the two would be the bigger one minus the smaller one. So, $P(265 < x < 280) = \text{ pnorm } (280, 272, 9) - \text{ pnorm } (265, 272, 9) = 0.595$. Thus 59.5% of all pregnancies last between 265 and 280 days.

e. This problem is asking 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.

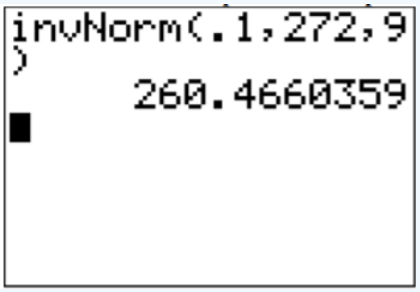


Figure 6.3.8: TI-83/84 Output for Example 6.3.1e

On R, the command is qnorm(area to the left, mean, standard deviation). For this example that would be qnorm(0.1, 272, 9)

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 actually became pregnant.

Example 6.3.2 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.

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 a 500 and a 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 the center of this normal curve is 514.

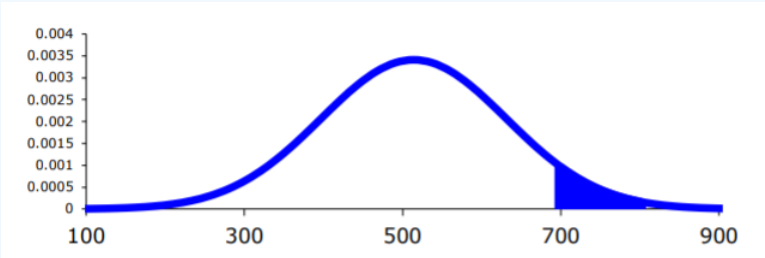


Figure for Example 6.3.2b

On TI-83/84: $P(x > 700) =$ normalcdf $(700, 1E99, 514, 117) \approx 0.056$

On R: $P(x > 700) = 1 -$ pnorm $(700, 514, 117) \approx 0.056$

There is a 5.6% chance that a person scored above a 700 on the mathematics SAT test. This is not unusual.

c. First translate the statement into a mathematical statement.

$$P\ (x<400)$$

Now, draw a picture. Remember the center of this normal curve is 514.

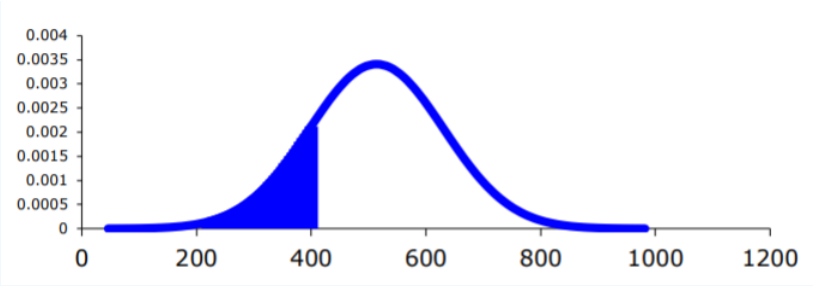


Figure for Example 6.3.2c

On TI-83/84: $P(x < 400) =$ normalcdf $(-1E99, 400, 514, 117) \approx 0.165$

On R: $P(x < 400) = \text{pnorm}(400, 514, 117) \approx 0.165$

So, there is a 16.5% chance that a person scores less than a 400 on the mathematics part of the SAT.

d. First translate the statement into a mathematical statement.

$$P\ (500<x<650)$$

Now, draw a picture. Remember the center of this normal curve is 514.

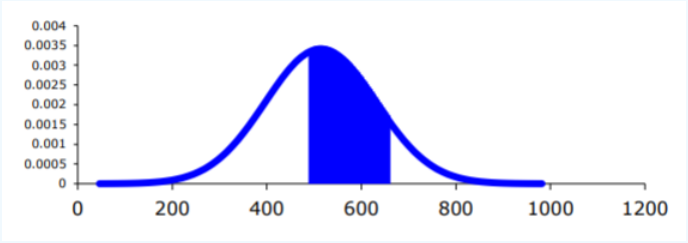


Figure for Example 6.3.2d

On TI-83/84: $P(500 < x < 650) =$ normalcdf $(500, 650, 514, 117) \approx 0.425$

On R: $P(500 < x < 650) =$ pnorm $(650, 514, 117) -$ pnorm $(500, 514, 117) \approx 0.425$

So, there is a 42.5% chance that a person has a mathematical SAT score between 500 and 650.

e. This problem is asking 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: $\text{invNorm}(.99, 514, 117) \approx 786$

On R: $\text{qnorm}(.99, 514, 117) \approx 786$

So, 1% of all people who took the SAT scored over about 786 points on the mathematics SAT.

Homework

Exercise 6.3.1

1. Find each of the probabilities, where z is a z-score from the standard normal distribution with mean of $\mu = 0$ and standard deviation $\sigma = 1$. Make sure you draw a picture for each problem.
   a. *P* (z<2.36)
   b. *P* (z>0.67)
   c. *P* (0<z<2.11)
   d. *P* (-2.78<z<1.97)
2. Find the z-score corresponding to the given area. Remember, z is distributed as the standard normal distribution with mean of $\mu = 0$ and standard deviation $\sigma = 1$.
   a. The area to the left of z is 15%.

b. The area to the right of z is 65%.
c. The area to the left of z is 10%.
d. The area to the right of z is 5%.
e. The area between -z and z is 95%. (Hint draw a picture and figure out the area to the left of the -z.)
f. The area between -z and z is 99%.

3. If a random variable that is normally distributed has a mean of 25 and a standard deviation of 3, convert the given value to a z-score.
   a. x = 23
   b. x = 33
   c. x = 19
   d. x = 45
4. According to the WHO MONICA Project the mean blood pressure for people in China is 128 mmHg with a standard deviation of 23 mmHg (Kuulasmaa, Hense & Tolonen, 1998). Assume that blood pressure is normally distributed.
   a. State the random variable.
   b. Find the probability that a person in China has blood pressure of 135 mmHg or more.
   c. Find the probability that a person in China has blood pressure of 141 mmHg or less.
   d. Find the probability that a person in China has blood pressure between 120 and 125 mmHg.
   e. Is it unusual for a person in China to have a blood pressure of 135 mmHg? Why or why not?
   f. What blood pressure do 90% of all people in China have less than?
5. The size of fish is very important to commercial fishing. A study conducted in 2012 found the length of Atlantic cod caught in nets in Karlskrona to have a mean of 49.9 cm and a standard deviation of 3.74 cm (Ovegard, Berndt & Lunneryd, 2012). Assume the length of fish is normally distributed.
   a. State the random variable.
   b. Find the probability that an Atlantic cod has a length less than 52 cm.
   c. Find the probability that an Atlantic cod has a length of more than 74 cm.
   d. Find the probability that an Atlantic cod has a length between 40.5 and 57.5 cm.
   e. If you found an Atlantic cod to have a length of more than 74 cm, what could you conclude?
   f. What length are 15% of all Atlantic cod longer than?
6. The mean cholesterol levels of women age 45-59 in Ghana, Nigeria, and Seychelles is 5.1 mmol/l and the standard deviation is 1.0 mmol/l (Lawes, Hoorn, Law & Rodgers, 2004). Assume that cholesterol levels are normally distributed.
   a. State the random variable.
   b. Find the probability that a woman age 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level above 6.2 mmol/l (considered a high level).
   c. Find the probability that a woman age 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level below 5.2 mmol/l (considered a normal level).
   d. Find the probability that a woman age 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level between 5.2 and 6.2 mmol/l (considered borderline high).
   e. If you found a woman age 45-59 in Ghana, Nigeria, or Seychelles having a cholesterol level above 6.2 mmol/l, what could you conclude?
   f. What value do 5% of all woman ages 45-59 in Ghana, Nigeria, or Seychelles have a cholesterol level less than?
7. In the United States, males between the ages of 40 and 49 eat on average 103.1 g of fat every day with a standard deviation of 4.32 g ("What we eat," 2012). Assume that the amount of fat a person eats is normally distributed.
   a. State the random variable.
   b. Find the probability that a man age 40-49 in the U.S. eats more than 110 g of fat every day.
   c. Find the probability that a man age 40-49 in the U.S. eats less than 93 g of fat every day.
   d. Find the probability that a man age 40-49 in the U.S. eats less than 65 g of fat every day.
   e. If you found a man age 40-49 in the U.S. who says he eats less than 65 g of fat every day, would you believe him? Why or why not?
   f. What daily fat level do 5% of all men age 40-49 in the U.S. eat more than?

8. A dishwasher has a mean life of 12 years with an estimated standard deviation of 1.25 years ("Appliance life expectancy," 2013). Assume the life of a dishwasher is normally distributed.
    a. State the random variable.
    b. Find the probability that a dishwasher will last more than 15 years.
    c. Find the probability that a dishwasher will last less than 6 years.
    d. Find the probability that a dishwasher will last between 8 and 10 years.
    e. If you found a dishwasher that lasted less than 6 years, would you think that you have a problem with the manufacturing process? Why or why not?
    f. A manufacturer of dishwashers only wants to replace free of charge 5% of all dishwashers. How long should the manufacturer make the warranty period?
9. The mean starting salary for nurses is $67,694 nationally ("Staff nurse -," 2013). The standard deviation is approximately $10,333. Assume that the starting salary is normally distributed.
    a. State the random variable.
    b. Find the probability that a starting nurse will make more than $80,000.
    c. Find the probability that a starting nurse will make less than $60,000.
    d. Find the probability that a starting nurse will make between $55,000 and $72,000.
    e. If a nurse made less than $50,000, would you think the nurse was under paid? Why or why not?
    f. What salary do 30% of all nurses make more than?
10. The mean yearly rainfall in Sydney, Australia, is about 137 mm and the standard deviation is about 69 mm ("Annual maximums of," 2013). Assume rainfall is normally distributed.
    a. State the random variable.
    b. Find the probability that the yearly rainfall is less than 100 mm.
    c. Find the probability that the yearly rainfall is more than 240 mm.
    d. Find the probability that the yearly rainfall is between 140 and 250 mm.
    e. If a year has a rainfall less than 100mm, does that mean it is an unusually dry year? Why or why not?
    f. What rainfall amount are 90% of all yearly rainfalls more than?

Answer

1. a. $P(z < 2.36) = 0.9909$, b. $P(z > 0.67) = 0.2514$, c. $P(0 < z < 2.11) = 0.4826$, d. $P(-2.78 < z < 1.97) = 0.9729$

3. a. -0.6667, b. -2.6667, c. -2, d. 6.6667

5. a. See solutions, b. $P(x < 52\text{cm}) = 0.7128$, c. $P(x > 74\text{cm}) = 5.852 \times 10^{-11}$, d. $P(40.5\text{cm} < x < 57.5\text{cm}) = 0.9729$, e. See solutions, f. 53.8 cm

7. a. See solutions, b. $P(x > 110\text{g}) = 0.0551$ c. $P(x < 93\text{g}) = 0.0097$, d. $P(x < 65\text{g}) \approx 0$ or 5.57×10^{-19}, e. See solutions, f. 110.2 g

9. a. See solutions, b. $P(x > \$80,000) = 0.1168$, c. $P(x > \$80,000) = 0.2283$, d. $P(\$55,000 < x < \$72,000) = 0.5519$, e. See solutions, f. $73,112

This page titled 6.3: Finding Probabilities for the Normal Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

6.4: Assessing Normality

The distributions you have seen up to this point have been assumed to be normally distributed, but how do you determine if it is normally distributed. One way is to take a sample and look at the sample to determine if it appears normal. If the sample looks normal, then most likely the population is also. Here are some guidelines that are use to help make that determination.

1. **Histogram**: Make a histogram. For a normal distribution, the histogram should be roughly bell-shaped. For small samples, this is not very accurate, and another method is needed. A distribution may not look normally distributed from the histogram, but it still may be normally distributed.
2. **Outliers**: For a normal distribution, there should not be more than one outlier. One way to check for outliers is to use a modified box plot. Outliers are values that are shown as dots outside of the rest of the values. If you don't have a modified box plot, outliers are those data values that are:
   Above Q3, the third quartile, by an amount greater than 1.5 times the interquartile range (IQR)
   Below Q1, the first quartile, by an amount greater than 1.5 times the interquartile range (IQR)

   Note

   If there is one outlier, that outlier could have a dramatic effect on the results especially if it is an extreme outlier. However, there are times where a distribution has more than one outlier, but it is still normally distributed. The guideline of only one outlier is just a guideline.

3. **Normal quantile plot (or normal probability plot)**: This plot is provided through statistical software on a computer or graphing calculator. If the points lie close to a line, the data comes from a distribution that is approximately normal. If the points do not lie close to a line or they show a pattern that is not a line, the data are likely to come from a distribution that is not normally distributed.

To create a histogram on the TI-83/84:

1. Go into the STAT menu, and then Chose 1: Edit



Figure 6.4.1*: STAT Menu on TI-83/84*

2. Type your data values into L1.
3. Now click STAT PLOT (2nd $Y=$).

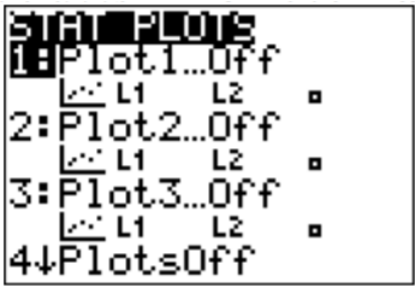


Figure 6.4.2*: STAT PLOT Menu on TI-83/84*

4. Use 1:Plot1. Press ENTER.

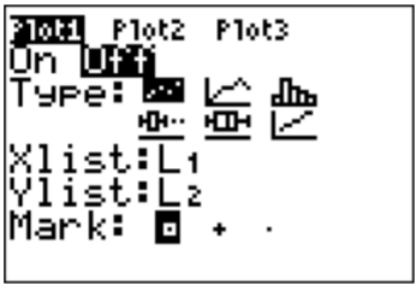


Figure 6.4.3*: Plot1 Menu on TI-83/84*

5. You will see a new window. The first thing you want to do is turn the plot on. At this point you should be on On, just press ENTER. It will make On dark.
6. Now arrow down to Type: and arrow right to the graph that looks like a histogram (3rd one from the left in the top row).
7. Now arrow down to Xlist. Make sure this says L1. If it doesn't, then put L1 there (2nd number 1). Freq: should be a 1.

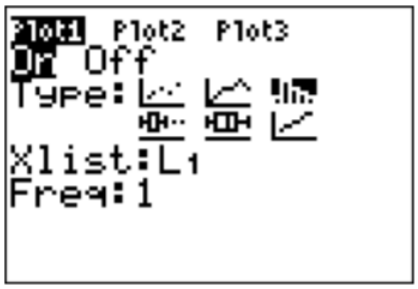


Figure 6.4.4*: Plot1 Menu on TI-83/84 Setup for Histogram*

8. Now you need to set up the correct window to graph on. Click on WINDOW. You need to set up the settings for the x variable. Xmin should be your smallest data value. Xmax should just be a value sufficiently above your highest data value, but not too high. Xscl is your class width that you calculated. Ymin should be 0 and Ymax should be above what you think the highest frequency is going to be. You can always change this if you need to. Yscl is just how often you would like to see a tick mark on the y-axis.
9. Now press GRAPH. You will see a histogram.

To find the IQR and create a box plot on the TI-83/84:

1. Go into the STAT menu, and then Choose 1:Edit



Figure 6.4.5*: STAT Menu on TI-83/84*

2. Type your data values into L1. If L1 has data in it, arrow up to the name L1, click CLEAR and then press ENTER. The column will now be cleared and you can type the data in.
3. Go into the STAT menu, move over to CALC and choose 1-Var Stats. Press ENTER, then type L1 (2nd 1) and then ENTER. This will give you the summary statistics. If you press the down arrow, you will see the five-number summary.

4. To draw the box plot press 2nd STAT PLOT.

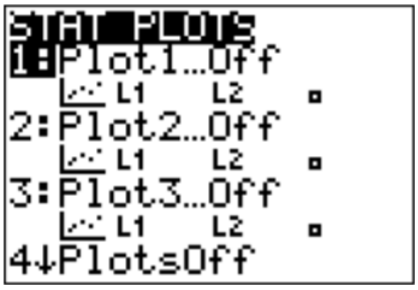


Figure 6.4.6*: STAT PLOT Menu on TI-83/84*

5. Use Plot1. Press ENTER

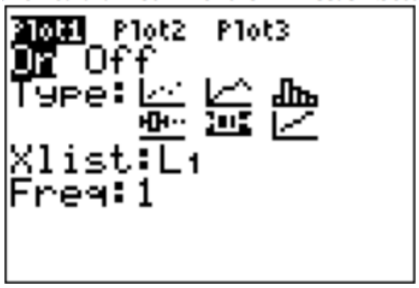


Figure 6.4.7*: Plot1 Menu on TI-83/84 Setup for Box Plot*

6. Put the cursor on On and press Enter to turn the plot on. Use the down arrow and the right arrow to highlight the boxplot in the middle of the second row of types then press ENTER. Set Data List to L1 (it might already say that) and leave Freq as 1.
7. Now tell the calculator the set up for the units on the x-axis so you can see the whole plot. The calculator will do it automatically if you press ZOOM, which is in the middle of the top row.



Figure 6.4.8*: ZOOM Menu on TI-83/84*

Then use the down arrow to get to 9:ZoomStat and press ENTER. The box plot will be drawn.

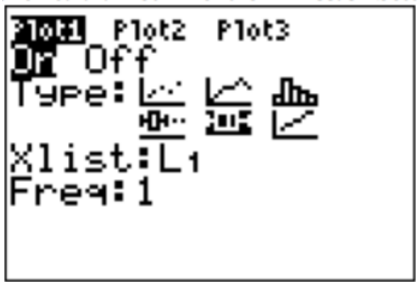


Figure 6.4.9*: ZOOM Menu on TI-83/84 with ZoomStat*

To create a normal quantile plot on the TI-83/84

1. Go into the STAT menu, and then Chose 1:Edit



Figure 6.4.10*: STAT Menu on TI-83/84*

2. Type your data values into L1. If L1 has data in it, arrow up to the name L1, click CLEAR and then press ENTER. The column will now be cleared and you can type the data in.
3. Now click STAT PLOT (2nd $Y=$). You have three stat plots to choose from.

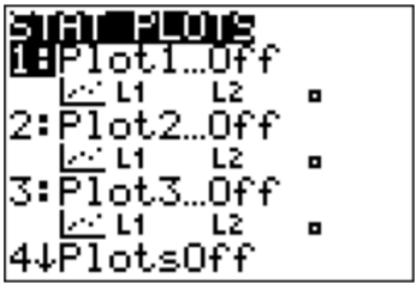


Figure 6.4.11*: STAT PLOT Menu on TI-83/84*

4. Use 1:Plot1. Press ENTER.
5. Put the cursor on the word On and press ENTER. This turns on the plot. Arrow down to Type: and use the right arrow to move over to the last graph (it looks like an increasing linear graph). Set Data List to L1 (it might already say that) and set Data Axis to Y. The Mark is up to you.



Figure 6.4.12*: Plot1 Menu on TI-83/84 Setup for Normal Quantile Plot*

6. Now you need to set up the correct window on which to graph. Click on WINDOW. You need to set up the settings for the *x* variable. Xmin should be -4. Xmax should be 4. Xscl should be 1. Ymin and Ymax are based on your data, the Ymin should be below your lowest data value and Ymax should be above your highest data value. Yscl is just how often you would like to see a tick mark on the *y*-axis.
7. Now press GRAPH. You will see the normal quantile plot.

To create a histogram on R:

Put the variable in using variable<-c(type in the data with commas between values) using a name for the variable that makes sense for the problem. The command for histogram is hist(variable). You can then copy the histogram into a word processing program. There are options that you can put in for title, and axis labels. See section 2.2 for the commands for those.

To create a modified boxplot on R:

Put the variable in using variable<-c(type in the data with commas between values) using a name for the variable that makes sense for the problem. The command for box plot is boxplot(variable). You can then copy the box plot into a word processing program. There are options that you can put in for title, horizontal orientation, and axis labels. See section 3.3 for the commands for those.

To create a normal quantile plot on R:

Put the variable in using variable<-c(type in the data with commas between values) using a name for the variable that makes sense for the problem. The command for normal quantile plot is qqnorm(variable). You can then copy the normal quantile plot into a word processing program.

Realize that your random variable may be normally distributed, even if the sample fails the three tests. However, if the histogram definitely doesn't look symmetric and bell shaped, there are outliers that are very extreme, and the normal probability plot doesn't look linear, then you can be fairly confident that the data set does not come from a population that is normally distributed.

Example 6.4.1 is it normal?

In Kiama, NSW, Australia, there is a blowhole. The data in table #6.4.1 are times in seconds between eruptions ("Kiama blowhole eruptions," 2013). Do the data come from a population that is normally distributed?

Table 6.4.1: Time (in Seconds) Between Kiama Blowhole Eruptions

83	51	87	60	28	95	8	27
15	10	18	16	29	54	91	8
17	55	10	35	47	77	36	17
21	36	18	40	10	7	34	27
28	56	8	25	68	146	89	18
73	69	9	37	10	82	29	8
60	61	61	18	169	25	8	26
11	83	11	42	17	14	9	12

a. State the random variable
b. Draw a histogram.
c. Find the number of outliers.
d. Draw the normal quantile plot.
e. Do the data come from a population that is normally distributed?

Solution

a. x = time in seconds between eruptions of Kiama Blowhole

b. The histogram produced is in *Figure* 6.4.13.

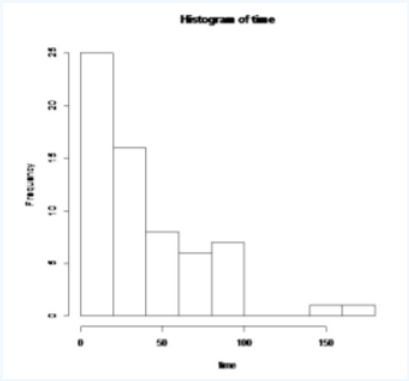


Figure 6.4.13: Histogram for Kiama Blowhole

This looks skewed right and not symmetric.

c. The box plot is in *Figure* 6.4.14.

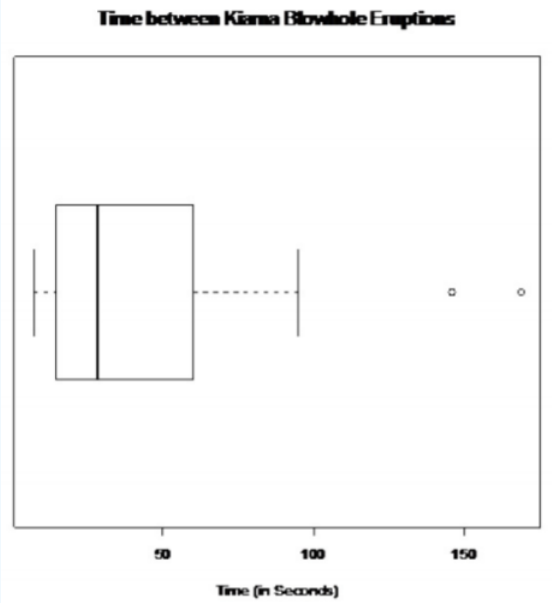


Figure 6.4.14: Modified Box Plot from TI-83/83 for Kiama Blowhole

There are two outliers. Instead using:

$IQR = Q3 - Q1 = 60 - 14.5 = 45.5$ seconds

$1.5 * IQR = 1.5 * 45.5 = 68.25$ seconds

$Q1 - 1.5^* IQR = 14.5 - 68.25 = -53.75$ seconds

$Q3 + 1.5 * IQR = 60 + 68.25 = 128.25$ seconds

Outliers are any numbers greater than 128.25 seconds and less than -53.75 seconds. Since all the numbers are measurements of time, then no data values are less than 0 or seconds for that matter. There are two numbers that are larger than 128.25 seconds, so there are two outliers. Two outliers are not real indications that the sample does not come from a normal distribution, but the fact that both are well above 128.25 seconds is an indication of an issue.

d. The normal quantile plot is in *Figure* 6.4.15.

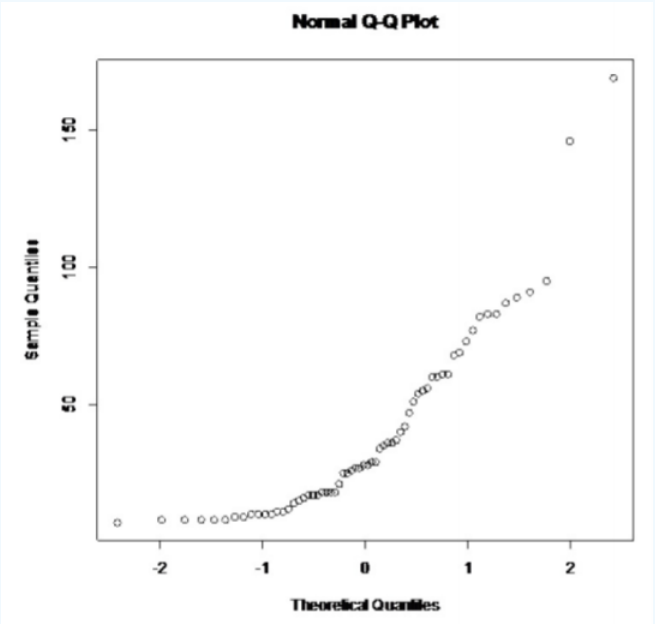


Figure 6.4.15: Normal Probability Plot

This graph looks more like an exponential growth than linear.

e. Considering the histogram is skewed right, there are two extreme outliers, and the normal probability plot does not look linear, then the conclusion is that this sample is not from a population that is normally distributed.

Example 6.4.2 is it normal?

One way to measure intelligence is with an IQ score. Example 6.4.2 contains 50 IQ scores. Determine if the sample comes from a population that is normally distributed.

Table 6.4.2: IQ Scores

78	92	96	100	67	105	109	75	127	111
93	114	82	100	125	67	94	74	81	98
102	108	81	96	103	91	90	96	86	92
84	92	90	103	115	93	85	116	87	106
85	88	106	104	102	98	116	107	102	89

a. State the random variable.
b. Draw a histogram.
c. Find the number of outliers.
d. Draw the normal quantile plot.
e. Do the data come from a population that is normally distributed?

Solution

a. x = IQ score

b. The histogram is in *Figure* 6.4.16.

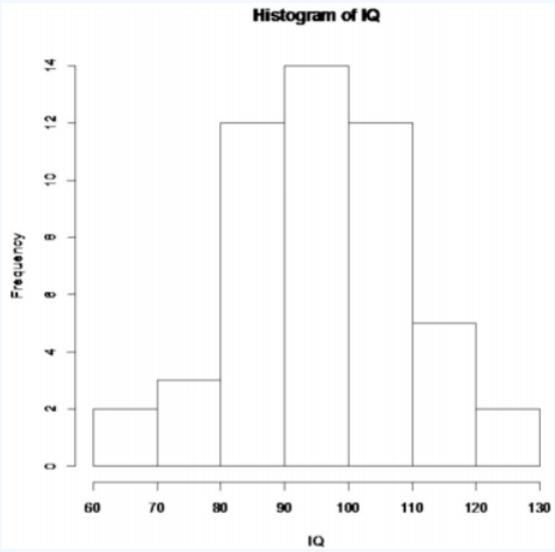


Figure 6.4.16: Histogram for IQ Score

This looks somewhat symmetric, though it could be thought of as slightly skewed right.

c. The modified box plot is in *Figure* 6.4.17.

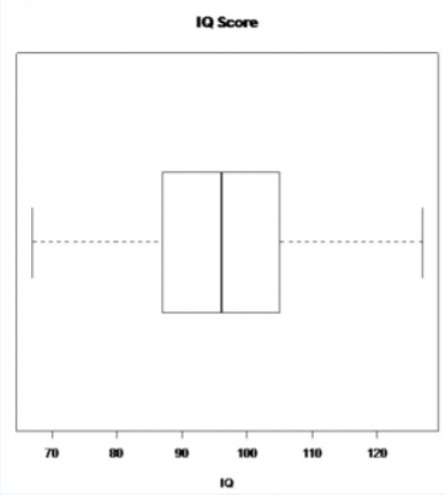


Figure 6.4.17: Output from TI-83/84 for IQ Score

There are no outliers.

Or using Outliers

$IQR = Q3 - Q1 = 105 - 87 = 18$

$1.5^{*}IQR = 1.5^{*}18 = 27$

$Q1 - 1.5IQR = 87 - 27 = 60$

$Q3 + 1.5IQR = 105 + 27 = 132$

are any numbers greater than 132 and less than 60. Since the maximum number is 127 and the minimum is 67, there are no outliers.

d. The normal quantile plot is in *Figure* 6.4.18.

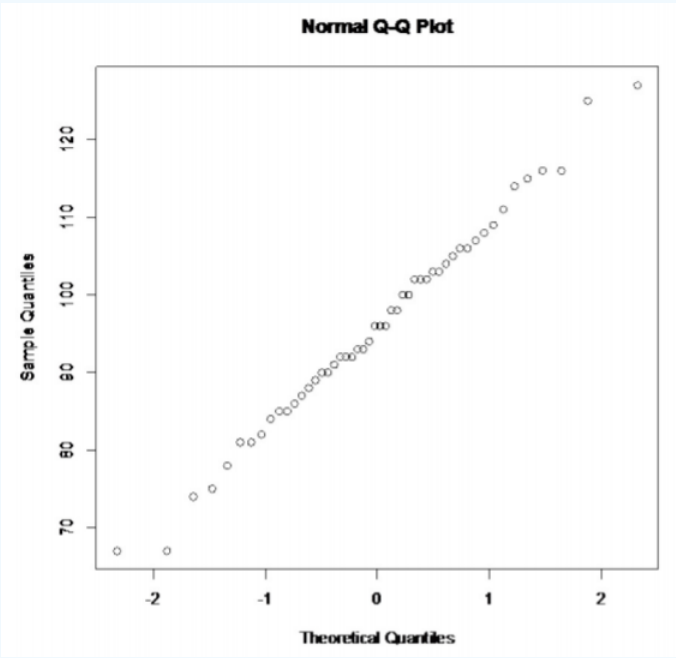


Figure 6.4.18: Normal Quantile Plot

This graph looks fairly linear.

e. Considering the histogram is somewhat symmetric, there are no outliers, and the normal probability plot looks linear, then the conclusion is that this sample is from a population that is normally distributed.

Homework

Exercise 6.4.1

1. Cholesterol data was collected on patients four days after having a heart attack. The data is in Example 6.4.3. Determine if the data is from a population that is normally distributed.

Table 6.4.3: Cholesterol Data Collected Four Days After a Heart Attack

218	234	214	116	200	276	146
182	238	288	190	236	244	258
240	294	220	200	220	186	352
202	218	248	278	248	270	242

2. The size of fish is very important to commercial fishing. A study conducted in 2012 collected the lengths of Atlantic cod caught in nets in Karlskrona (Ovegard, Berndt & Lunneryd, 2012). Data based on information from the study is in Example 6.4.4. Determine if the data is from a population that is normally distributed.

Table 6.4.4: Atlantic Cod Lengths

48	50	50	55	53	50	49	52
61	48	45	47	53	46	50	48
42	44	50	60	54	48	50	49
53	48	52	56	46	46	47	48
48	49	52	47	51	48	45	47

3. The WHO MONICA Project collected blood pressure data for people in China (Kuulasmaa, Hense & Tolonen, 1998). Data based on information from the study is in Example 6.4.5. Determine if the data is from a population that is normally distributed.

Table 6.4.5: Blood Pressure Values for People in China

114	141	154	137	131	132	133	156	119
138	86	122	112	114	177	128	137	140
171	129	127	104	97	135	107	136	118
92	182	150	142	97	140	106	76	115
119	125	162	80	138	124	132	143	119

4. Annual rainfalls for Sydney, Australia are given in Example 6.4.6. ("Annual maximums of," 2013). Can you assume rainfall is normally distributed?

Table 6.4.6: Annual Rainfall in Sydney, Australia

146.8	383	90.9	178.1	267.5	95.5	156.5	180
90.9	139.7	200.2	171.7	187.2	184.9	70.1	58
84.1	55.6	133.1	271.8	135.9	71.9	99.4	110.6
47.5	97.8	122.7	58.4	154.4	173.7	118.8	88
84.6	171.5	254.3	185.9	137.2	138.9	96.2	85
45.2	74.7	264.9	113.8	133.4	68.1	156.4	

Answer

1. Normally distributed
3. Normally distributed

This page titled 6.4: Assessing Normality is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

6.5: Sampling Distribution and the Central Limit Theorem

You now have most of the skills to start statistical inference, but you need one more concept.

First, it would be helpful to state what statistical inference is in more accurate terms.

Definition 6.5.1: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 make a decision about.

How is the statistic distributed? This is answered with a sampling distribution.

Definition 6.5.2: Sampling Distribution

Sampling Distribution: how a sample statistic is distributed when repeated trials of size *n* are taken.

Example 6.5.1 sampling distribution

Suppose you throw a penny and count how often a head comes up. The random variable is x = number of heads. The probability distribution (pdf) of this random variable is presented in *Figure* 6.5.1.

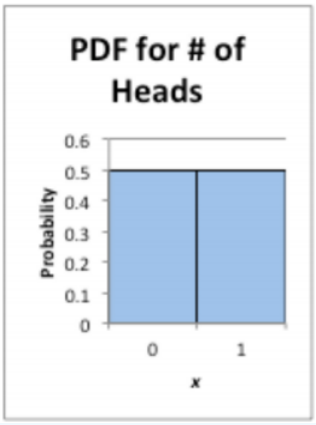


Figure 6.5.1: Distribution of Random Variable

Solution

Repeat this experiment 10 times, which means $n = 10$. Here is the data set:

{1, 1, 1, 1, 0, 0, 0, 0, 0, 0}. The mean of this sample is 0.4. Now take another sample. Here is that data set:

{1, 1, 1, 0, 1, 0, 1, 1, 0, 0}. The mean of this sample is 0.6. Another sample looks like:

{0, 1, 0, 1, 1, 1, 1, 1, 0, 1}. The mean of this sample is 0.7. Repeat this 40 times. You could get these means:

Table 6.5.1: Sample Means When n=10

0.4	0.6	0.7	0.3	0.3	0.2	0.5	0.5	0.5	0.5
0.4	0.4	0.5	0.7	0.7	0.6	0.4	0.4	0.4	0.6
0.7	0.7	0.3	0.5	0.6	0.3	0.3	0.8	0.3	0.6
0.4	0.3	0.5	0.6	0.5	0.6	0.3	0.5	0.6	0.2

Example 6.5.2 contains the distribution of these sample means (just count how many of each number there are and then divide by 40 to obtain the relative frequency).

Table 6.5.2: Distribution of Sample Means When n=10

Sample Mean	Probability
0.1	0
0.2	0.05
0.3	0.2
0.4	0.175
0.5	0.225
0.6	0.2
0.7	0.125
0.8	0.025
0.9	0

Figure 6.5.2 contains the histogram of these sample means.

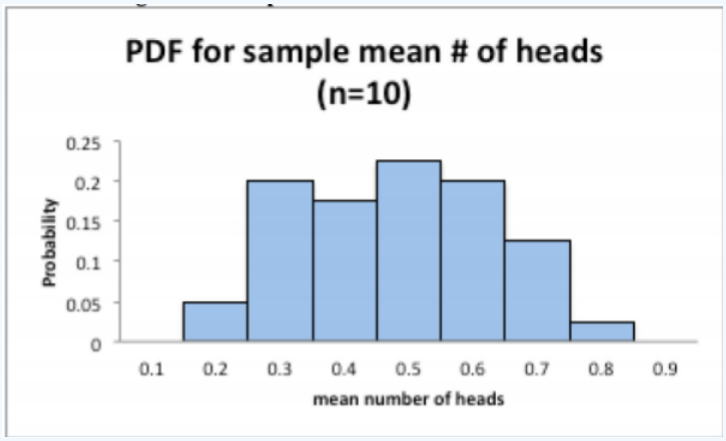


Figure 6.5.2: Histogram of Sample Means When n=10

This distribution (represented graphically by the histogram) is a sampling distribution. That is all a sampling distribution is. It is a distribution created from statistics.

Notice the histogram does not look anything like the histogram of the original random variable. It also doesn't look anything like a normal distribution, which is the only one you really know how to find probabilities. Granted you have the binomial, but the normal is better.

What does this distribution look like if instead of repeating the experiment 10 times you repeat it 20 times instead?

Example 6.5.3 contains 40 means when the experiment of flipping the coin is repeated 20 times.

Table 6.5.3: Sample Means When n=20

0.5	0.45	0.7	0.55	0.65	0.6	0.4	0.35	0.45	0.6
0.5	0.5	0.65	0.5	0.5	0.35	0.55	0.4	0.65	0.3
0.4	0.5	0.45	0.45	0.65	0.7	0.6	0.5	0.7	0.7
0.7	0.45	0.35	0.6	0.65	0.55	0.35	0.4	0.55	0.6

Example 6.5.3 contains the sampling distribution of the sample means.

Table 6.5.3: Distribution of Sample Means When n=20

Mean	Probability

Mean	Probability
0.1	0
0.2	0
0.3	0.125
0.4	0.2
0.5	0.3
0.6	0.25
0.7	0.125
0.8	0
0.9	0

This histogram of the sampling distribution is displayed in *Figure* 6.5.3.

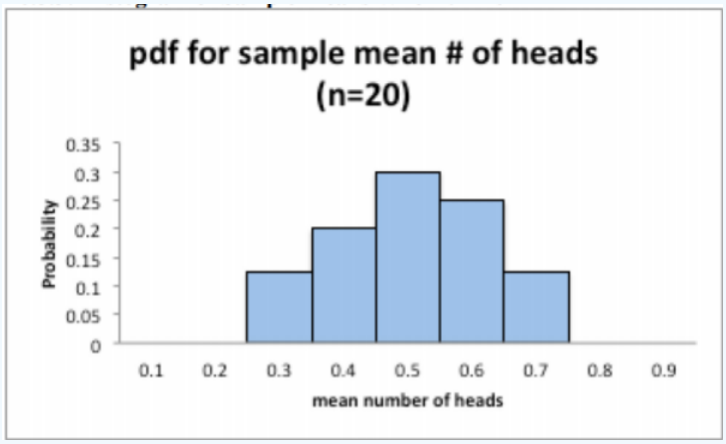


Figure 6.5.3: Histogram of Sample Means When n=20

Notice this histogram of the sample mean looks approximately symmetrical and could almost be called normal. What if you keep increasing *n*? What will the sampling distribution of the sample mean look like? In other words, what does the sampling distribution of $\overline{x}$ look like as *n* gets even larger?

This depends on how the original distribution is distributed. In Example 6.5.1, the random variable was uniform looking. But as *n* increased to 20, the distribution of the mean looked approximately normal. What if the original distribution was normal? How big would *n* have to be? Before that question is answered, another concept is needed.

Note

Suppose you have a random variable that has a population mean, μ, and a population standard deviation, σ. If a sample of size *n* is taken, then the sample mean, $\overline{x}$ has a mean $\mu_{\overline{x}} = \mu$ and standard deviation of $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation of $\overline{x}$ is lower because by taking the mean you are averaging out the extreme values, which makes the distribution of the original random variable spread out.

You now know the center and the variability of $\overline{x}$. You also want to know the shape of the distribution of $\overline{x}$. You hope it is normal, since you know how to find probabilities using the normal curve. The following theorem tells you the requirement to have $\overline{x}$ normally distributed.

Theorem 6.5.1 central limit theorem

Suppose a random variable is from any distribution. If a sample of size *n* is taken, then the sample mean, $\overline{x}$, becomes normally distributed as *n* increases.

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 really skewed one way or another.

Now that you know when the sample mean will look like a normal distribution, then 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($\overline{x}, \mu, \sigma/\text{sqrt}(n)$) to find the area to the left of $\overline{x}$. Remember to subtract pnorm($\overline{x}, \mu, \sigma/\text{sqrt}(n)$)) from 1 if you want the area to the right of $\overline{x}$.

Example 6.5.2 Finding probabilities for sample means

The birth weight of boy babies of European descent who were delivered at 40 weeks is normally distributed with a mean of 3687.6 g with a standard deviation of 410.5 g (Janssen, Thiessen, Klein, Whitfield, MacNab & Cullis-Kuhl, 2007). Suppose there were nine European descent boy babies born on a given day and the mean birth weight is calculated.

a. State the random variable.
b. What is the mean of the sample mean?
c. What is the standard deviation of the sample mean?
d. What distribution is the sample mean distributed as?
e. Find the probability that the mean weight of the nine boy babies born was less than 3500.4 g.
f. Find the probability that the mean weight of the nine babies born was less than 3452.5 g.

Solution

a. *x* = birth weight of boy babies (Note: the random variable is something you measure, and it is not the mean birth weight. Mean birth weight is calculated.)

b. $\mu_{\overline{x}} = \mu = 3687.6\text{g}$

c. $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{410.5}{\sqrt{9}} = \frac{410.5}{3} \approx 136.8\text{g}$

d. Since the original random variable is distributed normally, then the sample mean is distributed normally.

e. You are looking for the $P(\overline{x} < 3500.4)$. You use the normalcdf command on the calculator. Remember to use the standard deviation you found in part c. However to reduce rounding error, type the division into the command. On the TI-83/84 you would have

$$P(\overline{x} < 3500.4) = \text{ normalcdf } (-1E99, 3500.4, 3687.6, 410.5 \div \sqrt{9}) \approx 0.086$$

On R you would have

$$P(\overline{x} < 3500.4) = \text{ pnorm } (3500.4, 3687.6, 410.5/sqr(9)) \approx 0.086$$

There is an 8.6% chance that the mean birth weight of the nine boy babies born would be less than 3500.4 g. Since this is more than 5%, this is not unusual.

f. You are looking for the $P(\overline{x} < 3452.5)$.

On TI-83/84:

$$P(\overline{x} < 3452.5) = \text{ normalcdf } (-1E99, 3452.5, 3687.6, 410.5 \div \sqrt{9}) \approx 0.043$$

On R:

$$P(\overline{x} < 3452.5) = \text{ pnorm } (3452.5, 3687.6, 410.5 \div \sqrt{9}) \approx 0.043$$

There is a 4.3% chance that the mean birth weight of the nine boy babies born would be less than 3452.5 g. Since this is less than 5%, this would be an unusual event. If it actually happened, then you may think there is something unusual about this sample. Maybe some of the nine babies were born as multiples, which brings the mean weight down, or some or all of the babies were not of European descent (in fact the mean weight of South Asian boy babies is 3452.5 g), or some were born before 40 weeks, or the babies were born at high altitudes.

Example 6.5.3 finding probabilities for sample means

The age that American females first have intercourse is on average 17.4 years, with a standard deviation of approximately 2 years ("The Kinsey institute," 2013). This random variable is not normally distributed, though it is somewhat mound shaped.

a. State the random variable.
b. Suppose a sample of 35 American females is taken. Find the probability that the mean age that these 35 females first had intercourse is more than 21 years.

Solution

a. x = age that American females first have intercourse.

b. Even though the original random variable is not normally distributed, the sample size is over 30, by the central limit theorem the sample mean will be normally distributed. The mean of the sample mean is $\mu_{\overline{x}} = \mu = 17.4$ years. The standard deviation of the sample mean is $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2}{\sqrt{35}} \approx 0.33806$. You have all the information you need to use the normal command on your technology. Without the central limit theorem, you couldn't use the normal command, and you would not be able to answer this question.

On the TI-83/84:

$$P(\overline{x} > 21) = \text{ normalcdf } (21, 1E99, 17.4, 2 \div \sqrt{35}) \approx 9.0 \times 10^{-27}$$

On R:

$$P(\overline{x} > 21) = 1 - \text{ pnorm } (21, 17.4, 2/\text{ sqrt}(35)) \approx 9.0 \times 10^{-27}$$

The probability of a sample mean of 35 women being more than 21 years when they had their first intercourse is very small. This is extremely unlikely to happen. If it does, it may make you wonder about the sample. Could the population mean have increased from the 17.4 years that was stated in the article? Could the sample not have been random, and instead have been a group of women who had similar beliefs about intercourse? These questions, and more, are ones that you would want to ask as a researcher.

Homework

Exercise 6.5.1

1. A random variable is not normally distributed, but it is mound shaped. It has a mean of 14 and a standard deviation of 3.
   a. If you take a sample of size 10, can you say what the shape of the sampling distribution for the sample mean is? Why?
   b. For a sample of size 10, state the mean of the sample mean and the standard deviation of the sample mean.
   c. If you take a sample of size 35, can you say what the shape of the distribution of the sample mean is? Why?
   d. For a sample of size 35, state the mean of the sample mean and the standard deviation of the sample mean.
2. A random variable is normally distributed. It has a mean of 245 and a standard deviation of 21.
   a. If you take a sample of size 10, can you say what the shape of the distribution for the sample mean is? Why?
   b. For a sample of size 10, state the mean of the sample mean and the standard deviation of the sample mean.
   c. For a sample of size 10, find the probability that the sample mean is more than 241.
   d. If you take a sample of size 35, can you say what the shape of the distribution of the sample mean is? Why?
   e. For a sample of size 35, state the mean of the sample mean and the standard deviation of the sample mean.
   f. For a sample of size 35, find the probability that the sample mean is more than 241.

g. Compare your answers in part d and f. Why is one smaller than the other?

3. The mean starting salary for nurses is $67,694 nationally ("Staff nurse -," 2013). The standard deviation is approximately $10,333. The starting salary is not normally distributed but it is mound shaped. A sample of 42 starting salaries for nurses is taken.
   a. State the random variable.
   b. What is the mean of the sample mean?
   c. What is the standard deviation of the sample mean?
   d. What is the shape of the sampling distribution of the sample mean? Why?
   e. Find the probability that the sample mean is more than $75,000.
   f. Find the probability that the sample mean is less than $60,000.
   g. If you did find a sample mean of more than $75,000 would you find that unusual? What could you conclude?
4. According to the WHO MONICA Project the mean blood pressure for people in China is 128 mmHg with a standard deviation of 23 mmHg (Kuulasmaa, Hense & Tolonen, 1998). Blood pressure is normally distributed.
   a. State the random variable.
   b. Suppose a sample of size 15 is taken. State the shape of the distribution of the sample mean.
   c. Suppose a sample of size 15 is taken. State the mean of the sample mean.
   d. Suppose a sample of size 15 is taken. State the standard deviation of the sample mean.
   e. Suppose a sample of size 15 is taken. Find the probability that the sample mean blood pressure is more than 135 mmHg.
   f. Would it be unusual to find a sample mean of 15 people in China of more than 135 mmHg? Why or why not?
   g. If you did find a sample mean for 15 people in China to be more than 135 mmHg, what might you conclude?
5. The size of fish is very important to commercial fishing. A study conducted in 2012 found the length of Atlantic cod caught in nets in Karlskrona to have a mean of 49.9 cm and a standard deviation of 3.74 cm (Ovegard, Berndt & Lunneryd, 2012). The length of fish is normally distributed. A sample of 15 fish is taken.
   a. State the random variable.
   b. Find the mean of the sample mean.
   c. Find the standard deviation of the sample mean
   d. What is the shape of the distribution of the sample mean? Why?
   e. Find the probability that the sample mean length of the Atlantic cod is less than 52 cm.
   f. Find the probability that the sample mean length of the Atlantic cod is more than 74 cm.
   g. If you found sample mean length for Atlantic cod to be more than 74 cm, what could you conclude?
6. The mean cholesterol levels of women age 45-59 in Ghana, Nigeria, and Seychelles is 5.1 mmol/l and the standard deviation is 1.0 mmol/l (Lawes, Hoorn, Law & Rodgers, 2004). Assume that cholesterol levels are normally distributed.
   a. State the random variable.
   b. Find the probability that a woman age 45-59 in Ghana has a cholesterol level above 6.2 mmol/l (considered a high level).
   c. Suppose doctors decide to test the woman's cholesterol level again and average the two values. Find the probability that this woman's mean cholesterol level for the two tests is above 6.2 mmol/l.
   d. Suppose doctors being very conservative decide to test the woman's cholesterol level a third time and average the three values. Find the probability that this woman's mean cholesterol level for the three tests is above 6.2 mmol/l.
   e. If the sample mean cholesterol level for this woman after three tests is above 6.2 mmol/l, what could you conclude?
7. In the United States, males between the ages of 40 and 49 eat on average 103.1 g of fat every day with a standard deviation of 4.32 g ("What we eat," 2012). The amount of fat a person eats is not normally distributed but it is relatively mound shaped.
   a. State the random variable.
   b. Find the probability that a sample mean amount of daily fat intake for 35 men age 40-59 in the U.S. is more than 100 g.
   c. Find the probability that a sample mean amount of daily fat intake for 35 men age 40-59 in the U.S. is less than 93 g.
   d. If you found a sample mean amount of daily fat intake for 35 men age 40-59 in the U.S. less than 93 g, what would you conclude?

8. A dishwasher has a mean life of 12 years with an estimated standard deviation of 1.25 years ("Appliance life expectancy," 2013). The life of a dishwasher is normally distributed. Suppose you are a manufacturer and you take a sample of 10 dishwashers that you made.
   a. State the random variable.
   b. Find the mean of the sample mean.
   c. Find the standard deviation of the sample mean.
   d. What is the shape of the sampling distribution of the sample mean? Why?
   e. Find the probability that the sample mean of the dishwashers is less than 6 years.
   f. If you found the sample mean life of the 10 dishwashers to be less than 6 years, would you think that you have a problem with the manufacturing process? Why or why not?

Answer

1. a. See solutions, b. $\mu_{\bar{x}} = 14$, $\sigma_{\bar{x}} = 0.9487$, c. See solutions, d. $\mu_{\bar{x}} = 14$, $\sigma_{\bar{x}} = 0.5071$

3. a. See solutions, b. $\mu_{\bar{x}} = \$67,694$, c. $\sigma_{\bar{x}} = \$1594.42$, d. See solutions, e. $P(\bar{x} > \$75,000) = 2.302 \times 10^{-6}$, f. $P(\bar{x} < \$60,000) = 6.989 \times 10^{-7}$, g. See solutions

5. a. See solutions, b. $\mu_{\bar{x}} = 49.9\text{cm}$, c. $\sigma_{\bar{x}} = 0.9657\text{cm}$, d. See solutions, e. $P(\bar{x} < 52\text{cm}) = 0.9852$ f. $P(\bar{x} > 74\text{cm}) \approx 0$, g. See solutions

7. a. See solutions, b. $P(\bar{x} > 100\text{g}) = 0.99999$, c. $P(\bar{x} < 93\text{g}) \approx 0$ or 8.22×10^{-44}, d. See solutions

Data Sources:

Annual maximums of daily rainfall in Sydney. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/sydrain.html

Appliance life expectancy. (2013, November 8). Retrieved from http://www.mrappliance.com/expert/life-guide/

Bhat, R., & Kushtagi, P. (2006). A re-look at the duration of human pregnancy. *Singapore Med J*., 47(12), 1044-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17139400

College Board, SAT. (2012). *Total group profile report.* Retrieved from website: media.collegeboard.com/digita...lGroup2012.pdf

Greater Cleveland Regional Transit Authority, (2012). 2012 *annual report*. Retrieved from website: http://www.riderta.com/annual/2012

Janssen, P. A., Thiessen, P., Klein, M. C., Whitfield, M. F., MacNab, Y. C., & CullisKuhl, S. C. (2007). Standards for the measurement of birth weight, length and head circumference at term in neonates of european, chinese and south asian ancestry. *Open Medicine*, 1(2), e74-e88. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2802014/

Kiama blowhole eruptions. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/kiama.html

Kuulasmaa, K., Hense, H., & Tolonen, H. World Health Organization (WHO), WHO Monica Project. (1998). *Quality assessment of data on blood pressure in the who monica project* (ISSN 2242-1246). Retrieved from WHO MONICA Project e-publications website: http://www.thl.fi/publications/monica/bp/bpqa.htm

Lawes, C., Hoorn, S., Law, M., & Rodgers, A. (2004). High cholesterol. In M. Ezzati, A. Lopez, A. Rodgers & C. Murray (Eds.), *Comparative Quantification of Health Risks* (1 ed., Vol. 1, pp. 391-496). Retrieved from http://www.who.int/publications/cra/.../0391-0496.pdf

Ovegard, M., Berndt, K., & Lunneryd, S. (2012). Condition indices of atlantic cod (gadus morhua) biased by capturing method. *ICES Journal of Marine Science*, doi: 10.1093/icesjms/fss145

Staff nurse - RN salary. (2013, November 08). Retrieved from http://www1.salary.com/Staff-Nurse-RN-salary.html

The Kinsey institute - sexuality information links. (2013, November 08). Retrieved from www.iub.edu/~kinsey/resources/FAQ.html

US Department of Argriculture, Agricultural Research Service. (2012). *What we eat in America*. Retrieved from website: http://www.ars.usda.gov/Services/docs.htm?docid=18349

This page titled 6.5: Sampling Distribution and the Central Limit Theorem is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

7: One-Sample Inference

Now that you have all this information about descriptive statistics and probabilities, it is time to start inferential statistics. There are two branches of inferential statistics: hypothesis testing and confidence intervals.

Definition 7.1

Hypothesis Testing: making a decision about a parameter(s) based on a statistic(s).

Definition 7.2

Confidence Interval: estimating a parameter(s) based on a statistic(s).

7.1: Basics of Hypothesis Testing
7.2: One-Sample Proportion Test
7.3: One-Sample Test for the Mean

This page titled 7: One-Sample Inference is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

7.1: Basics of Hypothesis Testing

To understand the process of a hypothesis tests, you need to first have an understanding of what a hypothesis is, which is an educated guess about a parameter. Once you have the hypothesis, you collect data and use the data to make a determination to see if there is enough evidence to show that the hypothesis is true. However, in hypothesis testing you actually assume something else is true, 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 actually false. If you are able to say this assumption is false, then your hypothesis must be true. This is known as a proof by contradiction. You assume the opposite of your hypothesis is true and show that it can't be true. If this happens, then your hypothesis must be true. All hypothesis tests go through the same process. Once you have the process down, then the concept is much easier. It is easier to see the process by looking at an example. Concepts that are needed will be detailed in this example.

Example 7.1.1 basics of hypothesis testing

Suppose a manufacturer of the XJ35 battery claims the mean life of the battery is 500 days with a standard deviation of 25 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 can't get out of your contract.

What do you do?

Solution

Well first, you should know what you are trying to measure. Define the random variable.

Let x = life of a XJ35 battery

Now you are not just trying to find different x values. You are trying to find what the true mean is. Since you are trying to find it, it must be unknown. You don't think it is 500 days. If you did, you wouldn't be doing any testing. The true mean, μ, is unknown. That means you should define that too.

Let μ= mean life of a XJ35 battery

Now what?

You may want to collect a sample. What kind of sample?

You could ask the manufacturers to give you batteries, but there is a chance that there could be some bias in the batteries they pick. To reduce the chance of bias, it is best to take a random sample.

How big should the sample be?

A sample of size 30 or more means that you can use the central limit theorem. Pick a sample of size 30.

Example 7.1.1 contains the data for the sample you collected:

Table 7.1.1: Data on Battery Life

491	485	503	492	282	490
489	495	497	487	493	480
483	504	501	486	478	492
482	502	485	503	497	500
488	475	478	490	487	486

Now what should you do? Looking at the data set, you see some of the times are above 500 and some are below. But looking at all of the numbers is too difficult. It might be helpful to calculate the mean for this sample.

The sample mean is $\bar{x} = 490$ days. Looking at the sample mean, one might think that you are right. However, the standard deviation and the sample size also plays a role, so maybe you are wrong.

Before going any farther, it is time to formalize a few definitions.

You have a guess that the mean life of a battery is less than 500 days. This is opposed to what the manufacturer claims. There really are two hypotheses, which are just guesses here – the one that the manufacturer claims and the one that you believe. It is helpful to have names for them.

Definition 7.1.1

Null Hypothesis: historical value, claim, or product specification. The symbol used is H_o.

Definition 7.1.2

Alternate Hypothesis: what you want to prove. This is what you want to accept as true when you reject the null hypothesis. There are two symbols that are commonly used for the alternative hypothesis: H_A or H_I. The symbol H_A will be used in this book.

In general, the hypotheses look something like this:

$$H_o : \mu = \mu_o$$

$$H_A : \mu < \mu_o$$

where μ_o just represents the value that the claim says the population mean is actually equal to.

Also, H_A can be less than, greater than, or not equal to.

For this problem:

$H_o : \mu = 500$ days, since the manufacturer says the mean life of a battery is 500 days.

$H_A : \mu < 500$ days, since you believe that the mean life of the battery is less than 500 days.

Now back to the mean. You have a sample mean of 490 days. Is this small enough to believe that you are right and the manufacturer is wrong? How small does it have to be?

If you calculated a sample mean of 235, you would definitely believe the population mean is less than 500. But even if you had a sample mean of 435 you would probably believe that the true mean was less than 500. What about 475? Or 483? There is some point where you would stop being so sure that the population mean is less than 500. That point separates the values of where you are sure or pretty sure that the mean is less than 500 from the area where you are not so sure. How do you find that point?

Well it depends on how much error you want to make. Of course you don't want to make any errors, but unfortunately that is unavoidable in statistics. You need to figure out how much error you made with your sample. Take the sample mean, and find the probability of getting another sample mean less than it, assuming for the moment that the manufacturer is right. The idea behind this is that you want to know what is the chance that you could have come up with your sample mean even if the population mean really is 500 days.

You want to find $P(\overline{x} < 490 | H_o \text{ is true }) = P(\overline{x} < 490 | \mu = 500)$

To compute this probability, you need to know how the sample mean is distributed. Since the sample size is at least 30, then you know the sample mean is approximately normally distributed. Remember $\mu_{\overline{x}} = \mu$ and $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

A picture is always useful.

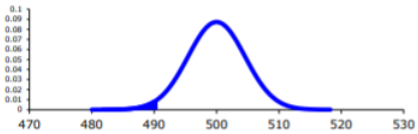


Figure 7.1.1

Before calculating the probability, it is useful to see how many standard deviations away from the mean the sample mean is. Using the formula for the z-score from chapter 6, you find

$$z = \frac{\overline{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{490 - 500}{25/\sqrt{30}} = -2.19$$

This sample mean is more than two standard deviations away from the mean. That seems pretty far, but you should look at the probability too.

On TI-83/84:

$$P(\overline{x} < 490 | \mu = 500) = \text{ normalcdf } (-1E99, 490, 500, 25 \div \sqrt{30}) \approx 0.0142$$

On R:

$$P(\overline{x} < 490 \mu = 500) = \text{ pnorm } (490, 500, 25/\text{ sqrt}(30)) \approx 0.0142$$

There is a 1.42% chance that you could find a sample mean less than 490 when the population mean is 500 days. This is really small, so the chances are that the assumption that the population mean is 500 days is wrong, and you can reject the manufacturer's claim. But how do you quantify really small? Is 5% or 10% or 15% really small? How do you decide?

Before you answer that question, a couple more definitions are needed.

Definition 7.1.3

Test Statistic: $z = \dfrac{\overline{x} - \mu_o}{\sigma/\sqrt{n}}$ since it is calculated as part of the testing of the hypothesis.

Definition 7.1.4

p – value: probability that the test statistic will take on more extreme values than the observed test statistic, given that the null hypothesis is true. It is the probability that was calculated above.

Now, how small is small enough? To answer that, you really want to know the types of errors you can make.

There are actually only two errors that can be made. The first error is if you say that H_o is false, when in fact it is true. This means you reject H_o when H_o was true. The second error is if you say that H_o is true, when in fact it is false. This means you fail to reject H_o when H_o is false. The following table organizes this for you:

Type of errors:

Table 7.1.2: Types of Errors

	H_o true	H_o false
Reject H_o	Type 1 error	No error
Fail to reject H_o	No error	Type II error

Thus

Definition 7.1.5

Type I Error is rejecting H_o when H_o is true, and

Definition 7.1.6

Type II Error is failing to reject H_o when H_o is false.

Since these are the errors, then one can define the probabilities attached to each error.

Definition 7.1.7

α = P(type I error) = P(rejecting H_o / H_o is true)

Definition 7.1.8

β = P(type II error) = P(failing to reject H_o / H_o is false)

α is also called the **level of significance**.

Another common concept that is used is Power = $1 - \beta$.

Now there is a relationship between α and β. They are not complements of each other. How are they related?

If α increases that means the chances of making a type I error will increase. It is more likely that a type I error will occur. It makes sense that you are less likely to make type II errors, only because you will be rejecting H_o more often. You will be failing to reject H_o less, and therefore, the chance of making a type II error will decrease. Thus, as α increases, β will decrease, and vice versa. That makes them seem like complements, but they aren't complements. What gives? Consider one more factor – sample size.

Consider if you have a larger sample that is representative of the population, then it makes sense that you have more accuracy then with a smaller sample. Think of it this way, which would you trust more, a sample mean of 490 if you had a sample size of 35 or sample size of 350 (assuming a representative sample)? Of course the 350 because there are more data points and so more accuracy. If you are more accurate, then there is less chance that you will make any error. By increasing the sample size of a representative sample, you decrease both α and β.

Summary of all of this:

1. For a certain sample size, *n*, if α increases, β decreases.
2. For a certain level of significance, α, if *n* increases, β decreases.

Now how do you find α and β? Well α is actually chosen. There are only three values that are usually picked for α: 0.01, 0.05, and 0.10. β is very difficult to find, so usually it isn't found. If you want to make sure it is small you take as large of a sample as you can afford provided it is a representative sample. This is one use of the Power. You want β to be small and the Power of the test is large. The Power word sounds good.

Which pick of α do you pick? Well that depends on what you are working on. Remember in this example you are the buyer who is trying to get out of a contract to buy these batteries. If you create a type I error, you said that the batteries are bad when they aren't, most likely the manufacturer will sue you. You want to avoid this. You might pick α to be 0.01. This way you have a small chance of making a type I error. Of course this means you have more of a chance of making a type II error. No big deal right? What if the batteries are used in pacemakers and you tell the person that their pacemaker's batteries are good for 500 days when they actually last less, that might be bad. If you make a type II error, you say that the batteries do last 500 days when they last less, then you have the possibility of killing someone. You certainly do not want to do this. In this case you might want to pick α as 0.10. If both errors are equally bad, then pick α as 0.05.

The above discussion is why the choice of α depends on what you are researching. As the researcher, you are the one that needs to decide what α level to use based on your analysis of the consequences of making each error is.

Note

If a type I error is really bad, then pick α = 0.01.

If a type II error is really bad, then pick α = 0.10

If neither error is bad, or both are equally bad, then pick α = 0.05

The main thing is to always pick the α before you collect the data and start the test.

The above discussion was long, but it is really important information. If you don't know what the errors of the test are about, then there really is no point in making conclusions with the tests. Make sure you understand what the two errors are and what the probabilities are for them.

Now it is time to go back to the example and put this all together. This is the basic structure of testing a hypothesis, usually called a hypothesis test. Since this one has a test statistic involving z, it is also called a z-test. And since there is only one sample, it is usually called a one-sample z-test.

Example 7.1.2 battery example revisited

1. State the random variable and the parameter in words.
2. State the null and alternative hypothesis and the level of significance.
3. State and check the assumptions for a hypothesis test.
    a. A random sample of size *n* is taken.

b. The population standard derivation is known.
c. The sample size is at least 30 or the population of the random variable is normally distributed.

4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x = life of battery

μ = mean life of a XJ35 battery

2. $H_o : \mu = 500$ days

$H_A : \mu < 500$ days

$\alpha = 0.10$ (from above discussion about consequences)

3. Every hypothesis has some assumptions that be met to make sure that the results of the hypothesis are valid. The assumptions are different for each test. This test has the following assumptions.

a. This occurred in this example, since it was stated that a random sample of 30 battery lives were taken.
b. This is true, since it was given in the problem.
c. The sample size was 30, so this condition is met.

4. The test statistic depends on how many samples there are, what parameter you are testing, and assumptions that need to be checked. In this case, there is one sample and you are testing the mean. The assumptions were checked above.

Sample statistic:

$$\overline{x} = 490$$

Test statistic:

$$z = \frac{\overline{x} - \mu_o}{\sigma/\sqrt{n}} = \frac{490 - 500}{25/\sqrt{30}} = -2.19$$

p-value:

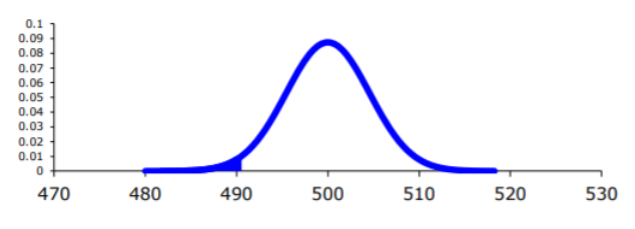


Figure 7.1.2

Using TI-83/84:

$$P(\overline{x} < 490 | \mu = 500) = \text{ normalcdf } (-1\text{E}99, 490, 500, 25/\sqrt{30}) \approx 0.0142$$

Using R:

$$P(\overline{x} < 490 | \mu = 500) = \text{pnorm}(490, 500, 25/\text{ sqrt}(30)) \approx 0.0142$$

5. Now what? Well, this p-value is 0.0142. This is a lot smaller than the amount of error you would accept in the problem -α = 0.10. That means that finding a sample mean less than 490 days is unusual to happen if H_o is true. This should make you think that H_o is not true. You should reject H_o.

Note

In fact, in general:

Reject H_o if the p-value < α and

Fail to reject H_o if the p-value $\geq \alpha$.

> 6. Since you rejected H_o, what does this mean in the real world? That is what goes in the interpretation. Since you rejected the claim by the manufacturer that the mean life of the batteries is 500 days, then you now can believe that your hypothesis was correct. In other words, there is enough evidence to show that the mean life of the battery is less than 500 days.

Now that you know that the batteries last less than 500 days, should you cancel the contract? Statistically, there is evidence that the batteries do not last as long as the manufacturer says they should. However, based on this sample there are only ten days less on average that the batteries last. There may not be practical significance in this case. Ten days do not seem like a large difference. In reality, if the batteries are used in pacemakers, then you would probably tell the patient to have the batteries replaced every year. You have a large buffer whether the batteries last 490 days or 500 days. It seems that it might not be worth it to break the contract over ten days. What if the 10 days was practically significant? Are there any other things you should consider? You might look at the business relationship with the manufacturer. You might also look at how much it would cost to find a new manufacturer. These are also questions to consider before making any changes. What this discussion should show you is that just because a hypothesis has statistical significance does not mean it has practical significance. The hypothesis test is just one part of a research process. There are other pieces that you need to consider.

That's it. That is what a hypothesis test looks like. All hypothesis tests are done with the same six steps. Those general six steps are outlined below.

1. State the random variable and the parameter in words. This is where you are defining what the unknowns are in this problem.
   x = random variable
   μ = mean of random variable, if the parameter of interest is the mean. There are other parameters you can test, and you would use the appropriate symbol for that parameter.
2. State the null and alternative hypotheses and the level of significance
   $H_o : \mu = \mu_o$, where μ_o is the known mean
   $H_A : \mu < \mu_o$
   $H_A : \mu > \mu_o$, use the appropriate one for your problem
   $H_A : \mu \neq \mu_o$
   Also, state your α level here.
3. State and check the assumptions for a hypothesis test.
   Each hypothesis test has its own assumptions. They will be stated when the different hypothesis tests are discussed.
4. Find the sample statistic, test statistic, and p-value.
   This depends on what parameter you are working with, how many samples, and the assumptions of the test. The p-value depends on your H_A. If you are doing the H_A with the less than, then it is a left-tailed test, and you find the probability of being in that left tail. If you are doing the H_A with the greater than, then it is a right-tailed test, and you find the probability of being in the right tail. If you are doing the H_A with the not equal to, then you are doing a two-tail test, and you find the probability of being in both tails. Because of symmetry, you could find the probability in one tail and double this value to find the probability in both tails.
5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. Interpretation
   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Sorry, one more concept about the conclusion and interpretation. First, the conclusion is that you reject H_o or you fail to reject H_o. Why was it said like this? It is because you never **accept** the null hypothesis. If you wanted to accept the null hypothesis, then why do the test in the first place? In the interpretation, you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true. You wouldn't want to go to all this work and then find out you wanted to accept the claim. Why go through the trouble? You always want to show that the alternative hypothesis is true. Sometimes you can do that and sometimes you can't. It doesn't mean you proved the null hypothesis; it just means you can't prove the alternative hypothesis. Here is an example to demonstrate this.

Example 7.1.3 conclusion in hypothesis tests

In the U.S. court system a jury trial could be set up as a hypothesis test. To really help you see how this works, let's use OJ Simpson as an example. In the court system, a person is presumed innocent until he/she is proven guilty, and this is your null hypothesis. OJ Simpson was a football player in the 1970s. In 1994 his ex-wife and her friend were killed. OJ Simpson was accused of the crime, and in 1995 the case was tried. The prosecutors wanted to prove OJ was guilty of killing his wife and her friend, and that is the alternative hypothesis

Solution

H_0: OJ is innocent of killing his wife and her friend

H_A: OJ is guilty of killing his wife and her friend

In this case, a verdict of not guilty was given. That does not mean that he is innocent of this crime. It means there was not enough evidence to prove he was guilty. Many people believe that OJ was guilty of this crime, but the jury did not feel that the evidence presented was enough to show there was guilt. The verdict in a jury trial is always guilty or not guilty!

The same is true in a hypothesis test. There is either enough or not enough evidence to show that alternative hypothesis. It is not that you proved the null hypothesis true.

When identifying hypothesis, it is important to state your random variable and the appropriate parameter you want to make a decision about. If count something, then the random variable is the number of whatever you counted. The parameter is the proportion of what you counted. If the random variable is something you measured, then the parameter is the mean of what you measured. (Note: there are other parameters you can calculate, and some analysis of those will be presented in later chapters.)

Example 7.1.4 stating hypotheses

Identify the hypotheses necessary to test the following statements:

a. The average salary of a teacher is more than $30,000.
b. The proportion of students who like math is less than 10%.
c. The average age of students in this class differs from 21.

Solution

a. x = salary of teacher

μ = mean salary of teacher

The guess is that $\mu > \$30,000$ and that is the alternative hypothesis.

The null hypothesis has the same parameter and number with an equal sign.

$$H_0 : \mu = \$30,000$$
$$H_A : \mu > \$30,000$$

b. x = number od students who like math

p = proportion of students who like math

The guess is that p < 0.10 and that is the alternative hypothesis.

$$H_0 : p = 0.10$$
$$H_A : p < 0.10$$

c. x = age of students in this class

μ = mean age of students in this class

The guess is that $\mu \neq 21$ and that is the alternative hypothesis.

$$H_0 : \mu = 21$$
$$H_A : \mu \neq 21$$

Example 7.1.5 Stating Type I and II Errors and Picking Level of Significance

a. The plant-breeding department at a major university developed a new hybrid raspberry plant called YumYum Berry. Based on research data, the claim is made that from the time shoots are planted 90 days on average are required to obtain the first berry with a standard deviation of 9.2 days. A corporation that is interested in marketing the product tests 60 shoots by planting them and recording the number of days before each plant produces its first berry. The sample mean is 92.3 days. The corporation wants to know if the mean number of days is more than the 90 days claimed. State the type I and type II errors in terms of this problem, consequences of each error, and state which level of significance to use.

b. A concern was raised in Australia that the percentage of deaths of Aboriginal prisoners was higher than the percent of deaths of non-indigenous prisoners, which is 0.27%. State the type I and type II errors in terms of this problem, consequences of each error, and state which level of significance to use.

Solution

a. x = time to first berry for YumYum Berry plant

μ = mean time to first berry for YumYum Berry plant

$$H_0 : \mu = 90$$
$$H_A : \mu > 90$$

Type I Error: If the corporation does a type I error, then they will say that the plants take longer to produce than 90 days when they don't. They probably will not want to market the plants if they think they will take longer. They will not market them even though in reality the plants do produce in 90 days. They may have loss of future earnings, but that is all.

Type II error: The corporation do not say that the plants take longer then 90 days to produce when they do take longer. Most likely they will market the plants. The plants will take longer, and so customers might get upset and then the company would get a bad reputation. This would be really bad for the company.

Level of significance: It appears that the corporation would not want to make a type II error. Pick a 10% level of significance, $\alpha = 0.10$.

b. x = number of Aboriginal prisoners who have died

p = proportion of Aboriginal prisoners who have died

$$H_o : p = 0.27\%$$
$$H_A : p > 0.27\%$$

Type I error: Rejecting that the proportion of Aboriginal prisoners who died was 0.27%, when in fact it was 0.27%. This would mean you would say there is a problem when there isn't one. You could anger the Aboriginal community, and spend time and energy researching something that isn't a problem.

Type II error: Failing to reject that the proportion of Aboriginal prisoners who died was 0.27%, when in fact it is higher than 0.27%. This would mean that you wouldn't think there was a problem with Aboriginal prisoners dying when there really is a problem. You risk causing deaths when there could be a way to avoid them.

Level of significance: It appears that both errors may be issues in this case. You wouldn't want to anger the Aboriginal community when there isn't an issue, and you wouldn't want people to die when there may be a way to stop it. It may be best to pick a 5% level of significance, $\alpha = 0.05$.

Note

Hypothesis testing is really easy if you follow the same recipe every time. The only differences in the various problems are the assumptions of the test and the test statistic you calculate so you can find the p-value. Do the same steps, in the same order, with the same words, every time and these problems become very easy.

Homework

Exercise 7.1.1

For the problems in this section, a question is being asked. This is to help you understand what the hypotheses are. You are not to run any hypothesis tests and come up with any conclusions in this section.

1. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made in a given time period and found that 11% of all lenses had defects of some type. Looking at the type of defects, they found in a three-month time period that out of 34,641 defective lenses, 5865 were due to scratches. Are there more defects from scratches than from all other causes? State the random variable, population parameter, and hypotheses.
2. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints ("Consumer fraud and," 2008). Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? State the random variable, population parameter, and hypotheses.
3. The Kyoto Protocol was signed in 1997, and required countries to start reducing their carbon emissions. The protocol became enforceable in February 2005. In 2004, the mean CO2 emission was 4.87 metric tons per capita. Is there enough evidence to show that the mean CO2 emission is lower in 2010 than in 2004? State the random variable, population parameter, and hypotheses.
4. The FDA regulates that fish that is consumed is allowed to contain 1.0 mg/kg of mercury. In Florida, bass fish were collected in 53 different lakes to measure the amount of mercury in the fish. The data for the average amount of mercury in each lake is in Example 7.1.5 ("Multi-disciplinary niser activity," 2013). Do the data provide enough evidence to show that the fish in Florida lakes has more mercury than the allowable amount? State the random variable, population parameter, and hypotheses.
5. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made in a given time period and found that 11% of all lenses had defects of some type. Looking at the type of defects, they found in a three-month time period that out of 34,641 defective lenses, 5865 were due to scratches. Are there more defects from scratches than from all other causes? State the type I and type II errors in this case, consequences of each error type for this situation from the perspective of the manufacturer, and the appropriate alpha level to use. State why you picked this alpha level.
6. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints ("Consumer fraud and," 2008). Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? State the type I and type II errors in this case, consequences of each error type for this situation from the perspective of the state of Arizona, and the appropriate alpha level to use. State why you picked this alpha level.
7. The Kyoto Protocol was signed in 1997, and required countries to start reducing their carbon emissions. The protocol became enforceable in February 2005. In 2004, the mean CO2 emission was 4.87 metric tons per capita. Is there enough evidence to show that the mean CO2 emission is lower in 2010 than in 2004? State the type I and type II errors in this case, consequences of each error type for this situation from the perspective of the agency overseeing the protocol, and the appropriate alpha level to use. State why you picked this alpha level.
8. The FDA regulates that fish that is consumed is allowed to contain 1.0 mg/kg of mercury. In Florida, bass fish were collected in 53 different lakes to measure the amount of mercury in the fish. The data for the average amount of mercury in each lake is in Example 7.1.5 ("Multi-disciplinary niser activity," 2013). Do the data provide enough evidence to show that the fish in Florida lakes has more mercury than the allowable amount? State the type I and type II errors in this case, consequences of each error type for this situation from the perspective of the FDA, and the appropriate alpha level to use. State why you picked this alpha level.

Answer

1. $H_o : p = 0.11, H_A : p > 0.11$
3. $H_o : \mu = 4.87$ metric tons per capita, $H_A : \mu < 4.87$ metric tons per capita
5. See solutions
7. See solutions

This page titled 7.1: Basics of Hypothesis Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

7.2: One-Sample Proportion Test

There are many different parameters that you can test. There is a test for the mean, such as was introduced with the z-test. There is also a test for the population proportion, p. This is where you might be curious if the proportion of students who smoke at your school is lower than the proportion in your area. Or you could question if the proportion of accidents caused by teenage drivers who do not have a drivers' education class is more than the national proportion.

To test a population proportion, there are a few things that need to be defined first. Usually, Greek letters are used for parameters and Latin letters for statistics. When talking about proportions, it makes sense to use *p* for proportion. The Greek letter for *p* is π, but that is too confusing to use. Instead, it is best to use *p* for the population proportion. That means that a different symbol is needed for the sample proportion. The convention is to use, $\hat{p}$, known as p-hat. This way you know that *p* is the population proportion, and that $\hat{p}$ is the sample proportion related to it.

Now proportion tests are about looking for the percentage of individuals who have a particular attribute. You are really looking for the number of successes that happen. Thus, a proportion test involves a binomial distribution.

Hypothesis Test for One Population Proportion (1-Prop Test)

1. State the random variable and the parameter in words.
   x = number of successes
   I = proportion of successes
2. State the null and alternative hypotheses and the level of significance
   $H_o : p = p_o$, where p_o is the known proportion
   $H_A : p < p_o$
   $H_A : p > p_o$, use the appropriate one for your problem
   $H_A : p \neq p_o$
   Also, state your α level here.
3. State and check the assumptions for a hypothesis test
   a. A simple random sample of size *n* is taken.
   b. The conditions for the binomial distribution are satisfied
   c. To determine the sampling distribution of $\hat{p}$, you need to show that $np \geq 5$ and $nq \geq 5$, where $q = 1 - p$. If this requirement is true, then the sampling distribution of $\hat{p}$ is well approximated by a normal curve.
4. Find the sample statistic, test statistic, and p-value
   Sample Proportion:
   $\hat{p} = \frac{x}{n} = \frac{\#\text{ of successes}}{\#\text{ of trials}}$
   Test Statistic:
   $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$
   p-value:
   TI-83/84: Use normalcdf(lower limit, upper limit, 0, 1)

   > **Note**
   > if $H_A : p < p_o$, then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : p > p_o$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : p \neq p_o$, then find the p-value for $H_A : p < p_o$, and multiply by 2.

   R: Use pnorm(z, 0, 1)

   > **Note**
   > If $H_A : p < p_o$, then you can use pnorm. If $H_A : p > p_o$, then you have to find pnorm and then subtract from 1. If $H_A : p \neq p_o$, then find the p-value for $H_A : p < p_o$, and multiply by 2.

5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail

to reject H_o.

6. Interpretation

   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Example 7.2.1 hypothesis test for one proportion using formula

A concern was raised in Australia that the percentage of deaths of Aboriginal prisoners was higher than the percent 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). Do the data provide enough evidence to show that the proportion of deaths of Aboriginal prisoners is more than 0.27%?

1. State the random variable and the parameter in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x = number of Aboriginal prisoners who die

p = proportion of Aboriginal prisoners who die

2. $H_o : p = 0.0027$
   $H_A : p > 0.0027$

Example 7.2.5b argued that the $\alpha = 0.05$.

3.

a. A simple random sample of 14,495 Aboriginal prisoners was taken. However, the sample was not a random sample, since it was data from six years. It is the numbers for all prisoners in these six years, but the six years were not picked at random. Unless there was something special about the six years that were chosen, the sample is probably a representative sample. 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 conditions for the binomial distribution are satisfied
c. In this case p = 0.0027 and n = 14,495. $np = 14495^*0.0027 \approx 39 \geq 5$ and $nq = 14495^*(1 - 0.0027) \approx 14456 \geq 5$. So, the sampling distribution for $\hat{p}$ is a normal distribution.

4. Sample Proportion:

x = 51

n = 14495

$$\hat{p} = \frac{x}{n} = \frac{51}{14495} \approx 0.003518$$

Test Statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.003518 - 0.0027}{\sqrt{\frac{0.0027(1 - 0.0027)}{14495}}} \approx 1.8979$$

p-value:

TI-83/84: p-value = $P(z > 1.8979) =$ normalcdf $(1.8979, 1E99, 0, 1) \approx 0.029$

R: p-value = $P(z > 1.8979) = 1 -$ pnorm $(1.8979, 0, 1) \approx 0.029$

5. Since the p-value < 0.05, then reject H_o.

6. There is enough evidence to show that the proportion of deaths of Aboriginal prisoners is more than for non-Aboriginal prisoners.

Example 7.2.2 hypothesis test for one proportion using technology

A researcher who is studying the effects of income levels on 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 breastfeed. In Tajikistan, considered a low-income country by the World Bank, researchers found that in a random sample of 500 new mothers that 125 were breastfeeding their infant. At the 5% level of significance, does this show that low-income countries have a higher incident of breastfeeding?

1. State you random variable and the parameter in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x = number of woman who breastfeed in a low-income country

p = proportion of woman who breastfeed in a low-income country

2. $H_o : p = 0.22$
   $H_A : p > 0.22$
   $\alpha = 0.05$

3.

a. A simple random sample of 500 breastfeeding habits of woman 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 conditions for the binomial distribution are satisfied
c. In this case, n = 500 and p = 0.22. $np = 500(0.22) = 110 \geq 5$ and $nq = 500(1 - 0.22) = 390 \geq 5$, so the sampling distribution of $\hat{p}$ is well approximated by a normal curve.

4. This time, all calculations will be done with technology. On the TI-83/84 calculator. Go into the STAT menu, then arrow over to TESTS. This test is a 1-propZTest. Then type in the information just as shown in *Figure* 7.2.1.

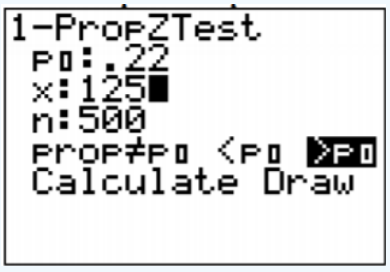


Figure 7.2.1: Setup for 1-Proportion Test

Once you press Calculate, you will see the results as in *Figure* 7.2.2.

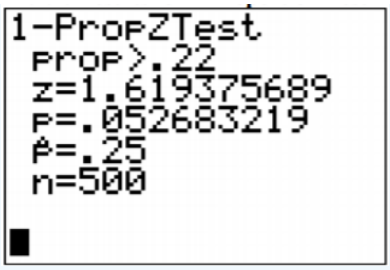


Figure 7.2.2: Results for 1-Proportion Test

The z in the results is the test statistic. The p = 0.052683219 is the p-value, and the $\hat{p} = 0.25$ is the sample proportion.

The p-value is approximately 0.053.

On R, the command is prop.test(x, n, po, alternative = "less" or "greater"), where po is what H_o says p equals, and you use less if your H_A is less and greater if your H_A is greater. If your H_A is not equal to, then leave off the alternative statement. So for this example, the command would be prop.test(125, 500, .22, alternative = "greater")

1-sample proportions test with continuity correction

data: 125 out of 500, null probability 0.22

X-squared = 2.4505, df = 1, p-value = 0.05874

alternative hypothesis: true p is greater than 0.22

95 percent confidence interval:

0.218598 1.000000

sample estimates:

p

0.25

Note

R does a continuity correction that the formula and the TI-83/84 calculator do not do. You can put in a command that says not to use the continuity correction, but it is correct to use it. Also, R doesn't give the z test statistic, so you don't need to worry about this. It does give a p-value that is slightly off from the formula and the calculator due to the continuity correction.

p-value = 0.05874

5. Since the p-value is more than 0.05, you fail to reject H_o.

6. There is not enough evidence to show that the proportion of women who breastfeed in low-income countries is more than in high-income countries.

Notice, the conclusion is that there wasn't enough evidence to show what H_1 said. The conclusion was not that you proved H_o true. There are many reasons why you can't say that H_o is true. It could be that the countries you chose were not very representative of what truly happens. If you instead looked at all high-income countries and compared them to low-income countries, you might have different results. It could also be that the sample you collected in the low-income country was not representative. It could also be that income level is not an indication of breastfeeding habits. There could be other factors involved. This is why you can't say that you have proven H_o is true. There are too many other factors that could be the reason that you failed to reject H_o.

Homework

Exercise 7.2.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made in a given time period and found that 11% of all lenses had defects of some type. Looking at the type of defects, they found in a three-month time period that out of 34,641 defective lenses, 5865 were due to scratches. Are there more defects from scratches than from all other causes? Use a 1% level of significance.
2. In July of 1997, Australians were asked if they thought unemployment would increase, and 47% thought that it would increase. In November of 1997, they were asked again. At that time 284 out of 631 said that they thought unemployment would increase ("Morgan gallup poll," 2013). At the 5% level, is there enough evidence to show that the proportion of Australians in November 1997 who believe unemployment would increase is less than the proportion who felt it would increase in July 1997?
3. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Arkansas had 1,601 complaints of identity theft out of 3,482 consumer complaints ("Consumer fraud and," 2008). Does this data provide enough evidence to show that Arkansas had a higher proportion of identity theft than 23%? Test at the 5% level.
4. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, 23% of all complaints in 2007 were for identity theft. In that year, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints ("Consumer fraud and," 2008). Does this data provide enough evidence to show that Alaska had a lower proportion of identity theft than 23%? Test at the 5% level.
5. In 2001, the Gallup poll found that 81% of American adults believed that there was a conspiracy in the death of President Kennedy. In 2013, the Gallup poll asked 1,039 American adults if they believe there was a conspiracy in the assassination, and found that 634 believe there was a conspiracy ("Gallup news service," 2013). Do the data show that the proportion of Americans who believe in this conspiracy has decreased? Test at the 1% level.
6. In 2008, there were 507 children in Arizona out of 32,601 who were diagnosed with Autism Spectrum Disorder (ASD) ("Autism and developmental," 2008). Nationally 1 in 88 children are diagnosed with ASD ("CDC features -," 2013). Is there sufficient data to show that the incident of ASD is more in Arizona than nationally? Test at the 1% level.

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Reject Ho.

3. Reject Ho.

5. Reject Ho.

This page titled 7.2: One-Sample Proportion Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

7.3: One-Sample Test for the Mean

It is time to go back to look at the test for the mean that was introduced in section 7.1 called the z-test. In the example, you knew what the population standard deviation, σ, was. What if you don't know σ?

You could just use the sample standard deviation, *s*, as an approximation of σ. That means the test statistic is now $\frac{\overline{x}-\mu}{s/\sqrt{n}}$. Great, now you can go and find the p-value using the normal curve. Or can you? Is this new test statistic normally distributed? Actually, it is not. How is it distributed? A man named W. S. Gossett figured out what this distribution is and called it the Student's t-distribution. There are some assumptions that must be made for this formula to be a Student's t-distribution. These are outlined in the following theorem. Note: the t-distribution is called the Student's t-distribution because that is the name he published under because he couldn't publish under his own name due to employer not wanting him to publish under his own name. His employer by the way was Guinness and they didn't want competitors knowing they had a chemist working for them. It is not called the Student's t-distribution because it is only used by students.

Theorem: If the following assumptions are met

a. A random sample of size *n* is taken.
b. The distribution of the random variable is normal or the sample size is 30 or more.

Then the distribution of $t=\frac{\overline{x}-\mu}{s/\sqrt{n}}$ is a Student's t-distribution with $n-1$ degrees of freedom.

Explanation of degrees of freedom:

Recall the formula for sample standard deviation is $s=\sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$. Notice the denominator is *n* - 1. This is the same as the degrees of freedom. This is no accident. The reason the denominator and the degrees of freedom are both *n* -1 comes from how the standard deviation is calculated. Remember, first you take each data value and subtract $\overline{x}$. If you add up all of these new values, you will get 0. This must happen. Since it must happen, the first *n* - 1 data values you have "freedom of choice", but the nth data value, you have no freedom to choose. Hence, you have *n* - 1 degrees of freedom. Another way to think about it is that if you five people and five chairs, the first four people have a choice of where they are sitting, but the last person does not. They have no freedom of where to sit. Only 5 - 1 =4 people have freedom of choice.

The Student's t-distribution is a bell-shape that is more spread out than the normal distribution. There are many t-distributions, one for each different degree of freedom.

Here is a graph of the normal distribution and the Student's t-distribution for df = 1 and df = 2.

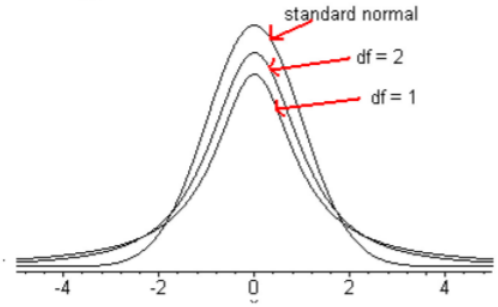


Figure 7.3.1: Typical Student t-Distributions

As the degrees of freedom increases, the student's t-distribution looks more like the normal distribution.

To find probabilities for the t-distribution, again technology can do this for you. There are many technologies out there that you can use. On the TI-83/84, the command is in the DISTR menu and is tcdf(. The syntax for this command is

tcdf(lower limit, upper limit, *df*)

On R: the command to find the area to the left of a t value is pt(t value, df)

Hypothesis Test for One Population Mean (t-Test)

1. State the random variable and the parameter in words.
   x = random variable
   μ = mean of random variable
2. State the null and alternative hypotheses and the level of significance
   $H_o : \mu = \mu_o$, where μ_o is the known mean
   $H_A : \mu < \mu_o$
   $H_A : \mu > \mu_o$, use the appropriate one for your problem
   $H_A : \mu \neq \mu_o$
   Also, state your α level here.
3. State and check the assumptions for a hypothesis test
   a. A random sample of size *n* is taken.
   b. The population of the random variable is normally distributed, though the t-test is fairly robust to the condition if the sample size is large. This means that if this condition isn't met, but your sample size is quite large (over 30), then the results of the t-test are valid.
   c. The population standard deviation, σ, is unknown.
4. Find the sample statistic, test statistic, and p-value
   Test Statistic:
   $t = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$
   with degrees of freedom *df* = *n* - 1
   p-value:
   Using TI-83/84: tcdf(lower limit, upper limit, *df*)

   Note

   If $H_A : \mu < \mu_o$, then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : \mu > \mu_o$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : \mu \neq \mu_o$, then find the p-value for $H_A : \mu < \mu_o$, and multiply by 2.

   Using R: pt(t value, df)

   Note

   If $H_A : \mu < \mu_o$, then the command is pt(t value, df). If $H_A : \mu > \mu_o$, then the command is $1 - \text{pt(t value, df)}$. If $H_A : \mu \neq \mu_o$, then find the p-value for $H_A : \mu < \mu_o$, and multiply by 2.
5. This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

How to check the assumptions of t-test:

In order for the t-test to be valid, the assumptions of the test must be true. Whenever you run a t-test, you must make sure the assumptions are true. You need to check them. Here is how you do this:

1. For the condition that the sample is a random sample, describe how you took the sample. Make sure your sampling technique is random.
2. For the condition that population of the random variable is normal, remember the process of assessing normality from chapter 6.

Note

If the assumptions behind this test are not valid, then the conclusions you make from the test are not valid. If you do not have a random sample, that is your fault. Make sure the sample you take is as random as you can make it following sampling techniques from chapter 1. If the population of the random variable is not normal, then take a sample larger than 30. If you

cannot afford to do that, or if it is not logistically possible, then you do different tests called non-parametric tests. There is an entire course on non-parametric tests, and they will not be discussed in this book.

Example 7.3.1 test of the mean using the formula

A random sample of 20 IQ scores of famous people was taken from the website of IQ of Famous People ("IQ of famous," 2013) and a random number generator was used to pick 20 of them. The data are in Example 7.3.1. Do the data provide evidence at the 5% level that the IQ of a famous person is higher than the average IQ of 100?

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 the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x = IQ score of a famous person

μ = mean IQ score of a famous person

2. $H_o : \mu = 100$
   $H_A : \mu > 100$
   $\alpha = 0.05$

3.

a. A random sample of 20 IQ scores was taken. This was said in the problem.
b. The population of IQ score is normally distributed. This was shown in Example 7.3.2.

4. Sample Statistic:

$\overline{x} = 145.4$
$s \approx 29.27$

Test Statistic:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{145.4 - 100}{\frac{29.27}{\sqrt{20}}} \approx 6.937$$

p-value:

df = *n* - 1 = 20 - 1 = 19

TI-83/84: p-value = $\text{tcdf}(6.937, 1E99, 19) = 6.5 \times 10^{-7}$

R: p-value = $1 - \text{pt}(6.937, 19) = 6.5 \times 10^{-7}$

5. Since the p-value is less than 5%, then reject H_o.

6. There is enough evidence to show that famous people have a higher IQ than the average IQ of 100.

Example 7.3.2 test of the mean using technology

In 2011, the average life expectancy for a woman in Europe was 79.8 years. The data in Example 7.3.2 are the life expectancies for men in European countries in 2011 ("WHO life expectancy," 2013). Do the data indicate that men's life expectancy is less than women's? Test at the 1% level.

Table 7.3.2: Life Expectancies for Men in European Countries in 2011

1. State the random variable and the parameter in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x = life expectancy for a European man in 2011

μ = mean life expectancy for European men in 2011

2. $H_o : \mu = 79.8 \text{ years}$
   $H_A : \mu < 79.8 \text{ years}$
   $\alpha = 0.01$

3.

a. A random sample of 53 life expectancies of European men in 2011 was taken. The data is actually all of the life expectancies for every country that is considered part of Europe by the World Health Organization. However, the information is still sample information since it is only for one year that the data was collected. It may not be a random sample, but that is probably not an issue in this case.

b. The distribution of life expectancies of European men in 2011 is normally distributed. To see if this condition has been met, look at the histogram, number of outliers, and the normal probability plot. (If you wish, you can look at the normal probability plot first. If it doesn't look linear, then you may want to look at the histogram and number of outliers at this point.)

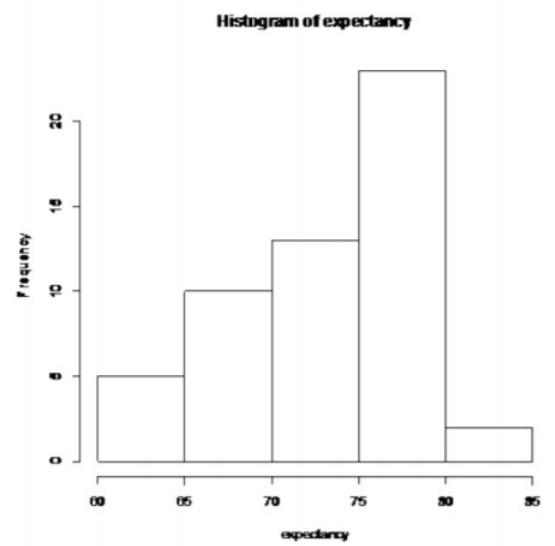


Figure 7.3.2: Histogram for Life Expectancies of European Men in 2011

Not bell shaped

Number of outliers:

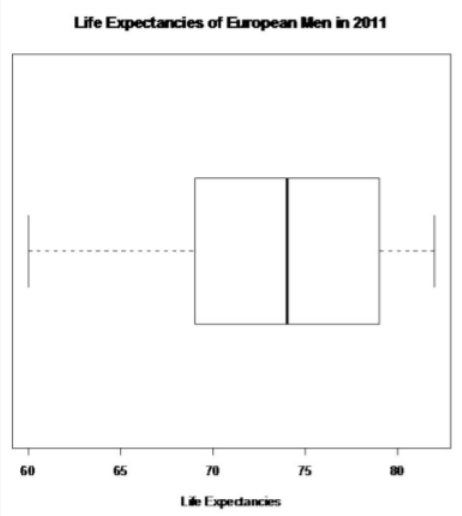


Figure 7.3.3: Modified Box Plot for Life Expectancies of European Men in 2011

or:

$IQR = 79 - 69 = 10$

$1.5 * IQR = 15$

$Q1 - 1.5 * IQR = 69 - 15 = 54$

$Q3 + 1.5 * IQR = 79 + 15 = 94$

Outliers are numbers below 54 and above 94. There are no outliers for this data set.

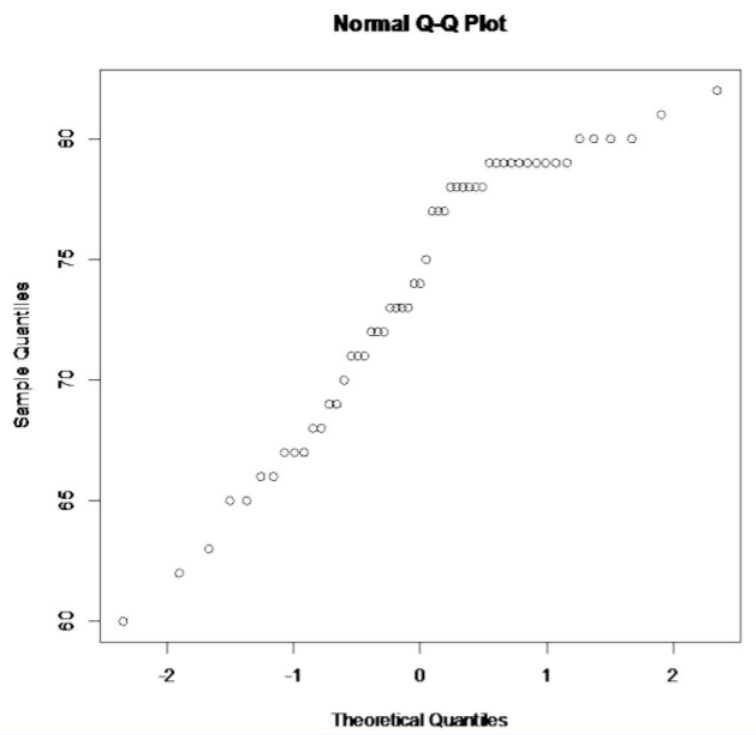


Figure 7.3.4: Normal Quantile Plot for Life Expectancies of European Men in 2011

Not linear

This population does not appear to be normally distributed. This sample is larger than 30, so it is good that the t-test is robust.

4. The calculations will be conducted using technology.

On the TI-83/84 calculator. Go into STAT and type the data into L1.

Then go into STAT and move over to TESTS. Choose T-Test. The setup for the calculator is in *Figure* 7.3.4.

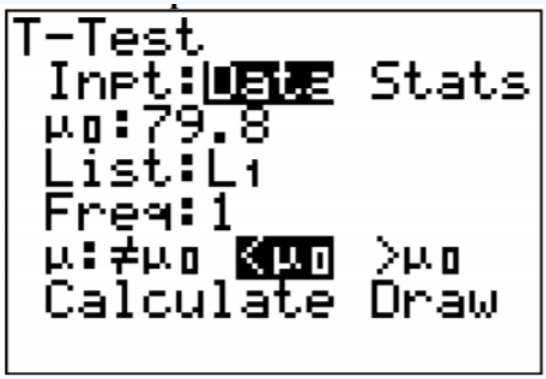


Figure 7.3.5: Setup for T-Test on TI-83/84 Calculator

Once you press ENTER on Calculate you will see the result shown in *Figure* 7.3.6.

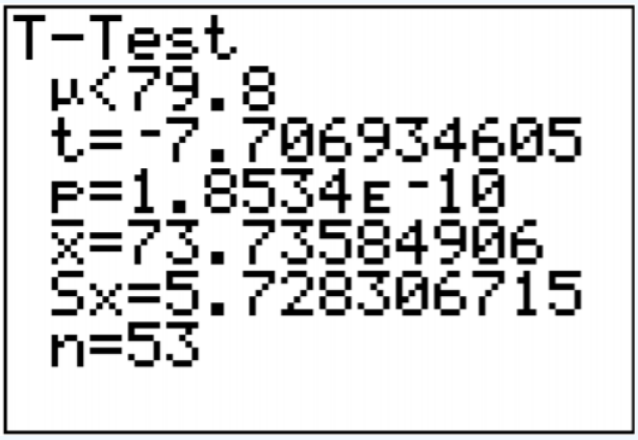


Figure 7.3.6: Result of T-Test on TI-83/84 Calculator

On R, the command is t.test(variable, mu = number in H_0, alternative = "less" or "greater"), where mu = what H_0 says the mean equals, and you use less if your H_A is less and greater if your H_A is greater. If your H_A is not equal to, then leave off the alternative statement. For this example, the command would be t.test(expectancy, mu=79.8, alternative = "less")

One Sample t-test

data: expectancy

t = -7.7069, df = 52, p-value = 1.853e-10

alternative hypothesis: true mean is less than 79.8

95 percent confidence interval:

-Inf 75.05357

sample estimates:

mean of x

73.73585

Most of the output you don't need. You need the test statistic and the p-value.

The *t* = -7.707 is the test statistic. The p-value is 1.8534×10^{-10}.

5. Since the p-value is less than 1%, then reject H_o.

6. There is enough evidence to show that the mean life expectancy for European men in 2011 was less than the mean life expectancy for European women in 2011 of 79.8 years.

Homework

Exercise 7.3.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. The Kyoto Protocol was signed in 1997, and required countries to start reducing their carbon emissions. The protocol became enforceable in February 2005. In 2004, the mean CO2 emission was 4.87 metric tons per capita. Table 7.3.3 contains a random sample of CO2 emissions in 2010 ("CO2 emissions," 2013). Is there enough evidence to show that the mean CO2 emission is lower in 2010 than in 2004? Test at the 1% level.

Table 7.3.3: CO2 Emissions (in metric tons per capita) in 2010

1.36	1.42	5.93	5.36	0.06	9.11	7.32
7.93	6.72	0.78	1.80	0.20	2.27	0.28
5.86	3.46	1.46	0.14	2.62	0.79	7.48
0.86	7.84	2.87	2.45			

2. The amount of sugar in a Krispy Kream glazed donut is 10 g. Many people feel that cereal is a healthier alternative for children over glazed donuts. Example 7.3.4 contains the amount of sugar in a sample of cereal that is geared towards children ("Healthy breakfast story," 2013). Is there enough evidence to show that the mean amount of sugar in children's cereal is more than in a glazed donut? Test at the 5% level.

Table 7.3.4: Sugar Amounts in Children's Cereal

10	14	12	9	13	13	13
11	12	15	9	10	11	3
6	12	15	12	12		

3. The FDA regulates that fish that is consumed is allowed to contain 1.0 mg/kg of mercury. In Florida, bass fish were collected in 53 different lakes to measure the amount of mercury in the fish. The data for the average amount of mercury in each lake is in Example 7.3.5 ("Multi-disciplinary niser activity," 2013). Do the data provide enough evidence to show that the fish in Florida lakes has more mercury than the allowable amount? Test at the 10% level.

Table 7.3.5: Average Mercury Levels (mg/kg) in Fish

1.23	1.33	0.04	0.44	1.20	0.27
0.48	0.19	0.83	0.81	0.81	0.5
0.49	1.16	0.05	0.15	0.19	0.77
1.08	0.98	0.63	0.56	0.41	0.73
0.34	0.59	0.34	0.84	0.50	0.34
0.28	0.34	0.87	0.56	0.17	0.18
0.19	0.04	0.49	1.10	0.16	0.10
0.48	0.21	0.86	0.52	0.65	0.27
0.94	0.40	0.43	0.25	0.27	

4. Stephen Stigler determined in 1977 that the speed of light is 299,710.5 km/sec. In 1882, Albert Michelson had collected measurements on the speed of light ("Student t-distribution," 2013). His measurements are given in Example 7.3.6. Is there evidence to show that Michelson's data is different from Stigler's value of the speed of light? Test at the 5% level.

Table 7.3.6: Speed of Light Measurements in (km/sec)

299883	299816	299778	299796	299682
299711	299611	299599	300051	299781
299578	299796	299774	299820	299772
299696	299573	299748	299748	299797
299851	299809	299723		

5. Example 7.3.7 contains pulse rates after running for 1 minute, collected from females who drink alcohol ("Pulse rates before," 2013). The mean pulse rate after running for 1 minute of females who do not drink is 97 beats per minute. Do the data show that the mean pulse rate of females who do drink alcohol is higher than the mean pulse rate of females who do not drink? Test at the 5% level.

Table 7.3.7: Pulse Rates of Woman Who Use Alcohol

176	150	150	115	129	160
120	125	89	132	120	120
68	87	88	72	77	84
92	80	60	67	59	64
88	74	68			

6. The economic dynamism, which is the index of productive growth in dollars for countries that are designated by the World Bank as middle-income are in Example 7.3.8 ("SOCR data 2008," 2013). Countries that are considered high-income have a mean economic dynamism of 60.29. Do the data show that the mean economic dynamism of middle-income countries is less than the mean for high-income countries? Test at the 5% level.

Table 7.3.8: Economic Dynamism of Middle Income Countries

25.8057	37.4511	51.915	43.6952	47.8506	43.7178	58.0767
41.1648	38.0793	37.7251	39.6553	42.0265	48.6159	43.8555
49.1361	61.9281	41.9543	44.9346	46.0521	48.3652	43.6252
50.9866	59.1724	39.6282	33.6074	21.6643		

7. In 1999, the average percentage of women who received prenatal care per country is 80.1%. Example 7.3.9 contains the percentage of woman receiving prenatal care in 2009 for a sample of countries ("Pregnant woman receiving," 2013). Do the data show that the average percentage of women receiving prenatal care in 2009 is higher than in 1999? Test at the 5% level.

Table 7.3.9: Percentage of Woman Receiving Prenatal Care

70.08	72.73	74.52	75.79	76.28	76.28
76.65	80.34	80.60	81.90	86.30	87.70
87.76	88.40	90.70	91.50	91.80	92.10
92.20	92.41	92.47	93.00	93.20	93.40
93.63	93.69	93.80	94.30	94.51	95.00

95.80	95.80	96.23	96.24	97.30	97.90
97.95	98.20	99.00	99.00	99.10	99.10
100.00	100.00	100.00	100.00	100.00	

8. Maintaining your balance may get harder as you grow older. A study was conducted to see how steady the elderly is on their feet. They had the subjects stand on a force platform and have them react to a noise. The force platform then measured how much they swayed forward and backward, and the data is in Example 7.3.10 ("Maintaining balance while," 2013). Do the data show that the elderly sway more than the mean forward sway of younger people, which is 18.125 mm? Test at the 5% level.

Table 7.3.10: Forward/Backward Sway (in mm) of Elderly Subjects

19	30	20	19	29	25	21	24	50

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Fail to reject Ho.
3. Fail to reject Ho.
5. Fail to reject Ho.
7. Reject Ho.

Data Sources:

Australian Human Rights Commission, (1996). *Indigenous deaths in custody 1989 - 1996.* Retrieved from website: www.humanrights.gov.au/public...deaths-custody

CDC features - new data on autism spectrum disorders. (2013, November 26). Retrieved from www.cdc.gov/features/countingautism/

Center for Disease Control and Prevention, Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network. (2008). *Autism and developmental disabilities monitoring network-2012*. Retrieved from website: www.cdc.gov/ncbddd/autism/doc...nityReport.pdf

CO2 emissions. (2013, November 19). Retrieved from http://data.worldbank.org/indicator/EN.ATM.CO2E.PC

Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-December 2007*. Retrieved from website: www.ftc.gov/opa/2008/02/fraud.pdf

Gallup news service. (2013, November 7-10). Retrieved from www.gallup.com/file/poll/1658...acy_131115.pdf

Healthy breakfast story. (2013, November 16). Retrieved from lib.stat.cmu.edu/DASL/Stories...Breakfast.html

IQ of famous people. (2013, November 13). Retrieved from http://www.kidsiqtestcenter.com/IQ-famous-people.html

Maintaining balance while concentrating. (2013, September 25). Retrieved from http://www.statsci.org/data/general/balaconc.html

Morgan Gallup poll on unemployment. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/gallup.html

Multi-disciplinary niser activity - mercury in bass. (2013, November 16). Retrieved from http://gozips.uakron.edu/~nmimoto/pa.../MercuryInBass - description.txt

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

SOCR data 2008 world countries rankings. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...ntriesRankings

Student t-distribution. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Stories/student.html

WHO life expectancy. (2013, September 19). Retrieved from www.who.int/gho/mortality_bur...n_trends/en/in dex.html

This page titled 7.3: One-Sample Test for the Mean is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

8: Estimation

In hypothesis tests, the purpose was to make a decision about a parameter, in terms of it being greater than, less than, or not equal to a value. But what if you want to actually know what the parameter is. You need to do estimation. There are two types of estimation – point estimator and confidence interval.

8.1: Basics of Confidence Intervals

8.2: One-Sample Interval for the Proportion

8.3: One-Sample Interval for the Mean

This page titled 8: Estimation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

8.1: Basics of Confidence Intervals

A point estimator is just the statistic that you have calculated previously. As an example, when you wanted to estimate the population mean, μ, the point estimator is the sample mean, $\bar{x}$. To estimate the population proportion, p, you use the sample proportion, $\hat{p}$. In general, if you want to estimate any population parameter, we will call it θ, you use the sample statistic, $\hat{\theta}$.

Point estimators are really easy to find, but they have some drawbacks. First, if you have a large sample size, then the estimate is better. But with a point estimator, you don't know what the sample size is. Also, you don't know how accurate the estimate is. Both of these problems are solved with a confidence interval.

Definition 8.1.1

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: $\hat{\theta} \pm E$, where $\hat{\theta}$ is the point estimator and 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. As an example, if you have a 95% confidence interval of 0.65 < p < 0.73, then you would say, "there is a 95% chance that the interval 0.65 to 0.73 contains the true population proportion." This means that if you have 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. The reason that this interpretation is wrong is that the true value is fixed out there somewhere. You are trying to capture it with this interval. So this is the chance is that your interval captures it, and not that the true value falls in the interval.

There is also a real world interpretation that depends on the situation. It is where you are telling people what numbers you found the parameter to lie between. So your real world is where you tell what values your parameter is between. There is no probability attached to this statement. That probability is in the statistical interpretation.

The common probabilities used for confidence intervals are 90%, 95%, and 99%. These are known as the confidence level. 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 area between the two tails. As an example, for a two-tailed test (H_A 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 - 2 α.

Example 8.1.1 stating the statistical and real world interpretations for a confidence interval

a. Suppose you have a 95% 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 $0.35 < p < 0.41$. State the statistical and real world interpretations of this statement

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 $0.35 < p < 0.41$ contains the proportion of Americans who have tried marijuana as of 2013. Real World Interpretation: The proportion of Americans who have tried marijuana as of 2013 is between 0.35 and 0.41.

One last thing to know 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 you are really good at that and that you have a 90% chance of hitting the bull's eye. Now the bull's eye is very small. Since you hit the bull's eye approximately 90% of the time, then you probably hit

inside the next ring out 95% of the time. You have a better chance of doing this, but the circle is bigger. You probably have a 99% chance of hitting the target, but that is a much bigger circle to hit. You can see, 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* 8.1.1.

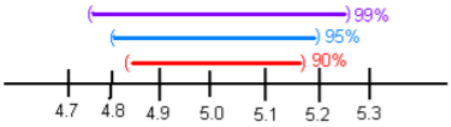


Figure 8.1.1: Affect of Confidence Level on Width

The higher level of confidence makes a wider interval. There's a trade off between width and confidence level. You can be really confident about your answer but your answer will not be very 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* 8.1.2 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 pretty close to the true value.

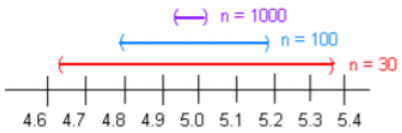


Figure 8.1.2: Affect of Sample Size on Width

Now you know everything you need to know about confidence intervals except for the actual formula. The formula depends on which parameter you are trying to estimate. With different situations you will be given the confidence interval for that parameter.

Homework

Exercise 8.1.1

1. Suppose you compute a confidence interval with a sample size of 25. What will happen to the confidence interval if the sample size increases to 50?
2. Suppose you compute a 95% confidence interval. What will happen to the confidence interval if you increase the confidence level to 99%?
3. Suppose you compute a 95% confidence interval. What will happen to the confidence interval if you decrease the confidence level to 90%?
4. Suppose you compute a confidence interval with a sample size of 100. What will happen to the confidence interval if the sample size decreases to 80?
5. A 95% confidence interval is 6353 km $< \mu <$ 6384 km, where μ is the mean diameter of the Earth. State the statistical interpretation.
6. A 95% confidence interval is 6353 km $< \mu <$ 6384 km, where μ is the mean diameter of the Earth. State the real world interpretation.
7. In 2013, Gallup conducted a poll and found a 95% confidence interval of $0.52 < p < 0.60$, where p is the proportion of Americans who believe it is the government's responsibility for health care. Give the real world interpretation.
8. In 2013, Gallup conducted a poll and found a 95% confidence interval of $0.52 < p < 0.60$, where p is the proportion of Americans who believe it is the government's responsibility for health care. Give the statistical interpretation.

Answer

1. Narrower
3. Narrower
5. See solutions
7. See solutions

This page titled 8.1: Basics of Confidence Intervals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

8.2: One-Sample Interval for the Proportion

Suppose you want to estimate the population proportion, *p*. As an example you may be curious what proportion of students at your school smoke. Or you could wonder what is the proportion of accidents caused by teenage drivers who do not have a drivers' education class.

Confidence Interval for One Population Proportion (1-Prop Interval)

1. State the random variable and the parameter in words.
   x = number of successes
   p = proportion of successes
2. State and check the assumptions for confidence interval
   a. A simple random sample of size *n* is taken.
   b. The condition for the binomial distribution are satisfied
   c. To determine the sampling distribution of $\hat{p}$, you need to show that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$, where $\hat{q} = 1 - \hat{p}$. If this requirement is true, then the sampling distribution of $\hat{p}$ is well approximated by a normal curve. (In reality this is not really true, since the correct assumption deals with *p*. However, in a confidence interval you do not know *p*, so you must use $\hat{p}$. This means you just need to show that $x \geq 5$ and $n - x \geq 5$.)
3. Find the sample statistic and the confidence interval
   Sample Proportion:
   $\hat{p} = \frac{x}{n} = \frac{\#\text{ of successes}}{\#\text{ of trials}}$
   Confidence Interval:
   $\hat{p} - E < p < \hat{p} + E$
   Where
   p = population proportion
   $\hat{p}$ = sample proportion
   n = number of sample values
   E = margin of error
   $z_c =$ = critical value
   $\hat{q} = 1 - \hat{p}$
   $E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}$
4. Statistical Interpretation: In general this looks like, "there is a C% chance that $\hat{p} - E < p < \hat{p} + E$ contains the true proportion."
5. Real World Interpretation: This is where you state what interval contains the true proportion.

The critical value is a value from the normal distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the sample proportion, and the interval has a probability of containing the true proportion, then you can think of this as the statement $P(\hat{p} - E < p < \hat{p} + E) = C$. You can use the invNorm command on the TI-83/84 calculator or qnorm command on R to find the critical value. The critical values will always be the same value, so it is easier to just look at table A.1 in the appendix.

Example 8.2.1 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 percent 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 a random sample, since it was data from six years. It is the numbers for all prisoners in these six years, but the six years were not picked at random. Unless there was something special about the six years that were chosen, the sample is probably a representative sample. 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 and both are greater than or equal to 5. The sampling distribution for $\hat{p}$ is a normal distribution.

3. Sample Proportion:

$$\hat{p} = \frac{x}{n} = \frac{51}{14495} \approx 0.003518$$

Confidence Interval:

$z_c = 1.96$, since 95% confidence level

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{0.003518(1-0.003518)}{14495}} \approx 0.000964$$

$$\hat{p} - E < p < \hat{p} + E$$

$$0.003518 - 0.000964 < p < 0.003518 + 0.000964$$

$$0.002554 < p < 0.004482$$

4. There is a 95% chance that $0.002554 < p < 0.004482$ contains the proportion of Aboriginal prisoners who died.

5. The proportion of Aboriginal prisoners who died is between 0.0026 and 0.0045.

You can also do the calculations for the confidence interval with technology. The following example shows the process on the TI-83/84.

Example 8.2.2 confidence interval for the population proportion using technology

A researcher studying the effects of income levels on 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 breastfeed. In Tajikistan, considered a low-income country by the World Bank, researchers found that in a random sample of 500 new mothers that 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 you 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 woman who breastfeed in a low-income country

p = proportion of woman who breastfeed in a low-income country

2.

a. A simple random sample of 500 breastfeeding habits of woman 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.

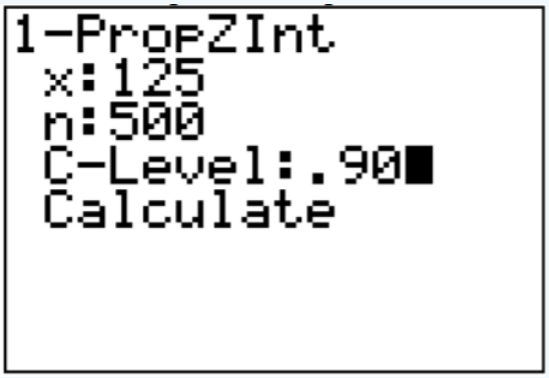


Figure 8.2.1: Setup for 1-Proportion Interval

Once you press Calculate, you will see the results as in *Figure* 8.2.2.

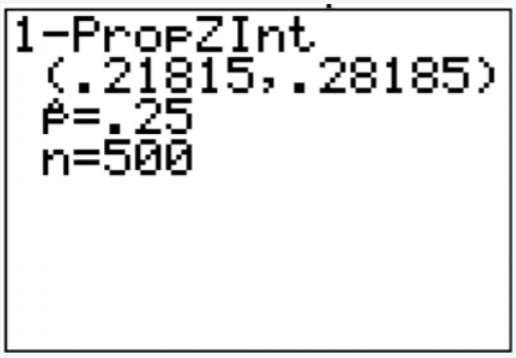


Figure 8.2.2: Results for 1-Proportion Interval

On R: the command is prop.test(x, n, conf.level = C), where C is given in decimal form. So for this example, the command is

prop.test(125, 500, conf.level = 0.90)

1-sample proportions test with continuity correction

data: 125 out of 500, null probability 0.5

X-squared = 124, df = 1, p-value < 2.2e-16

alternative hypothesis: true p is not equal to 0.5

90 percent confidence interval:

0.2185980 0.2841772

sample estimates:

p

0.25

Again, R does a continuity correction, so the answer is slightly off from the formula and the TI-83/84 calculator.

$0.219 < p < 0.284$

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.

Homework

Exercise 8.2.1

In each problem show all steps of the confidence interval. If some of the assumptions are not met, note that the results of the interval may not be correct and then continue the process of the confidence interval.

1. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they make. Looking at the type of defects, they found in a three-month time period that out of 34,641 defective lenses, 5865 were due to scratches. Find a 99% confidence interval for the proportion of defects that are from scratches.
2. In November of 1997, Australians were asked if they thought unemployment would increase. At that time 284 out of 631 said that they thought unemployment would increase ("Morgan gallup poll," 2013). Estimate the proportion of Australians in November 1997 who believed unemployment would increase using a 95% confidence interval?
3. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, Arkansas had 1,601 complaints of identity theft out of 3,482 consumer complaints ("Consumer fraud and," 2008). Calculate a 90% confidence interval for the proportion of identity theft in Arkansas.
4. According to the February 2008 Federal Trade Commission report on consumer fraud and identity theft, Alaska had 321 complaints of identity theft out of 1,432 consumer complaints ("Consumer fraud and," 2008). Calculate a 90% confidence interval for the proportion of identity theft in Alaska.
5. In 2013, the Gallup poll asked 1,039 American adults if they believe there was a conspiracy in the assassination of President Kennedy, and found that 634 believe there was a conspiracy ("Gallup news service," 2013). Estimate the proportion of American's who believe in this conspiracy using a 98% confidence interval.
6. In 2008, there were 507 children in Arizona out of 32,601 who were diagnosed with Autism Spectrum Disorder (ASD) ("Autism and developmental," 2008). Find the proportion of ASD in Arizona with a confidence level of 99%.

Answer

For all confidence intervals, just the interval using technology is given. See solution for the entire answer.

1. $0.1641 < p < 0.1745$

3. $0.4458 < p < 0.4739$

5. $0.5740 < p < 0.6452$

This page titled 8.2: One-Sample Interval for the Proportion is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

8.3: One-Sample Interval for the Mean

Suppose you want to estimate the mean height of Americans, or you want to estimate the mean salary of college graduates. A confidence interval for the mean would be the way to estimate these means.

Confidence Interval for One Population Mean (t-Interval)

1. State the random variable and the parameter in words.
   x = random variable
   μ = mean of random variable
2. State and check the assumptions for a hypothesis test
   a. A random sample of size n is taken.
   b. The population of the random variable is normally distributed, though the t-test is fairly robust to the assumption if the sample size is large. This means that if this assumption isn't met, but your sample size is quite large (over 30), then the results of the t-test are valid.
3. Find the sample statistic and confidence interval
   $\overline{x} - E < \mu < \overline{x} + E$
   where
   $E = t_c \frac{s}{\sqrt{n}}$
   $\overline{x}$ is the point estimator for μ
   t_c is the critical value where degrees of freedom: $df = n - 1$
   s is the sample standard deviation
   n is the sample size
4. Statistical Interpretation: In general this looks like, "there is a C% chance that the statement $\overline{x} - E < \mu < \overline{x} + E$ contains the true mean."
5. Real World Interpretation: This is where you state what interval contains the true mean.

The critical value is a value from the Student's t-distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the sample mean, and the interval has a probability of containing the true mean, then you can think of this as the statement $P(\overline{x} - E < \mu < \overline{x} + E) = C$. The critical values are found in table A.2 in the appendix.

How to check the assumptions of confidence interval:

In order for the confidence interval to be valid, the assumptions of the test must be true. Whenever you run a confidence interval, you must make sure the assumptions are true. You need to check them. Here is how you do this:

1. For the assumption that the sample is a random sample, describe how you took the sample. Make sure your sampling technique is random.
2. For the assumption that population is normal, remember the process of assessing normality from chapter 6.

Example 8.3.1 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 8.3.1 (this is the same data set that was used in Example 8.3.2). Find a 98% confidence interval for the IQ of a famous person.

Table 8.3.1: 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 score is normally distributed. This was shown in Example 8.3.2.

3. Sample Statistic:

$\overline{x} = 145.4$

$s \approx 29.27$

Now you need the degrees of freedom, *df* = *n* - 1 = 20 - 1 = 19 and the *C*, which is 98%. Now go to table A.2, go down the first column to 19 degrees of freedom. Then go over to the column headed with 98%. Thus $t_c = 2.539$. (See Example 8.3.2.)

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

Table 8.3.2: *Excerpt From Table A.2*

$$E = t_c \frac{s}{\sqrt{n}} = 2.539 \frac{29.27}{\sqrt{20}} \approx 16.6$$

$\overline{x} - E < \mu < \overline{x} + E$

$145.4 - 16.6 < \mu < 145.4 + 16.6$

$128.8 < \mu < 162$

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 8.3.2 confidence interval for the population mean using technology

The data in Example 8.3.3 are the life expectancies for men in European countries in 2011 ("WHO life expectancy," 2013). Find the 99% confident interval for the mean life expectancy of men in Europe.

Table 8.3.3: Life Expectancies for Men in European Countries in 2011

7365	79	67	78	69	66	78	74
71	74	79	75	77	71	78	78
68	78	78	71	81	79	80	80
62	65	69	68	79	79	79	73
79	79	72	77	67	70	63	82
72	72	77	79	80	80	67	73
73	60	65	79	66			

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 = life expectancy for a European man in 2011

μ = mean life expectancy for European men in 2011

2.

a. A random sample of 53 life expectancies of European men in 2011 was taken. The data is actually all of the life expectancies for every country that is considered part of Europe by the World Health Organization. However, the information is still sample information since it is only for one year that the data was collected. It may not be a random sample, but that is probably not an issue in this case.
b. The distribution of life expectancies of European men in 2011 is normally distributed. To see if this assumption has been met, look at the histogram, number of outliers, and the normal probability plot. (If you wish, you can look at the normal probability plot first. If it doesn't look linear, then you may want to look at the histogram and number of outliers at this point.)

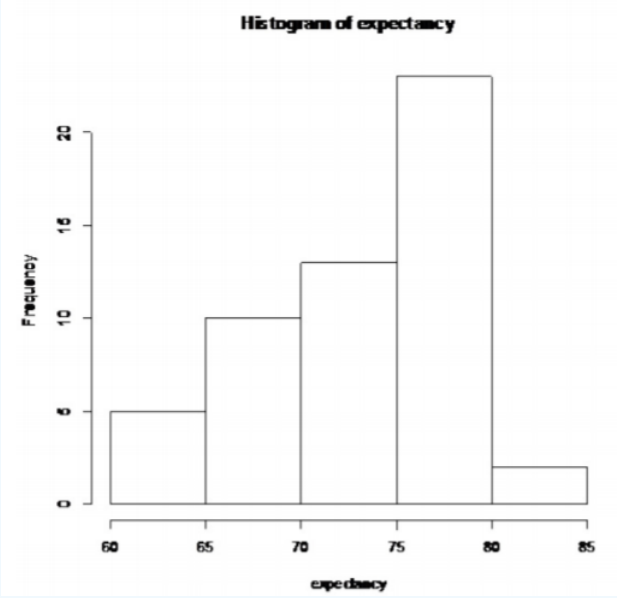


Figure 8.3.1: Histogram for Life Expectancies of European Men in 2011

Not normally distributed

Number of outliers:

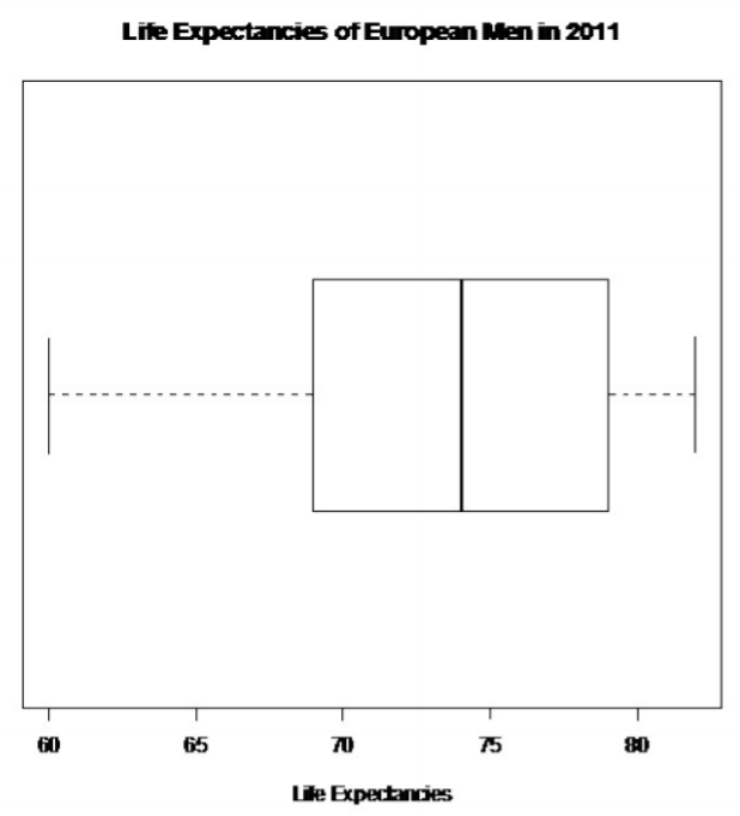


Figure 8.3.2: Modified Box Plot for Life Expectancies of European Men in 2011

$IQR = 79 - 69 = 10$

$1.5 * IQR = 15$

$Q1 - 1.5 * IQR = 69 - 15 = 54$

$Q3 + 1.5 * IQR = 79 + 15 = 94$

Outliers are numbers below 54 and above 94. There are no outliers for this data set.

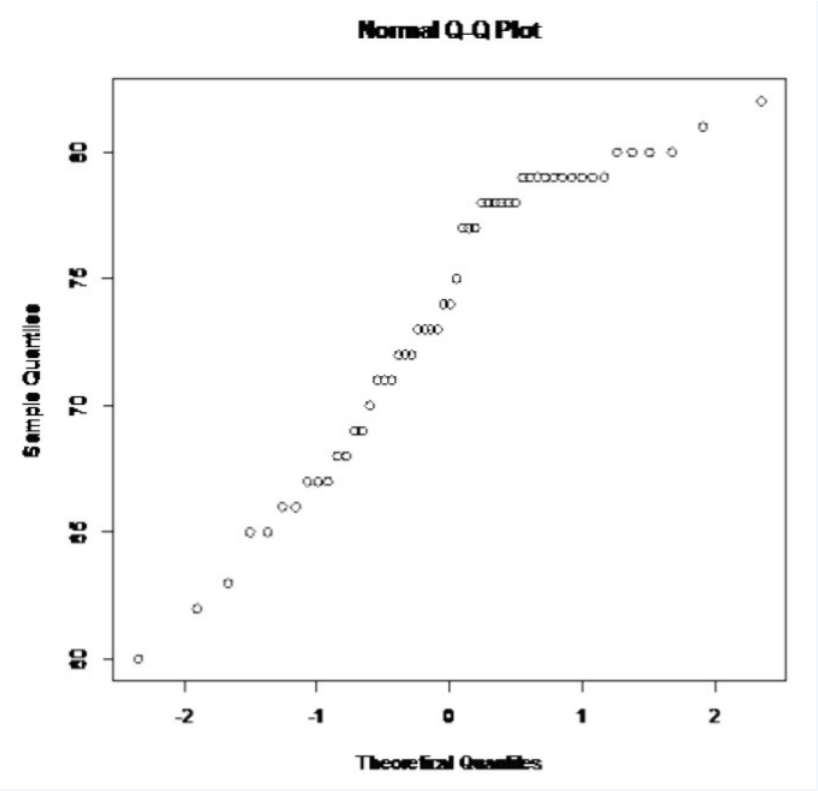


Figure 8.3.3: Normal Quantile Plot for Life Expectancies of European Men in 2011

Not linear

This population does not appear to be normally distributed. The t-test is robust for sample sizes larger than 30 so you can go ahead and calculate the interval.

3. Find the sample statistic and confidence interval

On the TI-83/84: Go into the STAT menu, and type the data into L1. Then go into STAT and over to TESTS. Choose TInterval.

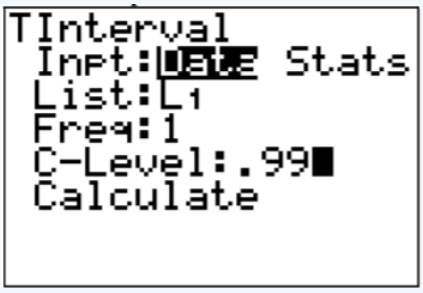


Figure 8.3.4: Setup for TInterval

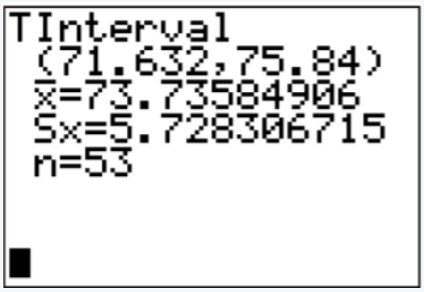


Figure 8.3.5: Results for TInterval

On R: t.test(variable, conf.level = C), where C is given in decimal form. So for this example it would be t.test(expectancy, conf.level = 0.99)

One Sample t-test

data: expectancy

t = 93.711, df = 52, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

99 percent confidence interval:

71.63204 75.83966

sample estimates:

mean of x

73.73585

71.6 years < μ 75.8 years

4. There is a 99% chance that 71.6 years < μ 75.8 years contains the mean life expectancy of European men.

5. The mean life expectancy of European men is between 71.6 and 75.8 years.

Homework

Exercise 8.3.1

In each problem show all steps of the confidence interval. If some of the assumptions are not met, note that the results of the interval may not be correct and then continue the process of the confidence interval.

1. The Kyoto Protocol was signed in 1997, and required countries to start reducing their carbon emissions. The protocol became enforceable in February 2005. Example 8.3.4 contains a random sample of CO2 emissions in 2010 ("CO2 emissions," 2013). Compute a 99% confidence interval to estimate the mean CO2 emission in 2010.

Table 8.3.4: CO2 Emissions (metric tons per capita) in 2010

1.36	1.42	5.93	5.36	0.06	9.11	7.32
7.93	6.72	0.78	1.80	0.20	2.27	0.28
5.86	3.46	1.46	0.14	2.62	0.79	7.48
0.86	7.84	2.87	2.45			

2. Many people feel that cereal is healthier alternative for children over glazed donuts. Example 8.3.5 contains the amount of sugar in a sample of cereal that is geared towards children ("Healthy breakfast story," 2013). Estimate the mean amount of sugar in children cereal using a 95% confidence level.

Table 8.3.5: Sugar Amounts (g) in Children's Cereal

10	14	12	9	13	13	13
11	12	15	9	10	11	3
6	12	15	12	12		

3. In Florida, bass fish were collected in 53 different lakes to measure the amount of mercury in the fish. The data for the average amount of mercury in each lake is in Example 8.3.6 ("Multi-disciplinary niser activity," 2013). Compute a 90% confidence interval for the mean amount of mercury in fish in Florida lakes.

Table 8.3.6: Average Mercury Levels (mg/kg) in Fish

1.23	1.33	0.04	0.44	1.20	0.27
0.48	0.19	0.83	0.81	0.81	0.5

0.49	1.16	0.05	0.15	0.19	0.77
1.08	0.98	0.63	0.56	0.41	0.73
0.34	0.59	0.34	0.84	0.50	0.34
0.28	0.34	0.87	0.56	0.17	0.18
0.19	0.04	0.49	1.10	0.16	0.10
0.48	0.21	0.86	0.52	0.65	0.27
0.94	0.40	0.43	0.25	0.27	

4. In 1882, Albert Michelson collected measurements on the speed of light ("Student t-distribution," 2013). His measurements are given in Example 8.3.7. Find the speed of light value that Michelson estimated from his data using a 95% confidence interval.

Table 8.3.7: Speed of Light Measurements in (km/sec)

299883	299816	299778	299796	299682
299711	299611	299599	300051	299781
299578	299796	299774	299820	299772
299696	299573	299748	299748	299797
299851	299809	299723		

5. Example 8.3.8 contains pulse rates after running for 1 minute, collected from females who drink alcohol ("Pulse rates before," 2013). The mean pulse rate after running for 1 minute of females who do not drink is 97 beats per minute. Do the data show that the mean pulse rate of females who do drink alcohol is higher than the mean pulse rate of females who do not drink? Test at the 5% level.

Table 8.3.8: Pulse Rates of Woman Who Use Alcohol

176	150	150	115	129	160
120	125	89	132	120	120
68	87	88	72	77	84
92	80	60	67	59	64
88	74	68			

6. The economic dynamism, which is the index of productive growth in dollars for countries that are designated by the World Bank as middle-income are in Example 8.3.9 ("SOCR data 2008," 2013). Countries that are considered high-income have a mean economic dynamism of 60.29. Do the data show that the mean economic dynamism of middle-income countries is less than the mean for high-income countries? Test at the 5% level.

Table 8.3.9: Economic Dynamism ($) of Middle Income Countries

25.8057	37.4511	51.915	43.6952	47.8506	43.7178	58.0767
41.1648	38.0793	37.7251	39.6553	42.0265	48.6159	43.8555
49.1361	61.9281	41.9543	44.9346	46.0521	48.3652	43.6252
50.9866	59.1724	39.6282	33.6074	21.6643		

7. In 1999, the average percentage of women who received prenatal care per country is 80.1%. Example 8.3.10 contains the percentage of woman receiving prenatal care in 2009 for a sample of countries ("Pregnant woman receiving," 2013). Do the data show that the average percentage of women receiving prenatal care in 2009 is higher than in 1999? Test at the 5% level.

Table 8.3.10: Percentage of Woman Receiving Prenatal Care

70.08	72.73	74.52	75.79	76.28	76.28
76.65	80.34	80.60	81.90	86.30	87.70
87.76	88.40	90.70	91.50	91.80	92.10
92.20	92.41	92.47	93.00	93.20	93.40
93.63	93.69	93.80	94.30	94.51	95.00
95.80	95.80	96.23	96.24	97.30	97.90
97.95	98.20	99.00	99.00	99.10	99.10
100.00	100.00	100.00	100.00	100.00	

8. Maintaining your balance may get harder as you grow older. A study was conducted to see how steady the elderly is on their feet. They had the subjects stand on a force platform and have them react to a noise. The force platform then measured how much they swayed forward and backward, and the data is in Example 8.3.11 ("Maintaining balance while," 2013). Do the data show that the elderly sway more than the mean forward sway of younger people, which is 18.125 mm? Test at the 5% level.

Table 8.3.11: Forward/Backward Sway (in mm) of Elderly Subjects

19	30	20	19	29	25	21	24	50

Answer

For all confidence intervals, just the interval using technology is given. See solution for the entire answer.

1. $1.7944 < \mu < 5.1152$ metric tons per capita

3. $0.44872 < \mu < 0.60562$ mg/kg

5. $87.2423 < \mu < 113.795$ beats/min

7. $88.8747\% < \mu < 93.0253\%$

Data Sources:

Australian Human Rights Commission, (1996). *Indigenous deaths in custody 1989 - 1996.* Retrieved from website: www.humanrights.gov.au/public...deaths-custody

CDC features - new data on autism spectrum disorders. (2013, November 26). Retrieved from www.cdc.gov/features/countingautism/

Center for Disease Control and Prevention, Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network. (2008). *Autism and developmental disabilities monitoring network-2012.* Retrieved from website: www.cdc.gov/ncbddd/autism/doc...nityReport.pdf

CO2 emissions. (2013, November 19). Retrieved from http://data.worldbank.org/indicator/EN.ATM.CO2E.PC

Federal Trade Commission, (2008). *Consumer fraud and identity theft complaint data: January-december 2007.* Retrieved from website: www.ftc.gov/opa/2008/02/fraud.pdf

Gallup news service. (2013, November 7-10). Retrieved from www.gallup.com/file/poll/1658...acy_131115.pdf

Healthy breakfast story. (2013, November 16). Retrieved from lib.stat.cmu.edu/DASL/Stories...Breakfast.html

Maintaining balance while concentrating. (2013, September 25). Retrieved from http://www.statsci.org/data/general/balaconc.html

Morgan Gallup poll on unemployment. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/gallup.html

Multi-disciplinary niser activity - mercury in bass. (2013, November 16). Retrieved from http://gozips.uakron.edu/~nmimoto/pa.../MercuryInBass - description.txt

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

Pulse rates before and after exercise. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/ms212.html

SOCR data 2008 world countries rankings. (2013, November 16). Retrieved from wiki.stat.ucla.edu/socr/index...ountriesRankin gs

Student t-distribution. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Stories/student.html

WHO life expectancy. (2013, September 19). Retrieved from www.who.int/gho/mortality_bur...n_trends/en/in dex.html

---

This page titled 8.3: One-Sample Interval for the Mean is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

9: Two-Sample Interference

Chapter 7 discussed methods of hypothesis testing about one-population parameters. Chapter 8 discussed methods of estimating population parameters from one sample using confidence intervals. This chapter will look at methods of confidence intervals and hypothesis testing for two populations. Since there are two populations, there are two random variables, two means or proportions, and two samples (though with paired samples you usually consider there to be one sample with pairs collected). Examples of where you would do this are:

- Testing and estimating the difference in testosterone levels of men before and after they had children (Gettler, McDade, Feranil & Kuzawa, 2011).
- Testing the claim that a diet works by looking at the weight before and after subjects are on the diet.
- Estimating the difference in proportion of those who approve of President Obama in the age group 18 to 26 year olds and the 55 and over age group.

All of these are examples of hypothesis tests or confidence intervals for two populations. The methods to conduct these hypothesis tests and confidence intervals will be explored in this method. As a reminder, all hypothesis tests are the same process. The only thing that changes is the formula that you use. Confidence intervals are also the same process, except that the formula is different.

9.1: Two Proportions
9.2: Paired Samples for Two Means
9.3: Independent Samples for Two Means
9.4: Which Analysis Should You Conduct?

This page titled 9: Two-Sample Interference is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Two Proportions

There are times you want to test a claim about two population proportions or construct a confidence interval estimate of the difference between two population proportions. As with all other hypothesis tests and confidence intervals, the process is the same though the formulas and assumptions are different.

Hypothesis Test for Two Populations Proportion (2-Prop Test)

1. State the random variables and the parameters in words.
   x_1= number of successes from group 1
   x_2 = number of successes from group 2
   p_1 = proportion of successes in group 1
   p_2 = proportion of successes in group 2
2. State the null and alternative hypotheses and the level of significance
   $H_o : p_1 = p_2$ or $H_o : p_1 - p_2 = 0$
   $H_A : p_1 < p_2$ $H_A : p_1 - p_2 < 0$
   $H_A : p_1 > p_2$ $H_A : p_1 - p_2 > 0$
   $H_A : p_1 \neq p_2$ $H_A : p_1 - p_2 \neq 0$
   Also, state your α level here.
3. State and check the assumptions for a hypothesis test
   a. A simple random sample of size n_1 is taken from population 1, and a simple random sample of size n_2 is taken from population 2.
   b. The samples are independent.
   c. The assumptions for the binomial distribution are satisfied for both populations.
   d. To determine the sampling distribution of $\hat{p}_1$, you need to show that $n_1p_1 \geq 5$ and $n_1q_1 \geq 5$, where $q_1 = 1 - p_1$. If this requirement is true, then the sampling distribution of $\hat{p}_1$ is well approximated by a normal curve. To determine the sampling distribution of $\hat{p}_2$, you need to show that $n_2p_2 \geq 5$ and $n_2q_2 \geq 5$, where $q_2 = 1 - p_2$. If this requirement is true, then the sampling distribution of $\hat{p}_2$ is well approximated by a normal curve. However, you do not know p_1 and p_2, so you need to use $\hat{p}_1$ and instead $\hat{p}_2$. This is not perfect, but it is the best you can do. Since $n_1\hat{p}_1 = n_1\frac{x_1}{n_1} = x_1$ (and similar for the other calculations) you just need to make sure that x_1, $n_1 - x_1$, $n_2 - x_2$,and are all more than 5.
4. Find the sample statistics, test statistic, and p-value
   Sample Proportion:
   $n_1 =$ size of sample 1 $n_2 =$ size of sample 2
   $\hat{p}_1 = \frac{x_1}{n_1}$(sample 1 proportion) $\hat{p}_2 = \frac{x_2}{n_2}$ (sample 2 proportion)
   $\hat{q}_1 = 1 - \hat{p}_1$ (complement of $\hat{p}_1$) $\hat{q}_2 = 1 - \hat{p}_2$ (complement of $\hat{p}_2$)
   Pooled Sample Proportion, $\overline{p}$:
   $\overline{p} = \frac{x_1 + x_2}{n_1 + n_2}$
   $\overline{q} = 1 - \overline{p}$
   Test Statistic:
   $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{pq}}{n_1} + \frac{\overline{pq}}{n_2}}}$$
   Usually $p_1 - p_2 = 0$, since $H_o : p_1 = p_2$
   p-value: On TI-83/84: use normalcdf(lower limit, upper limit, 0, 1)

> **Note**
>
> If $H_A : p_1 < p_2$ then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : p_1 > p_2$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : p_1 \neq p_2$, then find the p-value for $H_A : p_1 < p_2$, and multiply by 2.

On R: use pnorm(z, 0, 1)

Note

If $H_A : p_1 < p_2$, then use pnorm(z, 0, 1). If $H_A : p_1 > p_2$, then use 1 - pnorm(z, 0, 1). If $H_A : p_1 \neq p_2$, then find the p-value for $H_A : p_1 < p_2$, and multiply by 2.

5. Conclusion This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. Interpretation This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Confidence Interval for the Difference Between Two Population Proportion (2-Prop Interval)

The confidence interval for the difference in proportions has the same random variables and proportions and the same assumptions as the hypothesis test for two proportions. If you have already completed the hypothesis test, then you do not need to state them again. If you haven't completed the hypothesis test, then state the random variables and proportions and state and check the assumptions before completing the confidence interval step

1. Find the sample statistics and the confidence interval
   Sample Proportion:
   $n_1 =$ size of sample 1 $\quad n_2 =$ size of sample 2
   $\hat{p}_1 = \frac{x_1}{n_1}$(sample 1 proportion) $\quad \hat{p}_2 = \frac{x_2}{n_2}$ (sample 2 proportion)
   $\hat{q}_1 = 1 - \hat{p}_1$ (complement of $\hat{p}_1$) $\quad \hat{q}_2 = 1 - \hat{p}_2$ (complement of $\hat{p}_2$)
   Confidence Interval:
   The confidence interval estimate of the difference $p_1 - p_2$ is
   $(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$
   where the margin of error E is given by $E = z_c\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
   z_c = critical value
2. Statistical Interpretation: In general this looks like, "there is a C% chance that $(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$ contains the true difference in proportions."
3. Real World Interpretation: This is where you state how much more (or less) the first proportion is from the second proportion.

The critical value is a value from the normal distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the sample proportion, and the interval has a probability of being true, then you can think of this as the statement $P\left((\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E\right) = C$. So you can use the invNorm command on the TI-83/84 calculator or qnorm on R to find the critical value. These are always the same value, so it is easier to just look at the table A.1 in the Appendix.

Example 9.1.1 hypothesis test for two population proportions

Do husbands cheat on their wives more than wives cheat on their husbands ("Statistics brain," 2013)? Suppose you take a group of 1000 randomly selected husbands and find that 231 had cheated on their wives. Suppose in a group of 1200 randomly selected wives, 176 cheated on their husbands. Do the data show that the proportion of husbands who cheat on their wives are more than the proportion of wives who cheat on their husbands. Test at the 5% level.

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for a hypothesis test.
4. Find the sample statistics, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = number of husbands who cheat on his wife

x_2 = number of wives who cheat on her husband

p_1 = proportion of husbands who cheat on his wife

p_2 = proportion of wives who cheat on her husband

2. $H_o : p_1 = p_2$ or $H_o : p_1 - p_2 = 0$
   $H_A : p_1 > p_2$ $H_A : p_1 - p_2 > 0$
   $a = 0.05$

3.

a. A simple random sample of 1000 responses about cheating from husbands is taken. This was stated in the problem. A simple random sample of 1200 responses about cheating from wives is taken. This was stated in the problem.
b. The samples are independent. This is true since the samples involved different genders.
c. The properties of the binomial distribution are satisfied in both populations. This is true since there are only two responses, there are a fixed number of trials, the probability of a success is the same, and the trials are independent.
d. The sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ can be approximated with a normal distribution.
   $x_1 = 231, n_1 - x_1 = 1000 - 231 = 769, x_2 = 176$, and
   $n_2 - x_2 = 1200 - 176 = 1024$ are all greater than or equal to 5. So both sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ can be approximated with a normal distribution.

4. Sample Proportion:

$$n_1 = 1000 \qquad n_2 = 1200$$
$$\hat{p}_1 = \frac{231}{1000} = 0.231 \qquad \hat{p}_2 = \frac{176}{1200} \approx 0.1467$$
$$\hat{q}_1 = 1 - \frac{231}{1000} = \frac{769}{1000} = 0.769 \qquad \hat{q}_2 = 1 - \frac{176}{1200} = \frac{1024}{1200} \approx 0.8533$$

Pooled Sample Proportion, $\overline{p}$:

$$\overline{p} = \frac{231 + 176}{1000 + 1200} = \frac{407}{2200} = 0.185$$
$$\overline{q} = 1 - \frac{407}{2200} = \frac{1793}{2200} = 0.815$$

Test Statistic:

$$z = \frac{(0.231 - 0.1467) - 0}{\sqrt{\frac{0.185 * 0.815}{1000} + \frac{0.185 * 0.815}{1200}}}$$

$= 5.0704$

p-value:

On TI-83/84: normalcdf $(5.0704, 1E99, 0, 1) = 1.988 \times 10^{-7}$

On R: $1 -$ pnorm $(5.0704, 0, 1) = 1.988 \times 10^{-7}$

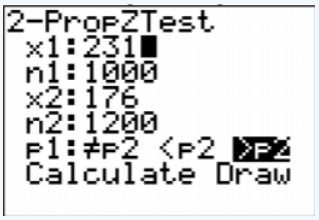


Figure 9.1.1: Setup for 2-PropZTest on TI-83/84 Calculator

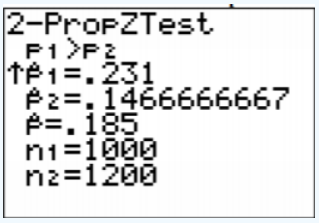


Figure 9.1.2: Results for 2-PropZTest on TI-83/84 Calculator

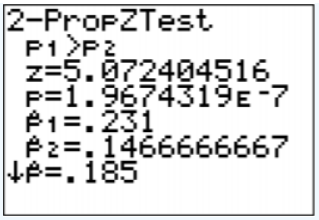


Figure 9.1.3: Results for 2-PropZTest on TI-83/84: Calculator

On R: prop.test$(c(x_1, x_2), c(n_1, n_2)$, alternative = "less" or "greater". For this example, prop.test(c(231,176), c(1000, 1200), alternative="greater")

2-sample test for equality of proportions with continuity correction

data: c(231, 176) out of c(1000, 1200)

X-squared = 25.173, df = 1, p-value = 2.621e-07

alternative hypothesis: greater

95 percent confidence interval:

0.05579805 1.00000000

sample estimates:

prop 1 prop 2

0.2310000 0.1466667

Note

The answer from R is the p-value. It is different from the formula or the TI-83/84 calculator due to a continuity correction that R does.

5. Conclusion

Reject H_o, since the p-value is less than 5%.

6. Interpretation This is enough evidence to show that the proportion of husbands having affairs is more than the proportion of wives having affairs.

Example 9.1.2 confidence interval for two population properties

Do more husbands cheat on their wives more than wives cheat on the husbands ("Statistics brain," 2013)? Suppose you take a group of 1000 randomly selected husbands and find that 231 had cheated on their wives. Suppose in a group of 1200 randomly selected wives, 176 cheated on their husbands. Estimate the difference in the proportion of husbands and wives who cheat on their spouses using a 95% confidence level.

1. State the random variables and the parameters in words.
2. State and check the assumptions for the confidence interval.
3. Find the sample statistics and the confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.1.1, but are reproduced here for reference.

x_1 = number of husbands who cheat on his wife

x_2 = number of wives who cheat on her husband

p_1 = proportion of husbands who cheat on his wife

p_2 = proportion of wives who cheat on her husband

2. The assumptions were stated and checked in Example 9.1.1.

3. Sample Proportion:

$n_1 = 1000$ $n_2 = 1200$

$\hat{p}_1 = \frac{231}{1000} = 0.231$ $\hat{p}_2 = \frac{176}{1200} \approx 0.1467$

$\hat{q}_1 = 1 - \frac{231}{1000} = \frac{769}{1000} = 0.769$ $\hat{q}_2 = 1 - \frac{176}{1200} = \frac{1024}{1200} \approx 0.8533$

Confidence Interval:

$z_c = 1.96$

$E = 1.96\sqrt{\frac{0.231 * 0.769}{1000} + \frac{0.1467 * 0.8533}{1200}} = 0.033$

The confidence interval estimate of the difference $p_1 - p_2$ is

$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$

$(0.231 - 0.1467) - 0.033 < p_1 - p_2 < (0.231 - 0.1467) + 0.033$

$0.0513 < p_1 - p_2 < 0.1173$

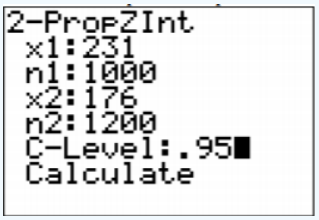


Figure 9.1.4: Setup for 2-PropZInt on TI-83/84 Calculator

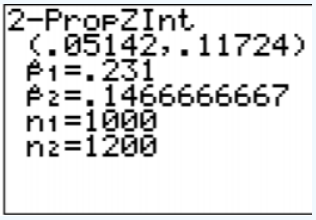


Figure 9.1.5: Results for 2-PropZInt on TI-83/84 Calculator

On R: prop.test$(c(x_1, x_2), c(n_1, n_2),$ conf.level $= C)$, where C is in decimal form. For this example, prop.test(c(231,176), c(1000, 1200), conf.level=0.95)

2-sample test for equality of proportions with continuity correction

data: c(231, 176) out of c(1000, 1200)

X-squared = 25.173, df = 1, p-value = 5.241e-07

alternative hypothesis: two.sided

95 percent confidence interval:

0.05050705 0.11815962

sample estimates:

prop 1 prop 2

0.2310000 0.1466667

Note

The answer from R is the confidence interval. It is different from the formula or the TI-83/84 calculator due to a continuity correction that R does.

4. Statistical Interpretation: There is a 95% chance that $0.0505 < p_1 - p_2 < 0.1182$ contains the true difference in proportions.

5. Real World Interpretation: The proportion of husbands who cheat is anywhere from 5.05% to 11.82% higher than the proportion of wives who cheat.

Homework

Exercise 9.1.1

In each problem show all steps of the hypothesis test or confidence interval. If some of the assumptions are not met, note that the results of the test or interval may not be correct and then continue the process of the hypothesis test or confidence interval.

1. Many high school students take the AP tests in different subject areas. In 2007, of the 144,796 students who took the biology exam 84,199 of them were female. In that same year, of the 211,693 students who took the calculus AB exam 102,598 of them were female ("AP exam scores," 2013). Is there enough evidence to show that the proportion of female students taking the biology exam is higher than the proportion of female students taking the calculus AB exam? Test at the 5% level.
2. Many high school students take the AP tests in different subject areas. In 2007, of the 144,796 students who took the biology exam 84,199 of them were female. In that same year, of the 211,693 students who took the calculus AB exam 102,598 of them were female ("AP exam scores," 2013). Estimate the difference in the proportion of female students taking the biology exam and female students taking the calculus AB exam using a 90% confidence level.
3. Many high school students take the AP tests in different subject areas. In 2007, of the 211,693 students who took the calculus AB exam 102,598 of them were female and 109,095 of them were male ("AP exam scores," 2013). Is there enough evidence to show that the proportion of female students taking the calculus AB exam is different from the proportion of male students taking the calculus AB exam? Test at the 5% level.
4. Many high school students take the AP tests in different subject areas. In 2007, of the 211,693 students who took the calculus AB exam 102,598 of them were female and 109,095 of them were male ("AP exam scores," 2013). Estimate using a 90% level the difference in proportion of female students taking the calculus AB exam versus male students taking the calculus AB exam.
5. Are there more children diagnosed with Autism Spectrum Disorder (ASD) in states that have larger urban areas over states that are mostly rural? In the state of Pennsylvania, a fairly urban state, there are 245 eight year olds diagnosed with ASD out of 18,440 eight year olds evaluated. In the state of Utah, a fairly rural state, there are 45 eight year olds diagnosed with ASD out of 2,123 eight year olds evaluated ("Autism and developmental," 2008). Is there enough evidence to show that the proportion of children diagnosed with ASD in Pennsylvania is more than the proportion in Utah? Test at the 1% level.
6. Are there more children diagnosed with Autism Spectrum Disorder (ASD) in states that have larger urban areas over states that are mostly rural? In the state of Pennsylvania, a fairly urban state, there are 245 eight year olds diagnosed with ASD out of 18,440 eight year olds evaluated. In the state of Utah, a fairly rural state, there are 45 eight year olds diagnosed with ASD out of 2,123 eight year olds evaluated ("Autism and developmental," 2008). Estimate the difference in proportion of children diagnosed with ASD between Pennsylvania and Utah. Use a 98% confidence level.
7. A child dying from an accidental poisoning is a terrible incident. Is it more likely that a male child will get into poison than a female child? To find this out, data was collected that showed that out of 1830 children between the ages one and four who pass away from poisoning, 1031 were males and 799 were females (Flanagan, Rooney & Griffiths, 2005). Do the data show that there are more male children dying of poisoning than female children? Test at the 1% level.
8. A child dying from an accidental poisoning is a terrible incident. Is it more likely that a male child will get into poison than a female child? To find this out, data was collected that showed that out of 1830 children between the ages one and four who pass away from poisoning, 1031 were males and 799 were females (Flanagan, Rooney & Griffiths, 2005). Compute a 99% confidence interval for the difference in proportions of poisoning deaths of male and female children ages one to four.

Answer

For all hypothesis tests, just the conclusion is given. For all confidence intervals, just the interval using technology (Software R) is given. See solution for the entire answer.

1. Reject Ho
2. $0.0941 < p_1 - p_2 < 0.0996$
3. Reject Ho
4. $-0.0332 < p_1 - p_2 < -0.0282$

5. Fail to reject Ho
6. $-0.01547 < p_1 - p_2 < -0.0001$
7. Reject Ho
8. $0.0840 < p_1 - p_2 < 0.1696$

This page titled 9.1: Two Proportions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

9.2: Paired Samples for Two Means

Are two populations the same? Is the average height of men taller than the average height of women? Is the mean weight less after a diet than before?

You can compare populations by comparing their means. You take a sample from each population and compare the statistics.

Anytime you compare two populations you need to know if the samples are independent or dependent. The formulas you use are different for different types of samples.

If how you choose one sample has no effect on the way you choose the other sample, the two samples are **independent**. The way to think about it is that in independent samples, the individuals from one sample are overall different from the individuals from the other sample. This will mean that sample one has no affect on sample two. The sample values from one sample are not related or paired with values from the other sample.

If you choose the samples so that a measurement in one sample is paired with a measurement from the other sample, the samples are **dependent** or **matched** or **paired**. (Often a before and after situation.) You want to make sure the there is a meaning for pairing data values from one sample with a specific data value from the other sample. One way to think about it is that in dependent samples, the individuals from one sample are the same individuals from the other sample, though there can be other reasons to pair values. This makes the sample values from each sample paired.

Example 9.2.1 independent or dependent samples

Determine if the following are dependent or independent samples.

a. Randomly choose 5 men and 6 women and compare their heights.
b. Choose 10 men and weigh them. Give them a new wonder diet drug and later weigh them again.
c. Take 10 people and measure the strength of their dominant arm and their non-dominant arm.

Solution

a. Independent, since there is no reason that one value belongs to another. The individuals are not the same for both samples. The individuals are definitely different. A way to think about this is that the knowledge that a man is chosen in one sample does not give any information about any of the woman chosen in the other sample.
b. Dependent, since each person's before weight can be matched with their after weight. The individuals are the same for both samples. A way to think about this is that the knowledge that a person weighs 400 pounds at the beginning will tell you something about their weight after the diet drug.
c. Dependent, since you can match the two arm strengths. The individuals are the same for both samples. So the knowledge of one person's dominant arm strength will tell you something about the strength of their non-dominant arm.

To analyze data when there are matched or paired samples, called dependent samples, you conduct a paired t-test. Since the samples are matched, you can find the difference between the values of the two random variables.

Hypothesis Test for Two Sample Paired *t*-Test

1. State the random variables and the parameters in words.
   x_1 = random variable 1
   x_2 = random variable 2
   μ_1 = mean of random variable 1
   μ_2 = mean of random variable 2
2. State the null and alternative hypotheses and the level of significance The usual hypotheses would be
   $H_o : \mu_1 = \mu_2$ or $H_o : \mu_1 - \mu_2 = 0$
   $H_A : \mu_1 < \mu_2$ $H_A : \mu_1 - \mu_2 < 0$
   $H_A : \mu_1 > \mu_2$ $H_A : \mu_1 - \mu_2 > 0$
   $H_A : \mu_1 \neq \mu_2$ $H_A : \mu_1 - \mu_2 \neq 0$
   However, since you are finding the differences, then you can actually think of $\mu_1 - \mu_2 = \mu_\sigma \mu_d =$ population mean value of the differences,
   So the hypotheses become

$H_o : \mu_d = 0$
$H_1 : \mu_d < 0$
$H_A : \mu_d > 0$
$H_A : \mu_d \neq 0$
Also, state your α level here.

3. State and check the assumptions for the hypothesis test
   a. A random sample of *n* pairs is taken.
   b. The population of the difference between random variables is normally distributed. In this case the population you are interested in has to do with the differences that you find. It does not matter if each random variable is normally distributed. It is only important if the differences you find are normally distributed. Just as before, the t-test is fairly robust to the assumption if the sample size is large. This means that if this assumption isn't met, but your sample size is quite large (over 30), then the results of the t-test are valid.
4. Find the sample statistic, test statistic, and p-value
   Sample Statistic:
   Difference: $d = x_1 - x_2$for each pair
   Sample mean of the differences: $\bar{d} = \frac{\sum d}{n}$
   Standard deviation of the differences: $s_d = \frac{\sum (d - \bar{d})^2}{n - 1}$
   Number of pairs: *n*
   Test Statistic:
   $t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$
   with degrees of freedom = *df* = *n* - 1

   Note

   $\mu_d = 0$ in most cases.

   p-value:
   On TI-83/84: Use tcdf (lower limit, upper limit, *df*)

   Note

   If $H_A : \mu_d < 0$, then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : \mu_d > 0$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : \mu_d \neq 0$, then find the p-value for $H_A : \mu_d < 0$, and multiply by 2.)

   On R: Use pt (*t*, *df*)

   Note

   If $H_A : \mu_d < 0$, use pt (*t*, *df*). If $H_A : \mu_d > 0$, use 1 - pt(*t*, *df*). If $H_A : \mu_d \neq 0$, then find the p-value for $H_A : \mu_d < 0$, and multiply by 2
5. This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Confidence Interval for Difference in Means from Paired Samples (t-Interval)

The confidence interval for the difference in means has the same random variables and means and the same assumptions as the hypothesis test for two paired samples. If you have already completed the hypothesis test, then you do not need to state them again. If you haven't completed the hypothesis test, then state the random variables and means, and state and check the assumptions before completing the confidence interval step.

1. Find the sample statistic and confidence interval
   Sample Statistic:
   Difference: $d = x_1 - x_2$
   Sample mean of the differences: $\overline{d} = \frac{\sum d}{n}$
   Standard deviation of the differences: $s_d = \frac{\sum (d - \overline{d})^2}{n - 1}$
   Number of pairs: n
   Confidence Interval:
   The confidence interval estimate of the difference $\mu_d = \mu_1 - \mu_2$ is
   $\overline{d} - E < \mu_d < \overline{d} + E$
   $E = t_c \frac{s_d}{\sqrt{n}}$
   t_c is the critical value where degrees of freedom $df = n - 1$
2. Statistical Interpretation: In general this looks like, "there is a C% chance that the statement $\overline{d} - E < \mu_d < \overline{d} + E$ contains the true mean difference."
3. Real World Interpretation: This is where you state what interval contains the true mean difference.

The critical value is a value from the Student's t-distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the sample mean, and the interval has a probability of containing the true mean difference, then you can think of this as the statement $P\left(\overline{d} - E < \mu_d < \overline{d} + E\right) = C$. To find the critical value, you use table A.2 in the Appendix.

How to check the assumptions of t-test and confidence interval:

In order for the t-test or confidence interval to be valid, the assumptions of the test must be met. So whenever you run a t-test or confidence interval, you must make sure the assumptions are met. So you need to check them. Here is how you do this:

1. For the assumption that the sample is a random sample, describe how you took the samples. Make sure your sampling technique is random and that the samples were dependent.
2. For the assumption that the population of the differences is normal, remember the process of assessing normality from chapter 6.

Example 9.2.2 hypothesis test for paired samples using the formula

A researcher wants to see if a weight loss program is effective. She measures the weight of 6 randomly selected women before and after the weight loss program (see Example 9.2.1). Is there evidence that the weight loss program is effective? Test at the 5% level.

Table 9.2.1: Data of Before and After Weights

Person	1	2	3	4	5	6
Weight before	165	172	181	185	168	175
Weight after	143	151	156	161	152	154

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for the hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = weight of a woman after the weight loss program

x_2 = weight of a woman before the weight loss program

μ_1 = mean weight of a woman after the weight loss program

μ_2 = mean weight of a woman before the weight loss program

$H_o : \mu_d = 0$

2. $H_A : \mu_d < 0$
   $\alpha = 0.05$
3.

a. A random sample of 6 pairs of weights before and after was taken. This was stated in the problem, since the women were chosen randomly.
b. The population of the difference in after and before weights is normally distributed. To see if this is true, look at the histogram, number of outliers, and the normal probability plot. (If you wish, you can look at the normal probability plot first. If it doesn't look linear, then you may want to look at the histogram and number of outliers at this point.)

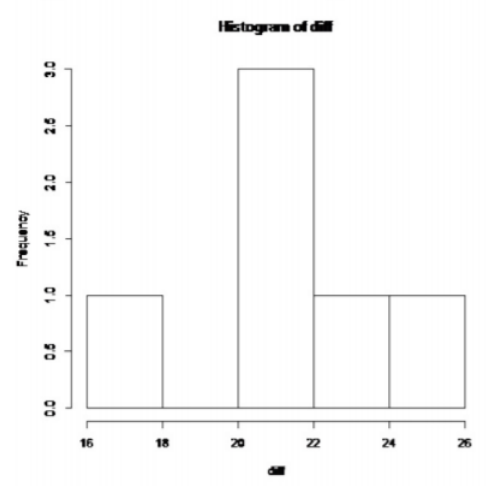


Figure 9.2.1: Histogram of Differences in Weights

This histogram looks somewhat bell shaped.

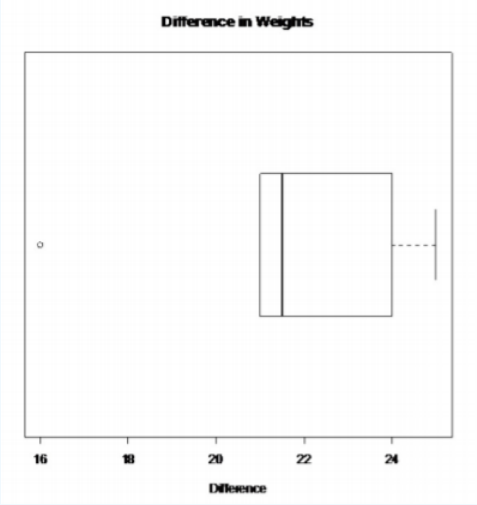


Figure 9.2.2: Modified Box Plot of Differences in Weights

There is only one outlier in the difference data set.

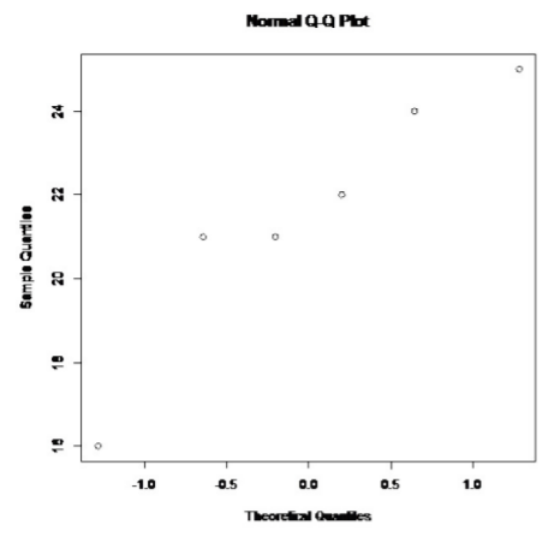


Figure 9.2.3: Normal Quantile Plot of Differences in Weights

The probability plot on the differences looks somewhat linear. So you can assume that the distribution of the difference in weights is normal.

4. Sample Statistics:

Table 9.2.2: Differences Between Before and After Weights

Person	1	2	3	4	5	6
Weight after, x_1	143	151	156	161	152	154
Weight before, x_2	165	172	181	185	168	175
$d = x_1 - x_2$	-22	-21	-25	-24	-16	-21

The mean and standard deviation are

$\overline{d} = -21.5$
$s_d = 3.15$

Test Statistic:

$$t = \frac{\overline{d} - \mu_d}{s_d/\sqrt{n}} = \frac{-21.5 - 0}{3.15/\sqrt{6}} = -16.779$$

p-value:

There are six pairs so the degrees of freedom are

$df = n - 1 = 6 - 1 = 5$

Since $H_1 : \mu_d < 0$, then p-value

Using TI-83/84: *tcdf* $(-1E99, -16.779, 5) \approx 6.87 \times 10^{-6}$

Using R: *pt* $(-16.779, 5) \approx 6.87 \times 10^{-6}$

5. Since the p-value < 0.05, reject H_o.

6. There is enough evidence to show that the weight loss program is effective.

Note

Just because the hypothesis test says the program is effective doesn't mean you should go out and use it right away. The program has statistical significance, but that doesn't mean it has practical significance. You need to see how much weight a person loses, and you need to look at how safe it is, how expensive, does it work in the long term, and other type questions. Remember to look at the practical significance in all situations. In this case, the average weight loss was 21.5 pounds, which is very practically significant. Do remember to look at the safety and expense of the drug also.

Example 9.2.3 hypothesis Test for Paired Samples Using Technology

The New Zealand Air Force purchased a batch of flight helmets. They then found out that the helmets didn't fit. In order to make sure that they order the correct size helmets, they measured the head size of recruits. To save money, they wanted to use cardboard calipers, but were not sure if they will be accurate enough. So they took 18 recruits and measured their heads with the cardboard calipers and also with metal calipers. The data in centimeters (cm) is in Example 9.2.3 ("NZ helmet size," 2013). Do the data provide enough evidence to show that there is a difference in measurements between the cardboard and metal calipers? Use a 5% level of significance.

Table 9.2.3: Data for Head Measurements

Cardboard	Metal
146	145
151	153
163	161
152	151
151	145
151	150
149	150
166	163
149	147
155	154
155	150
156	156
162	161
150	152
156	154
158	154
149	147
163	160

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for the hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = head measurement of recruit using cardboard caliper

x_2 = head measurement of recruit using metal caliper

μ_1 = mean head measurement of recruit using cardboard caliper

μ_2 = mean head measurement of recruit using metal caliper

2. $H_o : \mu_d = 0$
   $H_A : \mu_d \neq 0$
   $\alpha = 0.05$

3.

a. A random sample of 18 pairs of head measures of recruits with cardboard and metal caliper was taken. This was not stated, but probably could be safely assumed.
b. The population of the difference in head measurements between cardboard and metal calipers is normally distributed. To see if this is true, look at the histogram, number of outliers, and the normal probability plot. (If you wish, you can look at the normal probability plot first. If it doesn't look linear, then you may want to look at the histogram and number of outliers at this point.)

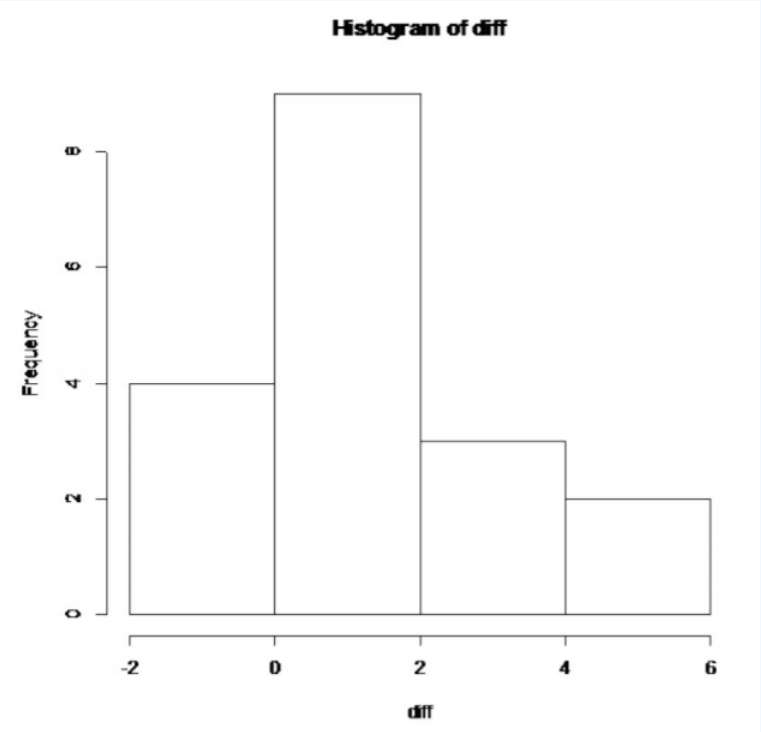


Figure 9.2.4: Histogram of Differences in Head Measurements

This histogram looks bell shaped.

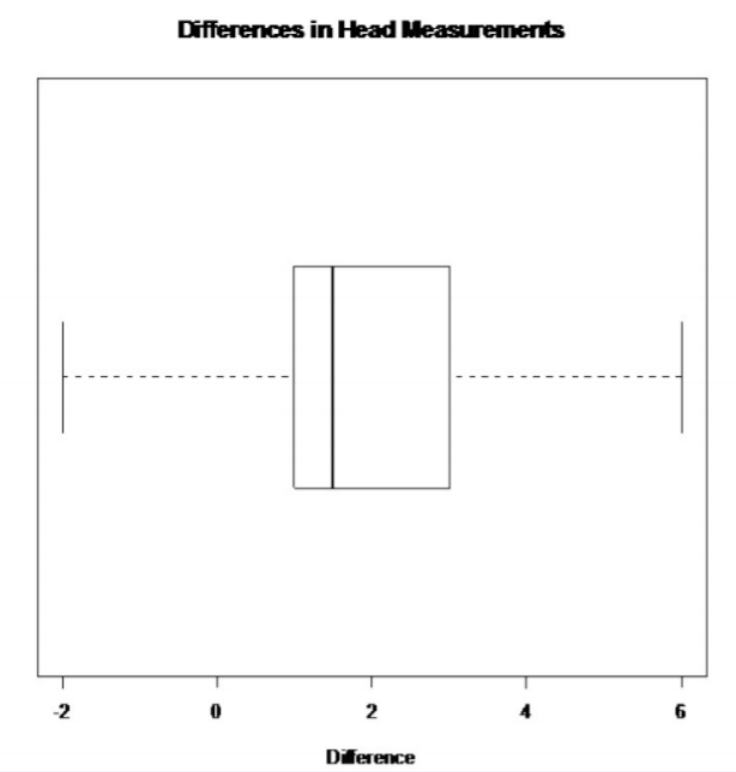


Figure 9.2.5: Modified Box Plot of Differences in Head Measurements

There are no outliers in the difference data set.

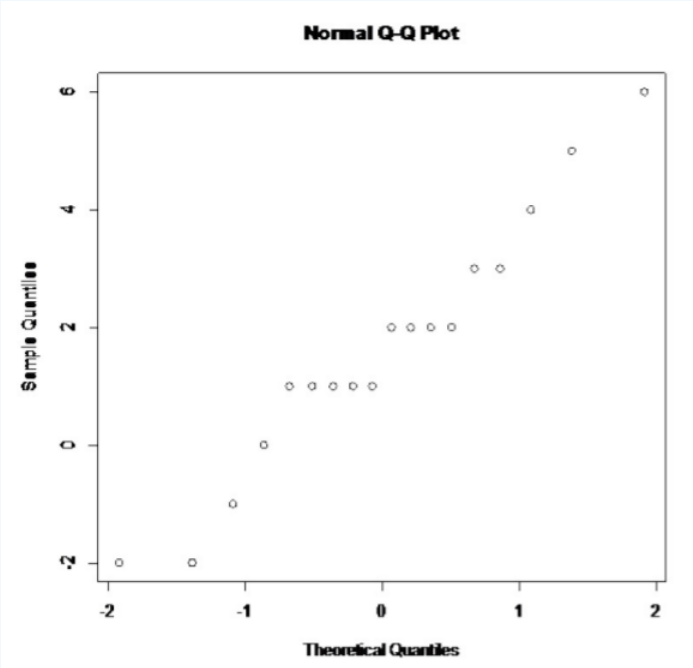


Figure 9.2.6: Normal Quantile Plot of Differences in Head Measurements

The probability plot on the differences looks somewhat linear.

So you can assume that the distribution of the difference in weights is normal.

4. Using the TI-83/84, put x_1 into L1 and x_2 into L2. Then go onto the name L3, and type *L*1-*L*2. The calculator will calculate the differences for you and put them in L3. Now go into STAT and move over to TESTS. Choose T-Test. The setup for the calculator is in *Figure* 9.2.7.

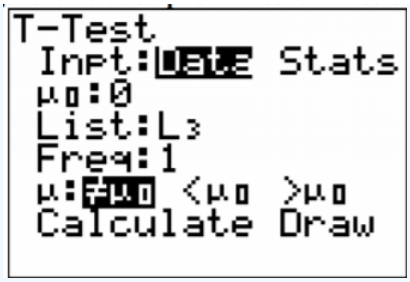


Figure 9.2.7: Setup for T-Test on TI-83/84 Calculator

Once you press ENTER on Calculate you will see the result shown in *Figure* 9.2.8.

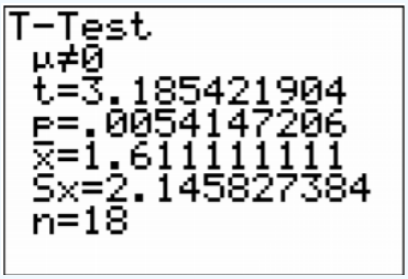


Figure 9.2.8: Results of T-Test on TI-83/84 Calculator

Using R: command is t.test(variable1, variable2, paired = TRUE, alternative = "less" or "greater"). For this example, the command would be t.test(cardboard, metal, paired = TRUE)

Paired t-test

data: cardboard and metal

t = 3.1854, df = 17, p-value = 0.005415

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.5440163 2.6782060

sample estimates:

mean of the differences

1.611111

The t = 3.185 is the test statistic. The p-value is 0.0054147206.

5. Since the p-value < 0.05, reject H_o.

6. There is enough evidence to show that the mean head measurements using the cardboard calipers are not the same as when using the metal calipers. So it looks like the New Zealand Air Force shouldn't use the cardboard calipers.

Example 9.2.4 confidence interval for paired samples using the formula

A researcher wants to estimate the mean weight loss that people experience using a new program. She measures the weight of 6 randomly selected women before and after the weight loss program (see Example 9.2.1). Find a 90% confidence interval for the mean the weight loss using the new program.

1. State the random variables and the parameters in words.
2. State and check the assumptions for the confidence interval.
3. Find the sample statistic and confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.2.2, but are reproduced here for reference.

x_1 = weight of a woman after the weight loss program

x_2 = weight of a woman before the weight loss program

μ_1 = mean weight of a woman after the weight loss program

μ_2 = mean weight of a woman before the weight loss program

2. The assumptions were stated and checked in Example 9.2.2.

3. Sample Statistics:

From Example 9.2.2

$\overline{d} = -21.5$
$s_d = 3.15$

The confidence level is 90%, so

C= 90%

There are six pairs, so the degrees of freedom are

df = *n* - 1 = 6 - 1 = 5

Now look in table A.2. Go down the first column to 5, then over to the column headed with 90%.

$t_c = 2.015$

$$E = t_c \frac{s_d}{\sqrt{n}} = 2.015 \frac{3.15}{\sqrt{6}} \approx 2.6$$

$$\overline{d} - E < \mu_d < \overline{d} + E$$

$$-21.5 - 2.6 < \mu_d < -21.5 + 2.6$$

$$-24.1 \text{ pounds } < \mu_d < -18.9 \text{ pounds}$$

4. There is a 90% chance that $-24.1 \text{ pounds } < \mu_d < -18.9 \text{ pounds}$ contains the true mean difference in weight loss.

5. The mean weight loss is between 18.9 and 24.1 pounds.

Note

The negative signs tell you that the first mean is less than the second mean, and thus a weight loss in this case.

Example 9.2.5 confidence interval for paired samples using technology

The New Zealand Air Force purchased a batch of flight helmets. They then found out that the helmets didn't fit. In order to make sure that they order the correct size helmets, they measured the head size of recruits. To save money, they wanted to use cardboard calipers, but were not sure if they will be accurate enough. So they took 18 recruits and measured their heads with the cardboard calipers and also with metal calipers. The data in centimeters (cm) is in Example 9.2.3 ("NZ helmet size," 2013). Estimate the mean difference in measurements between the cardboard and metal calipers using a 95% confidence interval.

1. State the random variables and the parameters in words.
2. State and check the assumptions for the hypothesis test.
3. Find the sample statistic and confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.2.3, but are reproduced here for reference.

x_1 = head measurement of recruit using cardboard caliper

x_2 = head measurement of recruit using metal caliper

μ_1 = mean head measurement of recruit using cardboard caliper

μ_2 = mean head measurement of recruit using metal caliper

2. The assumptions were stated and checked in Example 9.2.3.

3. Using the TI-83/84, put x_1 into L1 and x_2 into L2. Then go onto the name L3, and type *L*1 - *L*2. The calculator will now calculate the differences for you and put them in L3. Now go into STAT and move over to TESTS. Then chose TInterval. The setup for the calculator is in *Figure* 9.2.9.

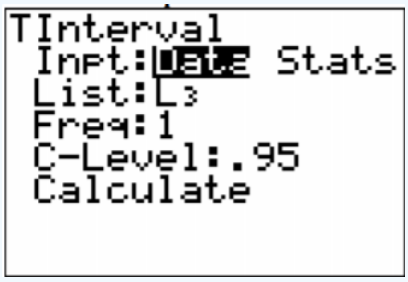


Figure 9.2.9: Setup for TInterval on TI-83/84 Calculator

Once you press ENTER on Calculate you will see the result shown in *Figure* 9.2.10.

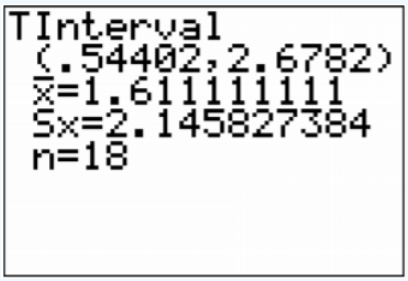


Figure 9.2.10: Results of TInterval on TI-83/84 Calculatori

Using R: the command is t.test(variable1, variable2, paired = TRUE, conf.level = C), where C is in decimal form. For this example the command would be

t.test(cardboard, metal, paired = TRUE, conf.level=0.95)

Paired t-test

data: cardboard and metal

t = 3.1854, df = 17, p-value = 0.005415

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.5440163 2.6782060

sample estimates:

mean of the differences

1.611111

So

$0.54\text{cm} < \mu_d < 2.68\text{cm}$

4. There is a 95% chance that $0.54\text{cm} < \mu_d < 2.68\text{cm}$ contains the true mean difference in head measurements between cardboard and metal calibers.

5. The mean difference in head measurements between the cardboard and metal calibers is between 0.54 and 2.68 cm. This means that the cardboard calibers measure on average the head of a recruit to be between 0.54 and 2.68 cm more in diameter

than the metal calibers. That makes it seem that the cardboard calibers are not measuring the same as the metal calibers. (The positive values on the confidence interval imply that the first mean is higher than the second mean.)

Examples 9.2.2 and *9.2.4* use the same data set, but one is conducting a hypothesis test and the other is conducting a confidence interval. Notice that the hypothesis test's conclusion was to reject H_o and say that there was a difference in the means, and the confidence interval does not contain the number 0. If the confidence interval did contain the number 0, then that would mean that the two means could be the same. Since the interval did not contain 0, then you could say that the means are different just as in the hypothesis test. This means that the hypothesis test and the confidence interval can produce the same interpretation. Do be careful though, you can run a hypothesis test with a particular significance level and a confidence interval with a confidence level that is not compatible with your significance level. This will mean that the conclusion from the confidence interval would not be the same as with a hypothesis test. So if you want to estimate the mean difference, then conduct a confidence interval. If you want to show that the means are different, then conduct a hypothesis test.

Homework

Exercise 9.2.1

In each problem show all steps of the hypothesis test or confidence interval. If some of the assumptions are not met, note that the results of the test or interval may not be correct and then continue the process of the hypothesis test or confidence interval.

1. The cholesterol level of patients who had heart attacks was measured two days after the heart attack and then again four days after the heart attack. The researchers want to see if the cholesterol level of patients who have heart attacks reduces as the time since their heart attack increases. The data is in Example 9.2.4 ("Cholesterol levels after," 2013). Do the data show that the mean cholesterol level of patients that have had a heart attack reduces as the time increases since their heart attack? Test at the 1% level.

Table 9.2.4: Cholesterol Levels in (mg/dL) of Heart Attack Patients

Patient	Cholesterol Level Day 2	Cholesterol Level Day 4
1	270	218
2	236	234
3	210	214
4	142	116
5	280	200
6	272	276
7	160	146
8	220	182
9	225	238
10	242	288
11	186	190
12	266	236
13	206	244
14	318	258
15	294	240
16	282	294
17	234	220
18	224	200

Patient	Cholesterol Level Day 2	Cholesterol Level Day 4
19	276	220
20	282	186
21	360	352
22	310	202
23	280	218
24	278	248
25	288	278
26	288	248
27	244	270
28	236	242

2. The cholesterol level of patients who had heart attacks was measured two days after the heart attack and then again four days after the heart attack. The researchers want to see if the cholesterol level of patients who have heart attacks reduces as the time since their heart attack increases. The data is in Example 9.2.4 ("Cholesterol levels after," 2013). Calculate a 98% confidence interval for the mean difference in cholesterol levels from day two to day four.
3. All Fresh Seafood is a wholesale fish company based on the east coast of the U.S. Catalina Offshore Products is a wholesale fish company based on the west coast of the U.S. Example 9.2.5 contains prices from both companies for specific fish types ("Seafood online," 2013) ("Buy sushi grade," 2013). Do the data provide enough evidence to show that a west coast fish wholesaler is more expensive than an east coast wholesaler? Test at the 5% level.

Table 9.2.5: Wholesale Prices of Fish in Dollars

Fish	All Fresh Seafood Prices	Catalina Offshore Product Prices
Cod	19.99	17.99
Tilapi	6.00	13.99
Farmed Salmon	19.99	22.99
Organic Salmon	24.99	24.99
Grouper Fillet	29.99	19.99
Tuna	28.99	31.99
Swordfish	23.99	23.99
Sea Bass	32.99	23.99
Striped Bass	29.99	14.99

4. All Fresh Seafood is a wholesale fish company based on the east coast of the U.S. Catalina Offshore Products is a wholesale fish company based on the west coast of the U.S. Example 9.2.5 contains prices from both companies for specific fish types ("Seafood online," 2013) ("Buy sushi grade," 2013). Find a 95% confidence interval for the mean difference in wholesale price between the east coast and west coast suppliers.
5. The British Department of Transportation studied to see if people avoid driving on Friday the 13th. They did a traffic count on a Friday and then again on a Friday the 13th at the same two locations ("Friday the 13th," 2013). The data for each location on the two different dates is in Example 9.2.6. Do the data show that on average fewer people drive on Friday the 13th? Test at the 5% level.

Table 9.2.6: Traffic Count

Dates	6th	13th

Dates	6th	13th
1990, July	139246	138548
1990, July	134012	132909
1991, September	137055	136018
1991, September	133732	131843
1991, December	123552	121641
1991, December	121139	118723
1992, March	128293	125532
1992, March	124631	120249
1992, November	124609	122770
1992, November	117584	117263

6. The British Department of Transportation studied to see if people avoid driving on Friday the 13th. They did a traffic count on a Friday and then again on a Friday the 13th at the same two locations ("Friday the 13th," 2013). The data for each location on the two different dates is in Example 9.2.6. Estimate the mean difference in traffic count between the 6th and the 13th using a 90% level.
7. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). The data is in Example 9.2.7. Do the data show that Reiki treatment reduces pain? Test at the 5% level.

Table 9.2.7: Pain Measures Before and After Reiki Treatment

VAS before	VAS after
6	3
2	1
2	0
9	1
3	0
3	2
4	1
5	2
2	2
3	0
5	1
2	2
3	0
5	1
1	0
6	4

VAS before	VAS after
6	1
4	4
4	1
7	6
2	1
4	3
8	8

8. To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). The data is in Example 9.2.7. Compute a 90% confidence level for the mean difference in VAS score from before and after Reiki treatment.
9. The female labor force participation rates (FLFPR) of women in randomly selected countries in 1990 and latest years of the 1990s are in Example 9.2.8 (Lim, 2002). Do the data show that the mean female labor force participation rate in 1990 is different from that in the latest years of the 1990s using a 5% level of significance?

Table 9.2.8: Female Labor Force Participation Rates

Region and country	FLFPR 25-54 1990	FLFPR 25-54 Latest years of 1990s
Iran	22.6	12.5
Morocco	41.4	34.5
Qatar	42.3	46.5
Syrian Arab Republic	25.6	19.5
United Arab Emirates	36.4	39.7
Cape Verde	46.7	50.9
Ghana	89.8	90.0
Kenya	82.1	82.6
Lesotho	51.9	68.0
South Africa	54.7	61.7
Bangladesh	73.5	60.6
Malaysia	49.0	50.2
Mongolia	84.7	71.3
Myanmar	72.1	72.3
Argentina	36.8	54
Belize	28.8	42.5
Bolivia	27.3	69.8
Brazil	51.1	63.2
Colombia	57.4	72.7
Ecuador	33.5	64

Region and country	FLFPR 25-54 1990	FLFPR 25-54 Latest years of 1990s
Nicaragua	50.1	42.5
Uruguay	59.5	71.5
Albania	77.4	78.8
Uzbekistan	79.6	82.8

10. The female labor force participation rates of women in randomly selected countries in 1990 and latest years of the 1990s are in Example 9.2.8 (Lim, 2002). Estimate the mean difference in the female labor force participation rate in 1990 to latest years of the 1990s using a 95% confidence level?
11. Example 9.2.9 contains pulse rates collected from males, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute. Do the data indicate that the pulse rate before exercise is less than after exercise? Test at the 1% level.

Table 9.2.9: Pulse Rate of Males Before and After Exercise

Pulse before	Pulse after
76	88
56	110
64	126
50	90
49	83
68	136
68	125
88	150
80	146
78	168
59	92
60	104
65	82
76	150
145	155
84	140
78	141
85	131
78	132

12. Example 9.2.9 contains pulse rates collected from males, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute. Compute a 98% confidence interval for the mean difference in pulse rates from before and after exercise.

Answer

For all hypothesis tests, just the conclusion is given. For all confidence intervals, just the interval using technology is given. See solution for the entire answer.

1. Reject Ho
2. $5.39857\text{mg/dL} < \mu_d < 41.1729\text{mg/dL}$
3. Fail to reject Ho
4. $-\$3.24216 < \mu_d < \$8.13327$
5. Reject Ho
6. $1154.09 < \mu_d < 2517.51$
7. Reject Ho
8. $1.499 < \mu_d < 3.001$
9. Fail to reject Ho
10. $-10.9096\% < \mu_d < 0.2596\%$
11. Reject Ho
12. $-62.0438 \text{ beats/min } < \mu_d < -37.1141 \text{ beats/min}$

This page titled 9.2: Paired Samples for Two Means is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

9.3: Independent Samples for Two Means

This section will look at how to analyze when two samples are collected that are independent. As with all other hypothesis tests and confidence intervals, the process is the same though the formulas and assumptions are different. The only difference with the independent t-test, as opposed to the other tests that have been done, is that there are actually two different formulas to use depending on if a particular assumption is met or not.

Hypothesis Test for Independent t-Test (2-Sample t-Test)

1. State the random variables and the parameters in words.
   x_1 = random variable 1
   x_2 = random variable 2
   μ_1 = mean of random variable 1
   μ_2= mean of random variable 2
2. State the null and alternative hypotheses and the level of significance The normal hypotheses would be

   $H_o : \mu_1 = \mu_2$ or $H_o : \mu_1 - \mu_2 = 0$
   $H_A : \mu_1 < \mu_2$ $H_A : \mu_1 - \mu_2 < 0$
   $H_A : \mu_1 > \mu_2$ $H_A : \mu_1 - \mu_2 > 0$
   $H_A : \mu_1 \neq \mu_2$ $H_A : \mu_1 - \mu_2 \neq 0$

   Also, state your α level here.
3. State and check the assumptions for the hypothesis test
   a. A random sample of size n_1 is taken from population 1. A random sample of size n_2 is taken from population 2.

   > **Note**
   >
   > The samples do not need to be the same size, but the test is more robust if they are.

   b. The two samples are independent.
   c. Population 1 is normally distributed. Population 2 is normally distributed. Just as before, the t-test is fairly robust to the assumption if the sample size is large. This means that if this assumption isn't met, but your sample sizes are quite large (over 30), then the results of the t-test are valid.
   d. The population variances are unknown and not assumed to be equal. The old assumption is that the variances are equal. However, this assumption is no longer an assumption that most statisticians use. This is because it isn't really realistic to assume that the variances are equal. So we will just assume the assumption of the variances being unknown and not assumed to be equal is true, and it will not be checked.
4. Find the sample statistic, test statistic, and p-value
   Sample Statistic:
   Calculate $\overline{x}_1, \overline{x}_2, s_1, s_2, n_1, n_2$
   Test Statistic:
   Since the assumption that $\sigma_1^2 = \sigma_2^2$ isn't being satisfied, then

   $$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   Usually $\mu_1 - \mu_2 = 0$, since $H_o : \mu_1 - \mu_2 = 0$
   Degrees of freedom: (the Welch–Satterthwaite equation)

   $$df = \frac{(A+B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}}$$

   where $A = \frac{s_1^2}{n_1}$ and $B = \frac{s_2^2}{n_2}$
   p-value:
   Using the TI-83/84: tcdf(lower limit, upper limit, *df*)

Note

If $H_A : \mu_1 - \mu_2 < 0$, then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : \mu_1 - \mu_2 > 0$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : \mu_1 - \mu_2 \neq 0$, then find the p-value for $H_A : \mu_1 - \mu_2 < 0$, and multiply by 2.

Using R: pt(t, *df*)

Note

If $H_A : \mu_1 - \mu_2 < 0$, then use pt(t, *df*). If $H_A : \mu_1 - \mu_2 > 0$, then use 1 - pt(t, *df*). If $H_A : \mu_1 - \mu_2 \neq 0$, then find the p-value for $H_A : \mu_1 - \mu_2 < 0$, and multiply by 2.

5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o
6. Interpretation This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Confidence Interval for the Difference in Means from Two Independent Samples (2 Samp T-Int)

The confidence interval for the difference in means has the same random variables and means and the same assumptions as the hypothesis test for independent samples. If you have already completed the hypothesis test, then you do not need to state them again. If you haven't completed the hypothesis test, then state the random variables and means and state and check the assumptions before completing the confidence interval step.

1. Find the sample statistic and confidence interval
   Sample Statistic:
   Calculate Confidence Interval: $\overline{x}_1, \overline{x}_2, s_1, s_2, n_1, n_2$
   The confidence interval estimate of the difference is $\mu_1 - \mu_2$
   Since the assumption that $\sigma_1^2 = \sigma_2^2$ isn't being satisfied, then
   $(\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E$
   where $E = t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
   where t_c is the critical value with degrees of freedom:
   Degrees of freedom: (the Welch–Satterthwaite equation)
   $df = \frac{(A+B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}}$
   where $A = \frac{s_1^2}{n_1}$ and $B = \frac{s_2^2}{n_2}$
2. Statistical Interpretation: In general this looks like, "there is a C% chance that $(\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E$ contains the true mean difference."
3. Real World Interpretation: This is where you state what interval contains the true difference in means, though often you state how much more (or less) the first mean is from the second mean.

The critical value is a value from the Student's t-distribution. Since a confidence interval is found by adding and subtracting a margin of error amount from the difference in sample means, and the interval has a probability of containing the true difference in means, then you can think of this as the statement $P\left((\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E\right) = C$. To find the critical value you use table A.2 in the Appendix.

How to check the assumptions of two sample t-test and confidence interval:

In order for the t-test or confidence interval to be valid, the assumptions of the test must be true. So whenever you run a t-test or confidence interval, you must make sure the assumptions are true. So you need to check them. Here is how you do this:

1. For the random sample assumption, describe how you took the two samples. Make sure your sampling technique is random for both samples.
2. For the independent assumption, describe how they are independent samples.

3. For the assumption about each population being normally distributed, remember the process of assessing normality from chapter 6. Make sure you assess each sample separately.
4. You do not need to check the equal variance assumption since it is not being assumed.

Example 9.3.1 hypothesis test for two means

The cholesterol level of patients who had heart attacks was measured two days after the heart attack. The researchers want to see if patients who have heart attacks have higher cholesterol levels over healthy people, so they also measured the cholesterol level of healthy adults who show no signs of heart disease. The data is in *Table* 9.3.1 ("Cholesterol levels after," 2013). Do the data show that people who have had heart attacks have higher cholesterol levels over patients that have not had heart attacks? Test at the 1% level.

Table 9.3.1: Cholesterol Levels in mg/dL

Cholesterol Level of Heart Attack Patients	Cholesterol Level of Healthy Individual
270	196
236	232
210	200
142	242
280	206
272	178
160	184
220	198
226	160
242	182
186	182
266	198
206	182
318	238
294	198
282	188
234	166
224	204
276	182
282	178
360	212
310	164
280	230
278	186
288	162
288	182
244	218

Cholesterol Level of Heart Attack Patients	Cholesterol Level of Healthy Individual
236	170
	200
	176

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for the hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = Cholesterol level of patients who had a heart attack

x_2 = Cholesterol level of healthy individuals

μ_1 = mean cholesterol level of patients who had a heart attack

μ_2 = mean cholesterol level of healthy individuals

2. The normal hypotheses would be

$H_o : \mu_1 = \mu_2$ or $H_o : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 > \mu_2$ $H_A : \mu_1 - \mu_2 > 0$
$\alpha = 0.01$

3.

a. A random sample of 28 cholesterol levels of patients who had a heart attack is taken. A random sample of 30 cholesterol levels of healthy individuals is taken. The problem does not state if either sample was randomly selected. So this assumption may not be valid.
b. The two samples are independent. This is because either they were dealing with patients who had heart attacks or healthy individuals.
c. Population of all cholesterol levels of patients who had a heart attack is normally distributed. Population of all cholesterol levels of healthy individuals is normally distributed.

Patients who had heart attacks:

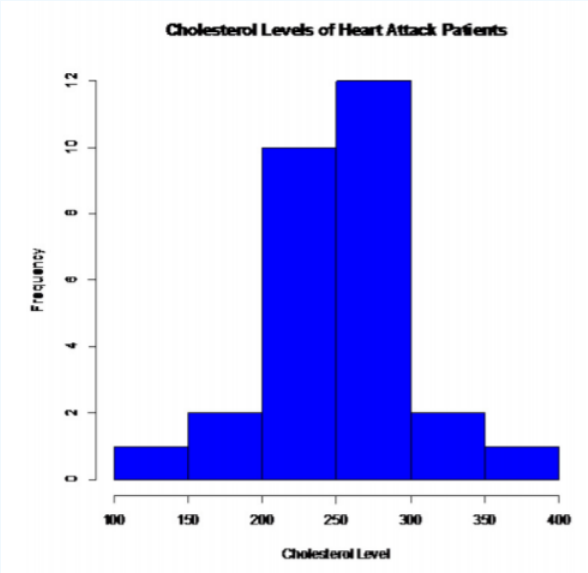


Figure 9.3.1: Histogram of Cholesterol Levels of Patients who had Heart Attacks

This looks somewhat bell shaped.

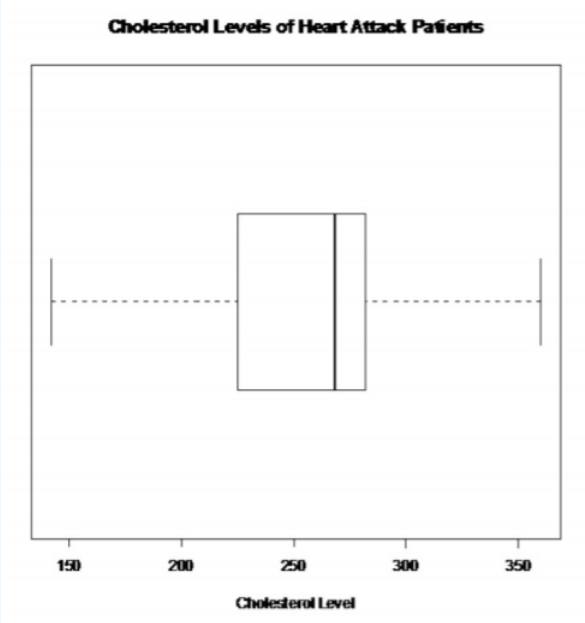


Figure 9.3.2: Modified Box Plot of Cholesterol Levels of Patients who had Heart Attacks

There are no outliers

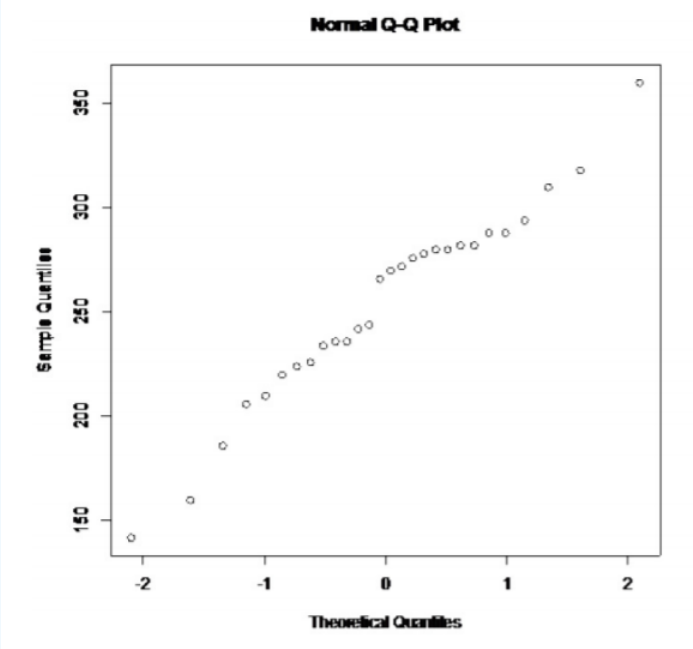


Figure 9.3.3: Normal Quantile Plot of Cholesterol Levels of Patients who had Heart Attacks

This looks somewhat linear.

So, the population of all cholesterol levels of patients who had heart attacks is probably somewhat normally distributed.

Healthy individuals:

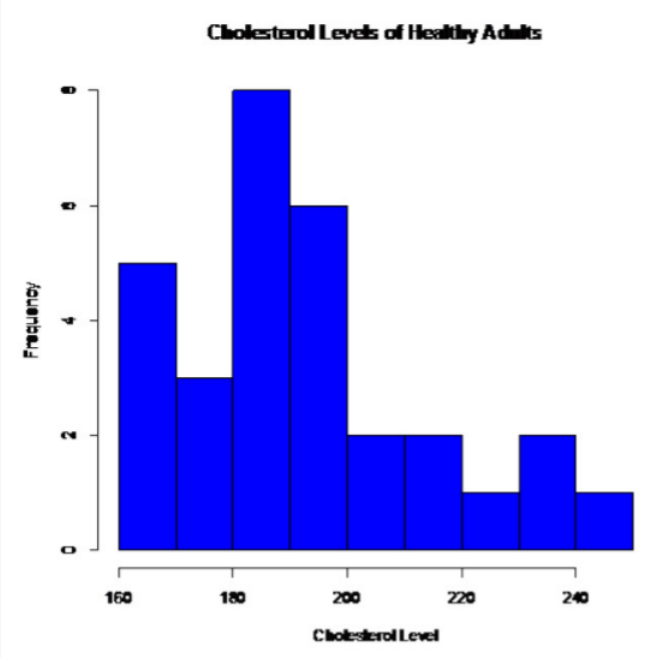


Figure 9.3.4: Histogram of Cholesterol Levels of Healthy Individuals

This does not look bell shaped.

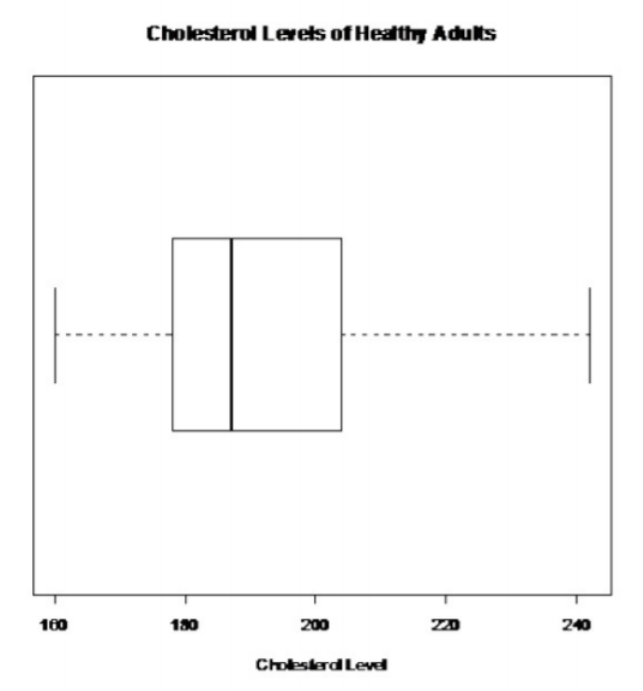


Figure 9.3.5: Modified Box Plot of Cholesterol Levels of Healthy Individuals

There are no outliers.

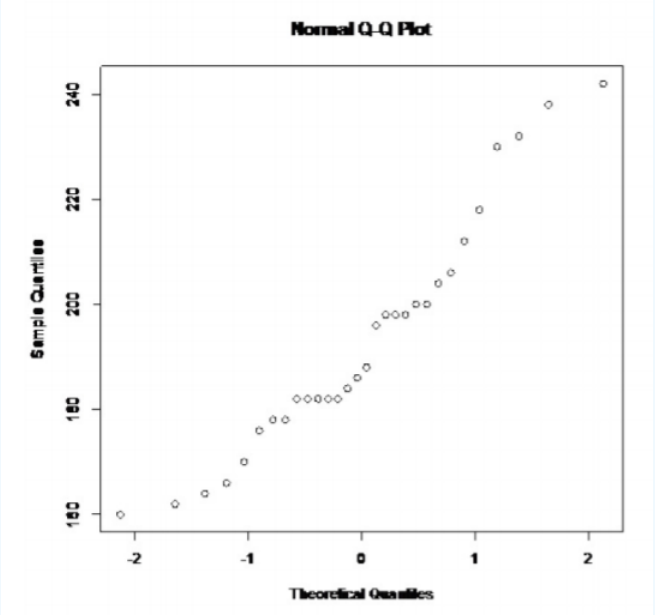


Figure 9.3.6: Normal Quantile Plot of Cholesterol Levels of Healthy Individuals

This doesn't look linear.

So, the population of all cholesterol levels of healthy individuals is probably not normally distributed.

This assumption is not valid for the second sample. Since the sample is fairly large, and the t-test is robust, it may not be an issue. However, just realize that the conclusions of the test may not be valid.

4. Sample Statistic:

$\overline{x}_1 \approx 252.32, \overline{x}_2 \approx 193.13, s_1 \approx 47.0642, s_2 \approx 22.3000, n_1 = 28, n_2 = 30$

Test Statistic:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(252.32 - 193.13) - 0}{\sqrt{\frac{47.0642^2}{28} + \frac{22.3000^2}{30}}}$$

≈ 6.051

Degrees of freedom: (the Welch-Satterthwaite equation)

$$A = \frac{s_1^2}{n_1} = \frac{47.0642^2}{28} \approx 79.1085$$

$$B = \frac{s_2^2}{n_2} = \frac{22.3000^2}{30} \approx 16.5763$$

$$df = \frac{(A+B)^2}{\frac{A^2}{n_1 - 1} + \frac{B^2}{n_2 - 1}} = \frac{(79.1085 + 16.5763)^2}{\frac{79.1085^2}{28 - 1} + \frac{16.5763^2}{30 - 1}} \approx 37.9493$$

p-value:

Using TI-83/84: $\text{tcdf}(6.051, 1E99, 37.9493) \approx 2.44 \times 10^{-7}$

Using R: $1 - \text{pt}(6.051, 37.9493) \approx 2.44 \times 10^{-7}$

Using Technology: Using the TI-83/84:



Figure 9.3.7: Setup for 2-SampTTest on TI-83/84 Calculator

Note

The Pooled question on the calculator is for whether you are assuming the variances are equal. Since this assumption is not being made, then the answer to this question is no. Pooled means that you assume the variances are equal and can pool the sample variances together.

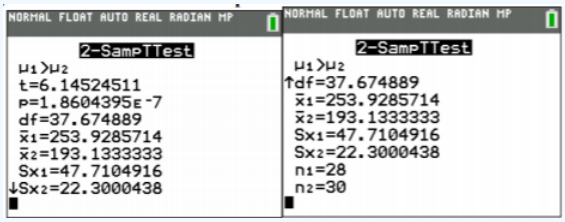


Figure 9.3.8: Results for 2-SampTTest on TI-83/84 Calculator

Using R: command in general: t.test(variable1, variable2, alternative = "less" or "greater")

For this example, the R command is:

t.test(heartattack, healthy, alternative="greater")

Welch Two Sample t-test

data: heartattack and healthy

t = 6.1452, df = 37.675, p-value = 1.86e-07

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

44.1124 Inf

sample estimates:

mean of x mean of y

253.9286 193.1333

The test statistic is t = 6.1452. The p-value is 1.86×10^{-7}

5. Reject H_o since the p-value < α.

6. This is enough evidence to show that patients who have had heart attacks have higher cholesterol level on average from healthy individuals. (Though do realize that some of assumptions are not valid, so this interpretation may be invalid.)

Example 9.3.2 confidence interval for $\mu_1 - \mu_2$

The cholesterol level of patients who had heart attacks was measured two days after the heart attack. The researchers want to see if patients who have heart attacks have higher cholesterol levels over healthy people, so they also measured the cholesterol level of healthy adults who show no signs of heart disease. The data is in Example 9.3.1 ("Cholesterol levels after," 2013). Find a 99% confidence interval for the mean difference in cholesterol levels between heart attack patients and healthy individuals.

1. State the random variables and the parameters in words.

2. State and check the assumptions for the hypothesis test.
3. Find the sample statistic and confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.3.1, but are reproduced here for reference.

x_1 = Cholesterol level of patients who had a heart attack

x_2 = Cholesterol level of healthy individuals

μ_1 = mean cholesterol level of patients who had a heart attack

μ_2 = mean cholesterol level of healthy individuals

2. The assumptions were stated and checked in Example 9.3.1.

3. Sample Statistic:

$\overline{x}_1 \approx 252.32, \overline{x_2} \approx 193.13, s_1 \approx 47.0642, s_2 \approx 22.3000, n_1 = 28, n_2 = 30$

Test Statistic:

Degrees of freedom: (the Welch–Satterthwaite equation)

$$A = \frac{s_1^2}{n_1} = \frac{47.0642^2}{28} \approx 79.1085$$

$$B = \frac{s_2^2}{n_2} = \frac{22.3000^2}{30} \approx 16.5763$$

$$df = \frac{(A+B)^2}{\frac{A^2}{n_1-1} + \frac{B^2}{n_2-1}} = \frac{(79.1085+16.5763)^2}{\frac{79.1085^2}{28-1} + \frac{16.5763^2}{30-1}} \approx 37.9493$$

Since this df is not in the table, round to the nearest whole number.

$t_c = 2.712$

$$E = t_c\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.712\sqrt{\frac{47.0642^2}{28} + \frac{22.3000^2}{30}} \approx 26.53$$

$$(\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E$$

$$(252.32 - 193.13) - 26.53 < \mu_1 - \mu_2 < (252.32 - 193.13) + 26.53$$

$$32.66\text{mg/dL} < \mu_1 - \mu_2 < 85.72\text{mg/dL}$$

Using Technology:
Using TI-83/84:

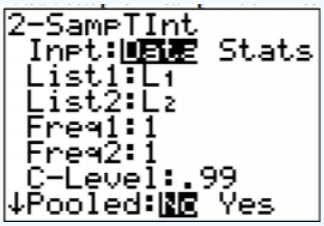


Figure 9.3.9: Setup for 2-SampTInt on TI-83/84 Calculator

Note

The Pooled question on the calculator is for whether you are assuming the variances are equal. Since this assumption is not being made, then the answer to this question is no. Pooled means that you assume the variances are equal and can pool the sample variances together.

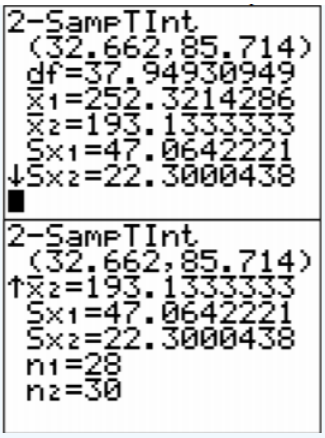


Figure 9.3.10: Results for 2-SampTInt on TI-83/84 Calculator

Using R: the commands is t.test(variable1, variable2, conf.level=C), where C is in decimal form.

For this example, the command is

t.test(heartattack, healthy, conf.level=.99)

Output:

Welch Two Sample t-test

data: heartattack and healthy

t = 6.1452, df = 37.675, p-value = 3.721e-07

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

33.95750 87.63298

sample estimates:

mean of x mean of y

253.9286 193.1333

The confidence interval is $33.96 < \mu_1 - \mu_2 < 87.63$

4. There is a 99% chance that $33.96 < \mu_1 - \mu_2 < 87.63$ contains the true difference in means.

5. The mean cholesterol level for patients who had heart attacks is anywhere from 32.66 mg/dL to 85,72 mg/dL more than the mean cholesterol level for healthy patients. (Though do realize that many of assumptions are not valid, so this interpretation may be invalid.)

If you do assume that the variances are equal, that is $\sigma_1^2 = \sigma_2^2$, then the test statistic is:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{where } s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

s_p = pooled standard deviation

The Degrees of Freedom is: $df = n_1 + n_2 - 2$

The confidence interval if you do assume that $\sigma_1^2 = \sigma_2^2$ has been met, is

$$(\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E$$

$$\text{where } E = t_c s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and $s_p = \sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{(n_1-1)+(n_2-1)}}$

Degrees of Freedom: $df = n_1 + n_2 - 2$

t_c is the critical value where C = 1 - α

To show that the variances are equal, just show that the ratio of your sample variances is not unusual (probability is greater than 0.05). In other words, make sure the following is true.

$P\left(F > s_1^2/s_2^2\right) \geq 0.05$ (or $P\left(F > s_2^2/s_1^2\right) \geq 0.05$ so that the larger variance is in the numerator). This probability is from an F-distribution. To find the probability on the TI-83/84 calculator use $\text{Fcdf}\left(s_1^2/s_2^2, 1E99, n_1-1, n_2-1\right)$. To find the probability on R, use $1 - \text{pf}\left(s_1^2/s_2^2, n_1-1, n_2-1\right)$.

Note

The F-distribution is very sensitive to the normal distribution. A better test for equal variances is Levene's test, though it is more complicated. It is best to do Levene's test when using statistical software (such as SPSS or Minitab) to perform the two-sample independent t-test.

Example 9.3.3 hypothesis test for two means

The amount of sodium in beef hotdogs was measured. In addition, the amount of sodium in poultry hotdogs was also measured ("SOCR 012708 id," 2013). The data is in Example 9.3.2. Is there enough evidence to show that beef has less sodium on average than poultry hotdogs? Use a 5% level of significance.

Table 9.3.2: Hotdog Data

Sodium in Beef Hotdogs	Sodium in Poultry Hotdogs
495	430
477	375
425	396
322	383
482	387
587	542
370	359
322	357
479	528
375	513
330	426
300	513
386	358
401	581
645	588
440	522
317	545
319	430
298	375

Sodium in Beef Hotdogs	Sodium in Poultry Hotdogs
253	396

1. State the random variables and the parameters in words.
2. State the null and alternative hypotheses and the level of significance.
3. State and check the assumptions for the hypothesis test.
4. Find the sample statistic, test statistic, and p-value.
5. Conclusion
6. Interpretation

Solution

1. x_1 = sodium level in beef hotdogs

x_2 = sodium level in poultry hotdogs

μ_1 = mean sodium level in beef hotdogs

μ_2 = mean sodium level in poultry hotdogs

2. The normal hypotheses would be

$H_o : \mu_1 = \mu_2$ or $H_o : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 < \mu_2$ $H_A : \mu_1 - \mu_2 < 0$
$\alpha = 0.05$

3.

a. A random sample of 20 sodium levels in beef hotdogs is taken. A random sample of 20 sodium levels in poultry hotdogs. The problem does not state if either sample was randomly selected. So this assumption may not be valid.
b. The two samples are independent since these are different types of hotdogs.
c. Population of all sodium levels in beef hotdogs is normally distributed. Population of all sodium levels in poultry hotdogs is normally distributed. Beef Hotdogs:

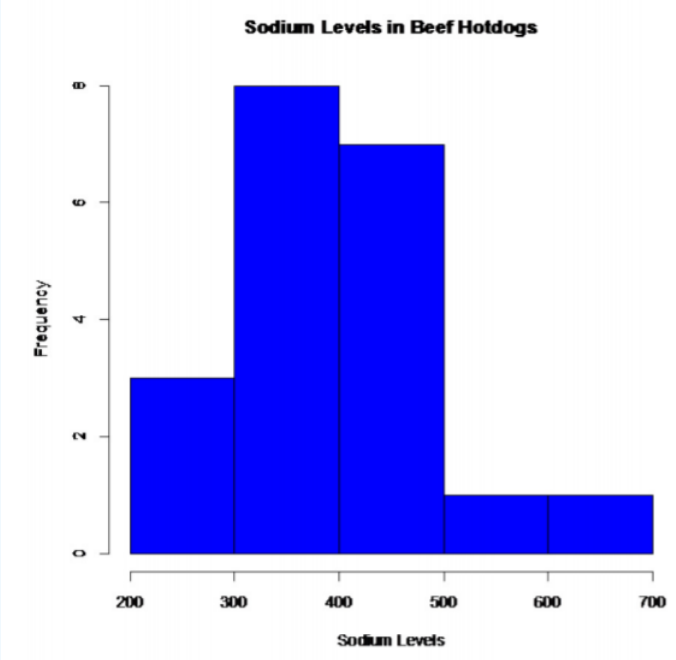


Figure 9.3.11: Histogram of Sodium Levels in Beef Hotdogs

This looks somewhat bell shaped.

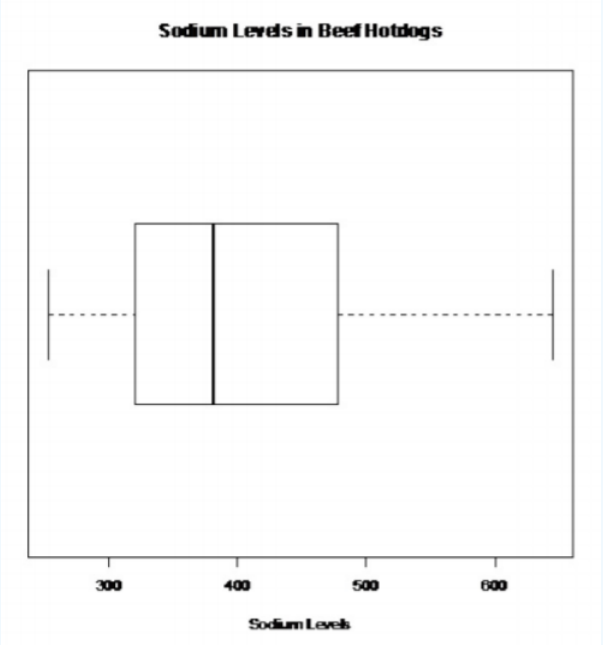


Figure 9.3.12: Modified Box Plot of Sodium Levels in Beef Hotdogs

There are no outliers.

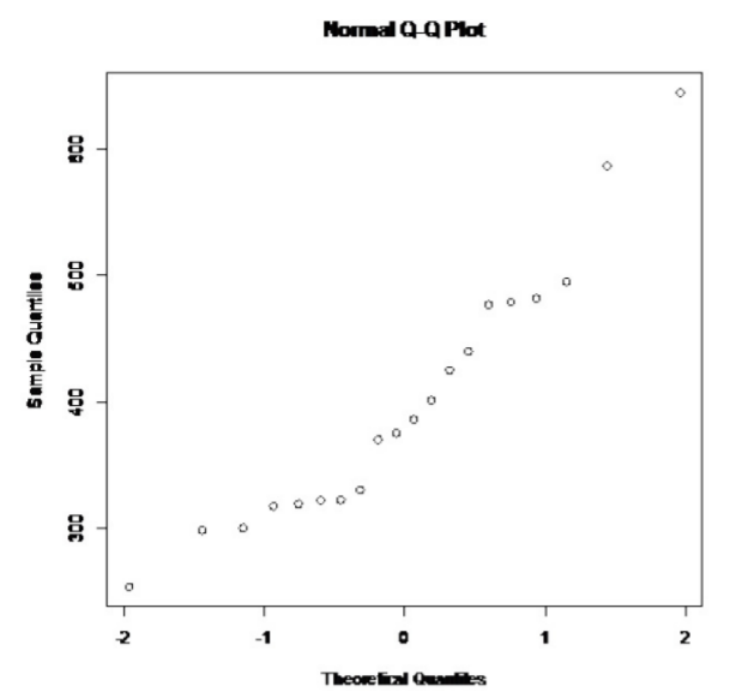


Figure 9.3.13: Normal Quantile Plot of Sodium Levels in Beef Hotdogs

This looks somewhat linear.

So, the population of all sodium levels in beef hotdogs may be normally distributed.

Poultry Hotdogs:

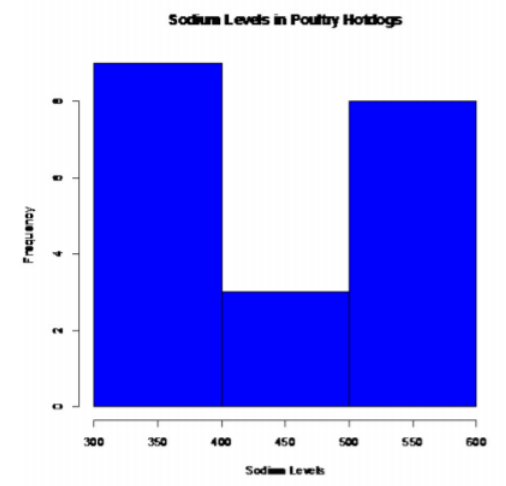


Figure 9.3.14: Histogram of Sodium Levels in Poultry Hotdogs

This does not look bell shaped.

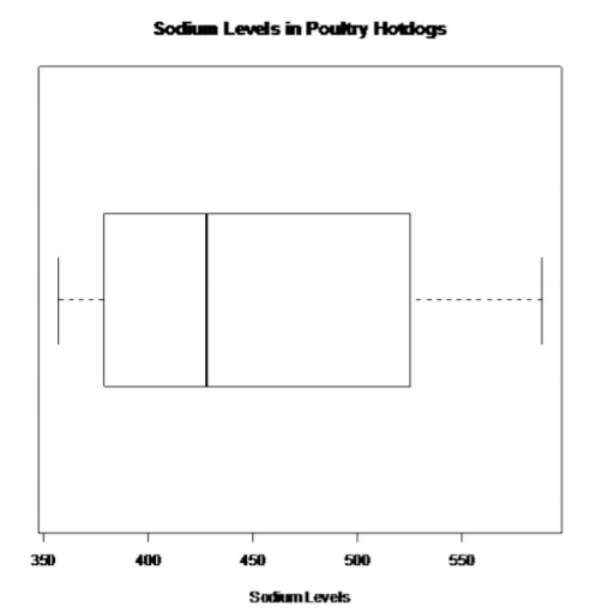


Figure 9.3.15: Modified Box Plot of Sodium Levels in Poultry Hotdogs

There are no outliers.

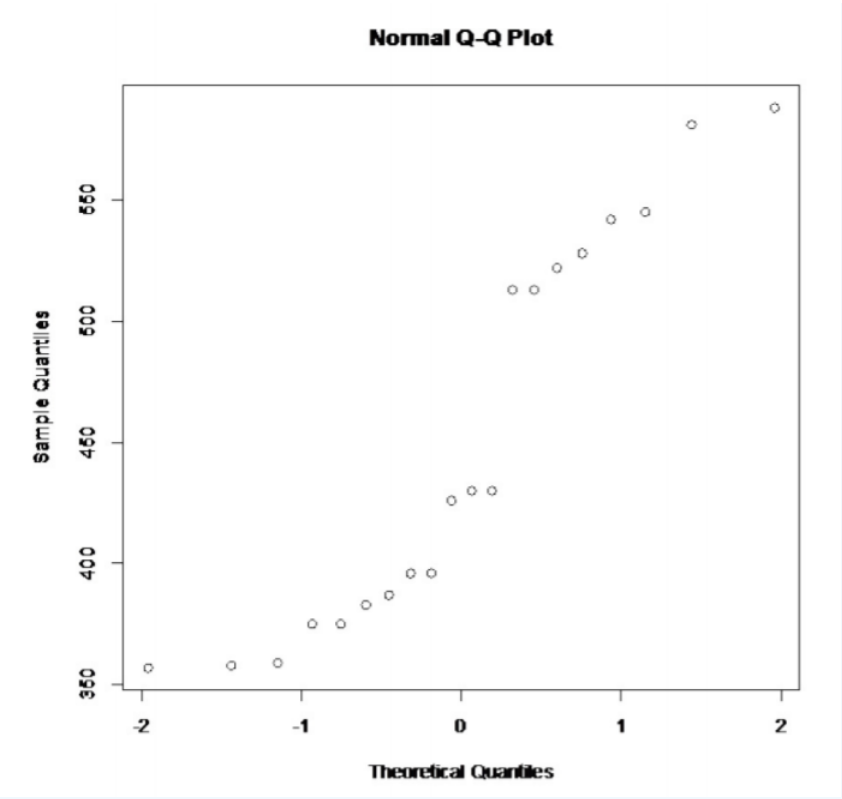


Figure 9.3.16: Normal Quantile Plot of Sodium Levels in Poultry Hotdogs

This does not look linear.

So, the population of all sodium levels in poultry hotdogs is probably not normally distributed.

This assumption is not valid. Since the samples are fairly large, and the t-test is robust, it may not be a large issue. However, just realize that the conclusions of the test may not be valid.

d. The population variances are equal, i.e. $\sigma_1^2 = \sigma_2^2$.

$s_1 \approx 102.4347$
$s_2 \approx 81.1786$
$\frac{s_1^2}{s_2^2} = \frac{102.4347^2}{81.1786^2} \approx 1.592$

Using TI-83/84: Fcdf $(1.592, 1E99, 19, 19) \approx 0.1597 \geq 0.05$

Using R: 1 - pf $(1.592, 19, 19) \approx 0.1597 \geq 0.05$

So you can say that these variances are equal.

4. Find the sample statistic, test statistic, and p-value

Sample Statistic:

$\overline{x}_1 = 401.15, \overline{x}_2 = 450.2, s_1 \approx 102.4347, s_2 \approx 81.1786, n_1 = 20, n_2 = 20$

Test Statistic:

The assumption $\sigma_1^2 = \sigma_2^2$ has been met, so

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

$$= \sqrt{\frac{102.4347^2 * 19 + 81.1786^2 * 19}{(20 - 1) + (20 - 1)}}$$

≈ 92.4198

Though you should try to do the calculations in the problem so you don't create round off error.

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{(401.15 - 450.2) - 0}{92.4198\sqrt{\frac{1}{20} + \frac{1}{20}}}$$

$$\approx -1.678$$

df = 20 + 20 - 2 = 38

p-value:

Using TI-83/84: tcdf $(-1E99, -1.678, 38) \approx 0.0508$

Using R: pt $(-1.678, 38) \approx 0.0508$

Using technology to find the t and p-value:

Using TI-83/84:



Figure 9.3.17: Setup for 2-SampTTest on TI-83/84 Calculator

Note

The Pooled question on the calculator is for whether you are using the pooled standard deviation or not. In this example, the pooled standard deviation was used since you are assuming the variances are equal. That is why the answer to the question is Yes.

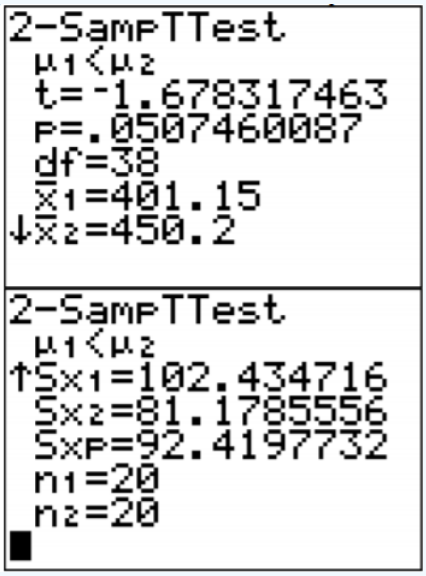


Figure 9.3.18: Results for 2-SampTTest on TI-83/84 Calculator

Using R: the command is t.test(variable1, variable2, alternative="less" or "greater")

For this example, the command is

t.test(beef, poultry, alternative="less", equalvar=TRUE)

Welch Two Sample t-test

data: beef and poultry

t = -1.6783, df = 36.115, p-value = 0.05096

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 0.2875363

sample estimates:

mean of x mean of y

401.15 450.20

The t = -1.6783 and the p-value = 0.05096.

5. Fail to reject H_o since the p-value > α.

6. This is not enough evidence to show that beef hotdogs have less sodium than poultry hotdogs. (Though do realize that many of assumptions are not valid, so this interpretation may be invalid.)

Example 9.3.4 confidence interval for $\mu_1 - \mu_2$

The amount of sodium in beef hotdogs was measured. In addition, the amount of sodium in poultry hotdogs was also measured ("SOCR 012708 id," 2013). The data is in Example 9.3.2. Find a 95% confidence interval for the mean difference in sodium levels between beef and poultry hotdogs.

1. State the random variables and the parameters in words.
2. State and check the assumptions for the hypothesis test.
3. Find the sample statistic and confidence interval.
4. Statistical Interpretation
5. Real World Interpretation

Solution

1. These were stated in Example 9.3.1, but are reproduced here for reference.

x_1 = sodium level in beef hotdogs

x_2 = sodium level in poultry hotdogs

μ_1 = mean sodium level in beef hotdogs

μ_2 = mean sodium level in poultry hotdogs

2. The assumptions were stated and checked in Example 9.3.3.

3. Sample Statistic:

$\overline{x}_1 = 401.15, \overline{x}_2 = 450.2, s_1 \approx 102.4347, s_2 \approx 81.1786, n_1 = 20, n_2 = 20$

Confidence Interval:

The confidence interval estimate of the difference $\mu_1 - \mu_2$ is

The assumption $\sigma_1^2 = \sigma_2^2$ has been met, so

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

$$= \sqrt{\frac{102.4347^2 * 19 + 81.1786^2 * 19}{(20 - 1) + (20 - 1)}}$$

≈ 92.4198

Though you should try to do the calculations in the formula for E so you don't create round off error.

$df = = n_1 + n_2 - 2 = 20 + 20 - 2 = 38$

$t_c = 2.024$

$E = t_c s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$= 2.024(92.4198)\sqrt{\frac{1}{20} + \frac{1}{20}}$

≈ 59.15

$(\overline{x}_1 - \overline{x}_2) - E < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + E$

$(401.15 - 450.2) - 59.15 < \mu_1 - \mu_2 < (401.15 - 450.2) + 59.15$

$-108.20\text{g} < \mu_1 - \mu_2 < 10.10\text{g}$

Using technology:

Using the TI-83/84:



Figure 9.3.19: Setup for 2-SampTInt on Ti-83/84 Calculator

Note

The Pooled question on the calculator is for whether you are using the pooled standard deviation or not. In this example, the pooled standard deviation was used since you are assuming the variances are equal. That is why the answer to the question is Yes.

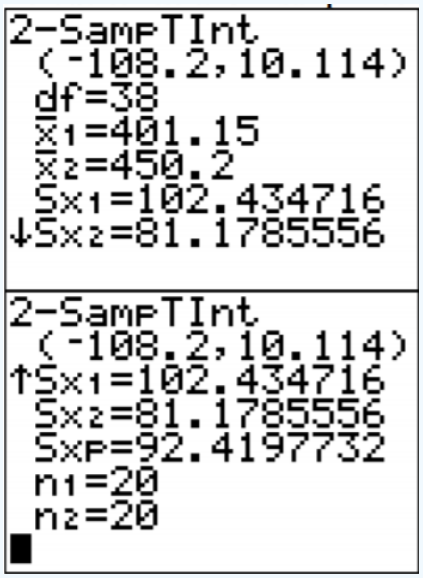


Figure 9.3.20: Results for 2-SampTInt on TI-83/84 Calculator

Using R: the command is t.test(variable1, variable2, equalvar=TRUE, conf.level=C), where C is in decimal form.

For this example, the command is

t.test(beef, poultry, conf.level=.95, equalvar=TRUE)

Welch Two Sample t-test

data: beef and poultry

t = -1.6783, df = 36.115, p-value = 0.1019

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-108.31592 10.21592

sample estimates:

mean of x mean of y

401.15 450.20

The confidence interval is $-108.32 < \mu_1 - \mu_2 < 10.22$.

4. There is a 95% chance that $-108.20\text{g} < \mu_1 - \mu_2 < 10.10\text{g}$ contains the true difference in means.

5. The mean sodium level of beef hotdogs is anywhere from 108.20 g less than the mean sodium level of poultry hotdogs to 10.10 g more. (The negative sign on the lower limit implies that the first mean is less than the second mean. The positive sign on the upper limit implies that the first mean is greater than the second mean.)

Realize that many of assumptions are not valid in this example, so the interpretation may be invalid.

Homework

Exercise 9.3.1

In each problem show all steps of the hypothesis test or confidence interval. If some of the assumptions are not met, note that the results of the test or interval may not be correct and then continue the process of the hypothesis test or confidence interval. Unless directed by your instructor, do not assume the variances are equal (except in problems 11 through 16).

1. The income of males in each state of the United States, including the District of Columbia and Puerto Rico, are given in Example 9.3.3, and the income of females is given in table #9.3.4 ("Median income of," 2013). Is there enough evidence to show that the mean income of males is more than of females? Test at the 1% level.

Table 9.3.3: Data of Income for Males

$42,951	$52,379	$42,544	$37,488	$49,281	$50,987	$60,705
$50,411	$66,760	$40,951	$43,902	$45,494	$41,528	$50,746
$45,183	$43,624	$43,993	$41,612	$46,313	$43,944	$56,708
$60,264	$50,053	$50,580	$40,202	$43,146	$41,635	$42,182
$41,803	$53,033	$60,568	$41,037	$50,388	$41,950	$44,660
$46,176	$41,420	$45,976	$47,956	$22,529	$48,842	$41,464
$40,285	$41,309	$43,160	$47,573	$44,057	$52,805	$53,046
$42,125	$46,214	$51,630				

Table 9.3.4: Data of Income for Females

$31,862	$40,550	$36,048	$30,752	$41,817	$40,236	$47,476	$40,500
$60,332	$33,823	$35,438	$37,242	$31,238	$39,150	$34,023	$33,745
$33,269	$32,684	$31,844	$34,599	$48,748	$46,185	$36,931	$40,416

$29,548	$33,865	$31,067	$33,424	$35,484	$41,021	$47,155	$32,316
$42,113	$33,459	$32,462	$35,746	$31,274	$36,027	$37,089	$22,117
$41,412	$31,330	$31,329	$33,184	$35,301	$32,843	$38,177	$40,969
$40,993	$29,688	$35,890	$34,381				

2. The income of males in each state of the United States, including the District of Columbia and Puerto Rico, are given in Example 9.3.3, and the income of females is given in Example 9.3.4 ("Median income of," 2013). Compute a 99% confidence interval for the difference in incomes between males and females in the U.S.
3. A study was conducted that measured the total brain volume (TBV) (in mm^3) of patients that had schizophrenia and patients that are considered normal. Example 9.3.5 contains the TBV of the normal patients and Example 9.3.6 contains the TBV of schizophrenia patients ("SOCR data oct2009," 2013). Is there enough evidence to show that the patients with schizophrenia have less TBV on average than a patient that is considered normal? Test at the 10% level.

Table 9.3.5: Total Brain Volume (in mm^3) of Normal Patients

1663407	1583940	1299470	1535137	1431890	1578698
1453510	1650348	1288971	1366346	1326402	1503005
1474790	1317156	1441045	1463498	1650207	1523045
1441636	1432033	1420416	1480171	1360810	1410213
1574808	1502702	1203344	1319737	1688990	1292641
1512571	1635918				

Table 9.3.6: Total Brain Volume (in mm^3) of Schizophrenia Patients

1331777	1487886	1066075	1297327	1499983	1861991
1368378	1476891	1443775	1337827	1658258	1588132
1690182	1569413	1177002	1387893	1483763	1688950
1563593	1317885	1420249	1363859	1238979	1286638
1325525	1588573	1476254	1648209	1354054	1354649
1636119					

4. A study was conducted that measured the total brain volume (TBV) (in mm^3) of patients that had schizophrenia and patients that are considered normal. Example 9.3.5 contains the TBV of the normal patients and Example 9.3.6 contains the TBV of schizophrenia patients ("SOCR data oct2009," 2013). Compute a 90% confidence interval for the difference in TBV of normal patients and patients with Schizophrenia.
5. The length of New Zealand (NZ) rivers that travel to the Pacific Ocean are given in Example 9.3.7 and the lengths of NZ rivers that travel to the Tasman Sea are given in Example 9.3.8 ("Length of NZ," 2013). Do the data provide enough evidence to show on average that the rivers that travel to the Pacific Ocean are longer than the rivers that travel to the Tasman Sea? Use a 5% level of significance.

Table 9.3.7: Lengths (in km) of NZ Rivers that Flow into the Pacific Ocean

209	48	169	138	64
97	161	95	145	90
121	80	56	64	209
64	72	288	322	

Table 9.3.8: Lengths (in km) of NZ Rivers that Flow into the Tasman Sea

76	64	68	64	37	32
32	51	56	40	64	56
80	121	177	56	80	35
72	72	108	48		

6. The length of New Zealand (NZ) rivers that travel to the Pacific Ocean are given in Example 9.3.7 and the lengths of NZ rivers that travel to the Tasman Sea are given in Example 9.3.8 ("Length of NZ," 2013). Estimate the difference in mean lengths of rivers between rivers in NZ that travel to the Pacific Ocean and ones that travel to the Tasman Sea. Use a 95% confidence level.
7. The number of cell phones per 100 residents in countries in Europe is given in Example 9.3.9 for the year 2010. The number of cell phones per 100 residents in countries of the Americas is given in Example 9.3.10 also for the year 2010 ("Population reference bureau," 2013). Is there enough evidence to show that the mean number of cell phones in countries of Europe is more than in countries of the Americas? Test at the 1% level.

Table 9.3.9: Number of Cell Phones per 100 Residents in Europe

100	76	100	130	75	84
112	84	138	133	118	134
126	188	129	93	64	128
124	122	109	121	127	152
96	63	99	95	151	147
123	95	67	67	118	125
110	115	140	115	141	77
98	102	102	112	118	118
54	23	121	126	47	

Table 9.3.10: Number of Cell Phones per 100 Residents in the America

158	117	106	159	53	50
78	66	88	92	42	3
150	72	86	113	50	58
70	109	37	32	85	101
75	69	55	115	95	73
86	157	100	119	81	113
87	105	96			

8. The number of cell phones per 100 residents in countries in Europe is given in Example 9.3.9 for the year 2010. The number of cell phones per 100 residents in countries of the Americas is given in Example 9.3.10 also for the year 2010 ("Population reference bureau," 2013). Find the 98% confidence interval for the difference in mean number of cell phones per 100 residents in Europe and the Americas.
9. A vitamin K shot is given to infants soon after birth. Nurses at Northbay Healthcare were involved in a study to see if how they handle the infants could reduce the pain the infants feel ("SOCR data nips," 2013). One of the measurements taken was how long, in seconds, the infant cried after being given the shot. A random sample was taken from the group that was given the shot using conventional methods (Example 9.3.11), and a random sample was taken from the group that was given

the shot where the mother held the infant prior to and during the shot (Example 9.3.12). Is there enough evidence to show that infants cried less on average when they are held by their mothers than if held using conventional methods? Test at the 5% level.

Table 9.3.11: Crying Time of Infants Given Shots Using Conventional Methods

63	0	2	46	33	33
29	23	11	12	48	15
33	14	51	37	24	70
63	0	73	39	54	52
39	34	30	55	58	18

Table 9.3.12: Crying Time of Infants Given Shots Using New Methods

0	32	20	23	14	19
60	59	64	64	72	50
44	14	10	58	19	41
17	5	36	73	19	46
9	43	73	27	25	18

10. A vitamin K shot is given to infants soon after birth. Nurses at Northbay Healthcare were involved in a study to see if how they handle the infants could reduce the pain the infants feel ("SOCR data nips," 2013). One of the measurements taken was how long, in seconds, the infant cried after being given the shot. A random sample was taken from the group that was given the shot using conventional methods (Example 9.3.11), and a random sample was taken from the group that was given the shot where the mother held the infant prior to and during the shot (Example 9.3.12). Calculate a 95% confidence interval for the mean difference in mean crying time after being given a vitamin K shot between infants held using conventional methods and infants held by their mothers.
11. Redo problem 1 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).
12. Redo problem 2 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).
13. Redo problem 7 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).
14. Redo problem 8 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).
15. Redo problem 9 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).
16. Redo problem 10 testing for the assumption of equal variances and then use the formula that utilizes the assumption of equal variances (follow the procedure in Example 9.3.3).

Answer

For all hypothesis tests, just the conclusion is given. For all confidence intervals, just the interval using technology is given. See solution for the entire answer.

1. Reject Ho
2. $\$65443.80 < \mu_1 - \mu_2 < \$13340.80$
3. Fail to reject Ho
4. $-51564.6\text{mm}^3 < \mu_1 - \mu_2 < 75656.6\text{mm}^3$

5. Reject Ho
6. $23.2818\text{km} < \mu_1 - \mu_2 < 103.67\text{km}$
7. Reject Ho
8. $4.3641 < \mu_1 - \mu_2 < 37.5276$
9. Fail to reject Ho
10. $-10.9726\text{s} < \mu_1 - \mu_2 < 11.3059\text{s}$
11. Reject Ho
12. $\$6544.98 < \mu_1 - \mu_2 < \$13339.60$
13. Reject Ho
14. $4.8267 < \mu_1 - \mu_2 < 37.0649$
15. Fail to reject Ho
16. $-10.9713\text{s} < \mu_1 - \mu_2 < 11.3047\text{s}$

This page titled 9.3: Independent Samples for Two Means is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

9.4: Which Analysis Should You Conduct?

One of the most important concept that you need to understand is deciding which analysis you should conduct for a particular situation. To help you to figure out the analysis to conduct, there are a series of questions you should ask yourself.

1. Does the problem deal with mean or proportion?
   Sometimes the problem states explicitly the words mean or proportion, but other times you have to figure it out based on the information you are given. If you counted number of individuals that responded in the affirmative to a question, then you are dealing with proportion. If you measured something, then you are dealing with mean.
2. Does the problem have one or two samples?
   So look to see if one group was measured or if two groups were measured. If you have the data sets, then it is usually easy to figure out if there is one or two samples, then there is either one data set or two data sets. If you don't have the data, then you need to decide if the problem describes collecting data from one group or from two groups.
3. If you have two samples, then you need to determine if the samples are independent or dependent.
   If the individuals are different for both samples, then most likely the samples are independent. If you can't tell, then determine if a data value from the first sample influences the data value in the second sample. In other words, can you pair data values together so you can find the difference, and that difference has meaning. If the answer is yes, then the samples are paired. Otherwise, the samples are independent.
4. Does the situation involve a hypothesis test or a confidence interval?
   If the problem talks about "do the data show", "is there evidence of", "test to see", then you are doing a hypothesis test. If the problem talks about "find the value", "estimate the" or "find the interval", then you are doing a confidence interval.

So if you have a situation that has two samples, independent samples, involving the mean, and is a hypothesis test, then you have a two-sample independent t-test. Now you look up the assumptions and the formula or technology process for doing this test. Every hypothesis test involves the same six steps, and you just have to use the correct assumptions and calculations. Every confidence interval has the same five steps, and again you just need to use the correct assumptions and calculations. So this is why it is so important to figure out what analysis you should conduct.

Data Sources:

AP exam scores. (2013, November 20). Retrieved from wiki.stat.ucla.edu/socr/index...08_APExamScore s

Buy sushi grade fish online. (2013, November 20). Retrieved from http://www.catalinaop.com/

Center for Disease Control and Prevention, Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network. (2008). *Autism and developmental disabilities monitoring network-2012*. Retrieved from website: www.cdc.gov/ncbddd/autism/doc...nityReport.pdf

Cholesterol levels after heart attack. (2013, September 25). Retrieved from http://www.statsci.org/data/general/cholest.html

Flanagan, R., Rooney, C., & Griffiths, C. (2005). Fatal poisoning in childhood, england & wales 1968-2000. *Forensic Science International*, 148:121-129, Retrieved from http://www.cdc.gov/nchs/data/ice/fat...ning_child.pdf

Friday the 13th datafile. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Datafil...aythe13th.html

Gettler, L. T., McDade, T. W., Feranil, A. B., & Kuzawa, C. W. (2011). Longitudinal evidence that fatherhood decreases testosterone in human males. *The Proceedings of the National Academy of Sciences, PNAS 2011*, doi: 10.1073/pnas.1105403108 Length of NZ rivers. (2013, September 25). Retrieved from http://www.statsci.org/data/oz/nzrivers.html

Lim, L. L. United Nations, International Labour Office. (2002). *Female labour-force participation*. Retrieved from website: www.un.org/esa/population/pub...ty/RevisedLIMp aper.PDF

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/...s.aspx?ind=137

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, 1(2), 108-13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9765732

Population reference bureau. (2013, October 8). Retrieved from http://www.prb.org/DataFinder/Topic/...gs.aspx?ind=25

Seafood online. (2013, November 20). Retrieved from http://www.allfreshseafood.com/

SOCR 012708 id data hotdogs. (2013, November 13). Retrieved from http://wiki.stat.ucla.edu/socr/index...D_Data_HotDogs

SOCR data nips infantvitK shotdata. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...tVitK_ShotData

SOCR data Oct2009 id ni. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index..._Oct2009_ID_NI

Statistics brain. (2013, November 30). Retrieved from http://www.statisticbrain.com/infidelity-statistics/

Student t-distribution. (2013, November 25). Retrieved from lib.stat.cmu.edu/DASL/Stories/student.html

This page titled 9.4: Which Analysis Should You Conduct? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

10: Regression and Correlation

The previous chapter looked at comparing populations to see if there is a difference between the two. That involved two random variables that are similar measures. This chapter will look at two random variables that are not similar measures, and see if there is a relationship between the two variables. To do this, you look at regression, which finds the linear relationship, and correlation, which measures the strength of a linear relationship.

Note

There are many other types of relationships besides linear that can be found for the data. This book will only explore linear, but realize that there are other relationships that can be used to describe data.

10.1: Regression
10.2: Correlation
10.3: Inference for Regression and Correlation

This page titled 10: Regression and Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Regression

When comparing two different variables, two questions come to mind: “Is there a relationship between two variables?” and “How strong is that relationship?” These questions can be answered using **regression** and **correlation**. Regression answers whether there is a relationship (again this book will explore linear only) and correlation answers how strong the linear relationship is. To introduce both of these concepts, it is easier to look at a set of data.

Example 10.1.1 if there is a relationship

Is there a relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is one a random sample was taken of beer’s alcohol content and calories ("Calories in beer," 2011), and the data is in Example 10.1.1.

Table 10.1.1: Alcohol and Calorie Content in Beer

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
O'Doul's	Anheuser Busch	0.40%	70
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution

To aid in figuring out if there is a relationship, it helps to draw a scatter plot of the data. It is helpful to state the random variables, and since in an algebra class the variables are represented as x and y, those labels will be used here. It helps to state which variable is x and which is y.

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

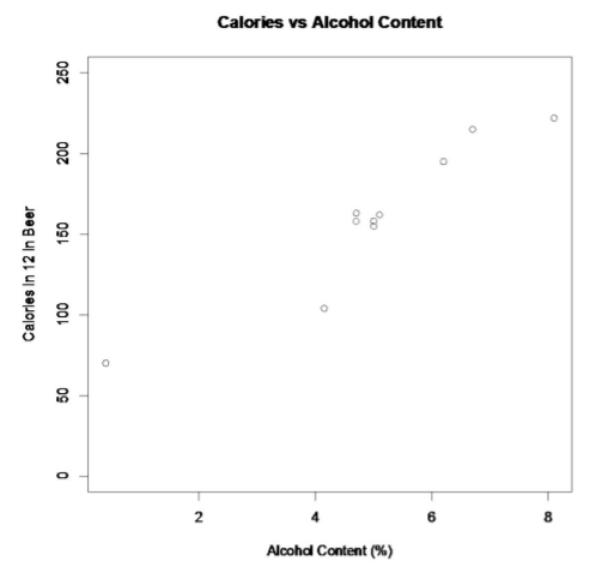


Figure 10.1.1: Scatter Plot of Beer Data

This scatter plot looks fairly linear. However, notice that there is one beer in the list that is actually considered a non-alcoholic beer. That value is probably an outlier since it is a non-alcoholic beer. The rest of the analysis will not include O'Doul's. You cannot just remove data points, but in this case it makes more sense to, since all the other beers have a fairly large alcohol content.

To find the equation for the linear relationship, the process of regression is used to find the line that best fits the data (sometimes called the best fitting line). The process is to draw the line through the data and then find the distances from a point to the line, which are called the residuals. The regression line is the line that makes the square of the residuals as small as possible, so the regression line is also sometimes called the least squares line. The regression line and the residuals are displayed in *Figure* 10.1.2.

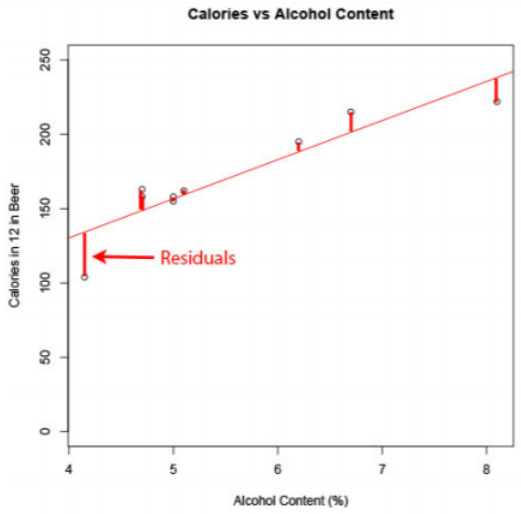


Figure 10.1.2: Scatter Plot of Beer Data with Regression Line and Residuals

To find the regression equation (also known as best fitting line or least squares line)

Given a collection of paired sample data, the regression equation is

$\hat{y} = a + bx$

where the slope = $b = \frac{SS_{xy}}{SS_x}$ and y-intercept = $a = \overline{y} - b\overline{x}$

Definition 10.1.1

The **residuals** are the difference between the actual values and the estimated values.

$$\text{residual} = y - \hat{y}$$

Definition 10.1.2

SS stands for sum of squares. So you are summing up squares. With the subscript *xy*, you aren't really summing squares, but you can think of it that way in a weird sense.

$$SS_{xy} = \sum(x - \overline{x})(y - \overline{y})$$
$$SS_x = \sum (x - \overline{x})^2$$
$$SS_y = \sum (y - \overline{y})^2$$

Note

The easiest way to find the regression equation is to use the technology.

The **independent variable**, also called the **explanatory variable** or **predictor variable**, is the *x*-value in the equation. The independent variable is the one that you use to predict what the other variable is. The **dependent variable** depends on what independent value you pick. It also responds to the explanatory variable and is sometimes called the **response variable**. In the alcohol content and calorie example, it makes slightly more sense to say that you would use the alcohol content on a beer to predict the number of calories in the beer.

Definition 10.1.3

The **population equation** looks like:

$$y = \beta_o + \beta_1 x$$
$$\beta_o = \text{ slope}$$
$$\beta_1 = y\text{-intercept}$$

$\hat{y}$ is used to predict *y*.

Assumptions of the regression line:

a. The set (x, y) of ordered pairs is a random sample from the population of all such possible (x, y) pairs.
b. For each fixed value of *x*, the *y*-values have a normal distribution. All of the *y* distributions have the same variance, and for a given *x*-value, the distribution of *y*-values has a mean that lies on the least squares line. You also assume that for a fixed *y*, each *x* has its own normal distribution. This is difficult to figure out, so you can use the following to determine if you have a normal distribution.
    i. Look to see if the scatter plot has a linear pattern.
    ii. Examine the residuals to see if there is randomness in the residuals. If there is a pattern to the residuals, then there is an issue in the data.

Example 10.1.2 find the equation of the regression line

a. Is there a positive relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear relationship, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.1.2.
b. Use the regression equation to find the number of calories when the alcohol content is 6.50%.
c. Use the regression equation to find the number of calories when the alcohol content is 2.00%.
d. Find the residuals and then plot the residuals versus the *x*-values.

Table 10.1.2: Alcohol and Caloric Content in Beer without Outlier

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163

Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
O'Doul's	Anheuser Busch	0.40%	70
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution

a. State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Assumptions check:

a. A random sample was taken as stated in the problem.
b. The distribution for each calorie value is normally distributed for every value of alcohol content in the beer.
   i. From Example 10.1.1, the scatter plot looks fairly linear.
   ii. The residual versus the x-values plot looks fairly random. (See *Figure* 10.1.5.)

It appears that the distribution for calories is a normal distribution.

To find the regression equation on the TI-83/84 calculator, put the x's in L1 and the y's in L2. Then go to STAT, over to TESTS, and choose LinRegTTest. The setup is in *Figure* 10.1.3. The reason that >0 was chosen is because the question was asked if there was a positive relationship. If you are asked if there is a negative relationship, then pick <0. If you are just asked if there is a relationship, then pick $\neq 0$. Right now the choice will not make a different, but it will be important later.

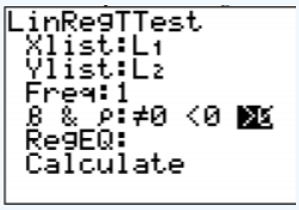


Figure 10.1.3: Setup for Linear Regression Test on TI-83/84

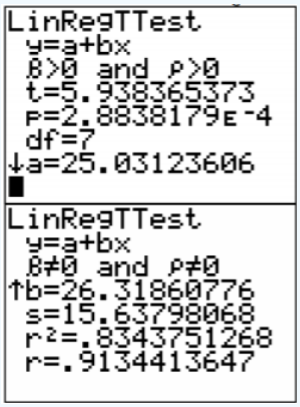


Figure 10.1.4: Results for Linear Regression Test on TI-83/84

From this you can see that

$\hat{y} = 25.0 + 26.3x$

To find the regression equation using R, the command is lm(dependent variable ~ independent variable), where ~ is the tilde symbol located on the upper left of most keyboards. So for this example, the command would be lm(calories ~ alcohol), and the output would be

Call:

lm(formula = calories ~ alcohol)

Coefficients:

(Intercept) alcohol

25.03 26.32

From this you can see that the *y*-intercept is 25.03 and the slope is 26.32. So the regression equation is $\hat{y} = 25.0 + 26.3x$.

Remember, this is an estimate for the true regression. A different random sample would produce a different estimate.

b. $x_o = 6.50$
$\hat{y} = 25.0 + 26.3(6.50) = 196 \text{ calories}$

If you are drinking a beer that is 6.50% alcohol content, then it is probably close to 196 calories. Notice, the mean number of calories is 170 calories. This value of 196 seems like a better estimate than the mean when looking at the original data. The regression equation is a better estimate than just the mean.

c. $x_o = 2.00$
$\hat{y} = 25.0 + 26.3(2.00) = 78 \text{ calories}$

If you are drinking a beer that is 2.00% alcohol content, then it has probably close to 78 calories. This doesn't seem like a very good estimate. This estimate is what is called extrapolation. It is not a good idea to predict values that are far outside the range of the original data. This is because you can never be sure that the regression equation is valid for data outside the original data.

d. To find the residuals, find $\hat{y}$ for each *x*-value. Then subtract each $\hat{y}$ from the given *y* value to find the residuals. Realize that these are sample residuals since they are calculated from sample values. It is best to do this in a spreadsheet.

Table 10.1.3: Residuals for Beer Calories

x	y	$\hat{y} = 25.0 + 26.3x$	$y - \hat{y}$
4.70	163	148.61	14.390
6.70	215	201.21	13.790
8.10	222	238.03	-16.030
4.15	104	134.145	-30.145
5.10	162	159.13	2.870
5.00	158	156.5	1.500
5.00	155	156.5	-1.500
4.70	158	148.61	9.390
6.20	195	188.06	6.940

Notice the residuals add up to close to 0. They don't add up to exactly 0 in this example because of rounding error. Normally the residuals add up to 0.

You can use R to get the residuals. The command is

lm.out = lm(dependent variable ~ independent variable) – this defines the linear model with a name so you can use it later.

Then residual(lm.out) – produces the residuals.

For this example, the command would be

lm(calories~alcohol)

Call:

lm(formula = calories ~ alcohol)

Coefficients:

(Intercept) alcohol

25.03 26.32

> residuals(lm.out)

1	2	3	4	5	6	7	8	9
14.271307	13.634092	−16.211959	−30.253458	2.743864	1.375725	−1.624275	9.271307	6.793396

So the first residual is 14.271307 and it belongs to the first x value. The residual 13.634092 belongs to the second x value, and so forth.

You can then graph the residuals versus the independent variable using the plot command. For this example, the command would be plot(alcohol, residuals(lm.out), main="Residuals for Beer Calories versus Alcohol Content", xlab="Alcohol Content", ylab="Residuals"). Sometimes it is useful to see the x-axis on the graph, so after creating the plot, type the command abline(0,0).

The graph of the residuals versus the x-values is in *Figure* 10.1.5. They appear to be somewhat random.

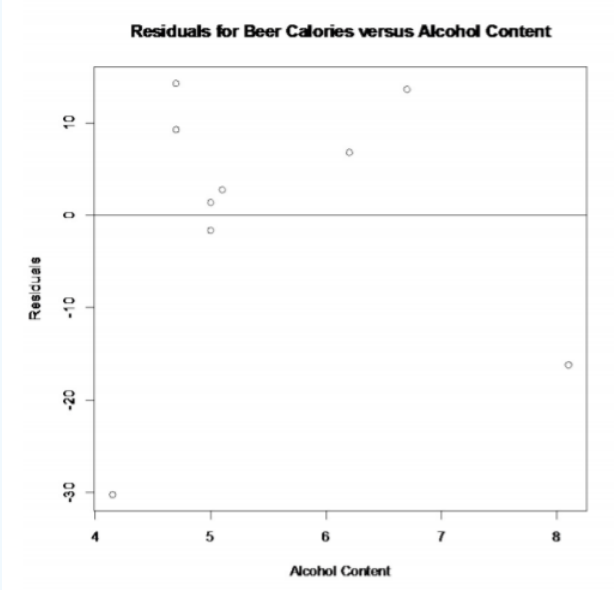


Figure 10.1.5: Residuals of Beer Calories versus Content

Notice, that the 6.50% value falls into the range of the original *x*-values. The processes of predicting values using an *x* within the range of original *x*-values is called **interpolating**. The 2.00% value is outside the range of original *x*-values. Using an *x*-value that is outside the range of the original *x*-values is called **extrapolating**. When predicting values using interpolation, you can usually feel pretty confident that that value will be close to the true value. When you extrapolate, you are not really sure that the predicted value is close to the true value. This is because when you interpolate, you know the equation that predicts, but when you extrapolate, you are not really sure that your relationship is still valid. The relationship could in fact change for different *x*-values.

An example of this is when you use regression to come up with an equation to predict the growth of a city, like Flagstaff, AZ. Based on analysis it was determined that the population of Flagstaff would be well over 50,000 by 1995. However, when a census was undertaken in 1995, the population was less than 50,000. This is because they extrapolated and the growth factor they were using had obviously changed from the early 1990's. Growth factors can change for many reasons, such as employment growth,

employment stagnation, disease, articles saying great place to live, etc. Realize that when you extrapolate, your predicted value may not be anywhere close to the actual value that you observe.

What does the slope mean in the context of this problem?

$$m = \frac{\Delta y}{\Delta x} = \frac{\Delta \text{ calories}}{\Delta \text{ alcohol content}} = \frac{26.3 \text{ calories}}{1\%}$$

The calories increase 26.3 calories for every 1% increase in alcohol content.

The *y*-intercept in many cases is meaningless. In this case, it means that if a drink has 0 alcohol content, then it would have 25.0 calories. This may be reasonable, but remember this value is an extrapolation so it may be wrong.

Consider the residuals again. According to the data, a beer with 6.7% alcohol has 215 calories. The predicted value is 201 calories.

$$\text{Residual} = \text{actual} - \text{predicted}$$

$$= 215 - 201$$

$$= 14$$

This deviation means that the actual value was 14 above the predicted value. That isn't that far off. Some of the actual values differ by a large amount from the predicted value. This is due to variability in the dependent variable. The larger the residuals the less the model explains the variability in the dependent variable. There needs to be a way to calculate how well the model explains the variability in the dependent variable. This will be explored in the next section.

The following example demonstrates the process to go through when using the formulas for finding the regression equation, though it is better to use technology. This is because if the linear model doesn't fit the data well, then you could try some of the other models that are available through technology.

Example 10.1.3 calculating the regression equation with the formula

Is there a relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is one a random sample was taken of beer's alcohol content and calories ("Calories in beer," 2011), and the data are in Example 10.1.2. Find the regression equation from the formula.

Solution

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Table 10.1.4: Calculations for Regression Equation

Alcohol Content	Calories	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
4.70	163	-0.8167	-7.2222	0.6669	52.1065	5.8981
6.70	215	1.1833	44.7778	1.4003	2005.0494	52.9870
8.10	222	2.5833	51.7778	6.6736	2680.9383	133.7595
4.15	104	-1.3667	-66.2222	1.8678	4385.3827	90.5037
5.10	162	-0.4167	-8.2222	0.1736	67.6049	3.4259
5.00	158	-0.5167	-12.2222	0.2669	149.3827	6.3148
5.00	155	-0.5167	-15.2222	0.2669	231.7160	7.8648
4.70	158	-0.8167	-12.2222	0.6669	149.3827	9.9815
6.20	195	0.6833	24.7778	0.4669	613.9383	16.9315
5.516667 = $\overline{x}$	170.2222 = $\overline{y}$			12.45 = SS_x	10335.5556 = SS_y	327.6667 = SS_{xy}

slope: $b = \frac{SS_{xy}}{SS_x} = \frac{327.6667}{12.45} \approx 26.3$

y-intercept: $a = \overline{y} - b\overline{x} = 170.222 - 26.3(5.516667) \approx 25.0$

Regression equation: $\hat{y} = 25.0 + 26.3x$

Homework

Exercise 10.1.1

For each problem, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the regression analysis with and without the suspected outlier points to determine if their removal affects the regression. The data sets in this section are used in the homework for sections 10.2 and 10.3 also.

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in Example 10.1.5 ("Prediction of height," 2013). Create a scatter plot and find a regression equation between the height of a person and the length of their metacarpal. Then use the regression equation to find the height of a person for a metacarpal length of 44 cm and for a metacarpal length of 55 cm. Which height that you calculated do you think is closer to the true height of the person? Why?

Table 10.1.5: Data of Metacarpal versus Height

Length of Metacarpal (cm)	Height of Person (cm)
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

2. Example 10.1.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Create a scatter plot and find a regression equation between house value and rental income. Then use the regression equation to find the rental income a house worth $230,000 and for a house worth $400,000. Which rental income that you calculated do you think is closer to the true rental income? Why?

Table 10.1.6: Data of House Value versus Rental

Value	Rental	Value	Rental	Value	Rental	Value	Rental
81000	6656	77000	4576	75000	7280	67500	6864
95000	7904	94000	8736	90000	6240	85000	7072
121000	12064	115000	7904	110000	7072	104000	7904
135000	8320	130000	9776	126000	6240	125000	7904
145000	8320	140000	9568	140000	9152	135000	7488
165000	13312	165000	8528	155000	7488	148000	8320
178000	11856	174000	10400	170000	9568	170000	12688
200000	12272	200000	10608	194000	11232	190000	8320

Value	Rental	Value	Rental	Value	Rental	Value	Rental
214000	8528	208000	10400	200000	10400	200000	8320
240000	10192	240000	12064	240000	11648	225000	12480
289000	11648	270000	12896	262000	10192	244500	11232
325000	12480	310000	12480	303000	12272	300000	12480

3. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example 10.1.7. Create a scatter plot of the data and find a linear regression equation between fertility rate and life expectancy. Then use the regression equation to find the life expectancy for a country that has a fertility rate of 2.7 and for a country with fertility rate of 8.1. Which life expectancy that you calculated do you think is closer to the true life expectancy? Why?

Table 10.1.7: Data of Fertility Rates versus Life Expectancy

Fertility Rate	Life Expectancy
1.7	77.2
5.8	55.4
2.2	69.9
2.1	76.4
1.8	75.0
2.0	78.2
2.6	73.0
2.8	70.8
1.4	82.6
2.6	68.9
1.5	81.0
6.9	54.2
2.4	67.1
1.5	73.3
2.5	74.2
1.4	80.7
2.9	72.1
2.1	78.3
4.7	62.9
6.8	54.4
5.2	55.9
4.2	66.0
1.5	76.0
3.9	72.3

4. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information are available for the year 2011 is in Example 10.1.8. Create a scatter plot of the data and find a regression equation between percentage spent on health expenditure and the percentage of women receiving prenatal care. Then use the regression equation to find the percent of women receiving prenatal care for a country that spends 5.0% of GDP on health expenditure and for a country that spends 12.0% of GDP. Which prenatal care percentage that you calculated do you think is closer to the true percentage? Why?

Table 10.1.8: Data of Health Expenditure versus Prenatal Care

Health Expenditure (% of GDP)	Prenatal Care (%)
9.6	47.9
3.7	54.6
5.2	93.7
5.2	84.7
10.0	100.0
4.7	42.5
4.8	96.4
6.0	77.1
5.4	58.3
4.8	95.4
4.1	78.0
6.0	93.3
9.5	93.3
6.8	93.7
6.1	89.8

5. The height and weight of baseball players are in Example 10.1.9 ("MLB heightsweights," 2013). Create a scatter plot and find a regression equation between height and weight of baseball players. Then use the regression equation to find the weight of a baseball player that is 75 inches tall and for a baseball player that is 68 inches tall. Which weight that you calculated do you think is closer to the true weight? Why?

Table 10.1.9: Heights and Weights of Baseball Players

Height (inches)	Weight (pounds)
76	212
76	224
72	180
74	210
75	215
71	200
77	235
78	235
77	194

Height (inches)	Weight (pounds)
76	185
72	180
72	170
75	220
74	228
73	210
72	180
70	185
73	190
71	186
74	200
74	200
75	210
79	240
72	208
75	180

6. Different species have different body weights and brain weights are in Example 10.1.10. ("Brain2bodyweight," 2013). Create a scatter plot and find a regression equation between body weights and brain weights. Then use the regression equation to find the brain weight for a species that has a body weight of 62 kg and for a species that has a body weight of 180,000 kg. Which brain weight that you calculated do you think is closer to the true brain weight? Why?

Table 10.1.10: Body Weights and Brain Weights of Species

Species	Body Weight (kg)	Brain Weight (kg)
Newborn Human	3.20	0.37
Adult Human	73.00	1.35
Pithecanthropus Man	70.00	0.93
Squirrel	0.80	0.01
Hamster	0.15	0.00
Chimpanzee	50.00	0.42
Rabbit	1.40	0.01
Dog (Beagle)	10.00	0.07
Cat	4.50	0.03
Rat	0.40	0.00
Bottle-Nosed Dolphin	400.00	1.50
Beaver	24.00	0.04
Gorilla	320.00	0.50

Species	Body Weight (kg)	Brain Weight (kg)
Tiger	170.00	0.26
Owl	1.50	0.00
Camel	550.00	0.76
Elephant	4600.00	6.00
Lion	187.00	0.24
Sheep	120.00	0.14
Walrus	800.00	0.93
Horse	450.00	0.50
Cow	700.00	0.44
Giraffe	950.00	0.53
Green Lizard	0.20	0.00
Sperm Whale	35000.00	7.80
Turtle	3.00	0.00
Alligator	270.00	0.01

7. A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in Example 10.1.11. Create a scatter plot and find a regression equation between amount of calories and amount of sodium. Then use the regression equation to find the amount of sodium a beef hotdog has if it is 170 calories and if it is 120 calories. Which sodium level that you calculated do you think is closer to the true sodium level? Why?

Table 10.1.11: Calories and Sodium Levels in Beef Hotdogs

Calories	Sodium
186	495
181	477
176	425
149	322
184	482
190	587
158	370
139	322
175	479
148	375
152	330
111	300
141	386
153	401
190	645

Calories	Sodium
157	440
131	317
149	319
135	298
132	253

8. Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in Example 10.1.12 ("OECD economic development," 2013). Create a scatter plot and find a regression equation between percent of labor force in agriculture and per capita income. Then use the regression equation to find the per capita income in a country that has 21 percent of labor in agriculture and in a country that has 2 percent of labor in agriculture. Which per capita income that you calculated do you think is closer to the true income? Why?

Table 10.1.12: Percent of Labor in Agriculture and Per Capita Income for European Countries

Country	Percent in Agriculture	Per Capita Income
Sweden	14	1644
Switzerland	11	1361
Luxembourg	15	1242
U. Kingdom	4	1105
Denmark	18	1049
W. Germany	15	1035
France	20	1013
Belgium	6	1005
Norway	20	977
Iceland	25	839
Netherlands	11	810
Austria	23	681
Ireland	36	529
Italy	27	504
Greece	56	324
Spain	42	290
Portugal	44	238
Turkey	79	177

9. Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in Example 10.1.13 ("Smoking and cancer," 2013). Create a scatter plot and find a regression equation between cigarette smoking and deaths of bladder cancer. Then use the regression equation to find the number of deaths from bladder cancer when the cigarette sales were 20 per capita and when the cigarette sales were 6 per capita. Which number of deaths that you calculated do you think is closer to the true number? Why?

Table 10.1.13: Number of Cigarettes and Number of Bladder Cancer Deaths in 1960

Cigarette Sales (per Capita)	Bladder Cancer Deaths (per 100 thousand)	Cigarette Sales (per Capita)	Bladder Cancer Deaths (per 100 Thousand)
18.20	2.90	42.40	6.54
25.82	3.52	28.64	5.98
18.24	2.99	21.16	2.90
28.60	4.46	29.14	5.30
31.10	5.11	19.96	2.89
33.60	4.78	26.38	4.47
40.46	5.60	23.44	2.93
28.27	4.46	23.78	4.89
20.10	3.08	29.18	4.99
27.91	4.75	18.06	3.25
26.18	4.09	20.94	3.64
22.12	4.23	20.08	2.94
21.84	2.91	22.57	3.21
23.44	2.86	14.00	3.31
21.58	4.65	25.89	4.63
28.92	4.79	21.17	4.04
25.91	5.21	21.25	5.14
26.92	4.69	22.86	4.78
24.96	5.27	28.04	3.20
22.06	3.72	30.34	3.46
16.08	3.06	23.75	3.95
27.56	4.04	23.32	3.72

10. The weight of a car can influence the mileage that the car can obtain. A random sample of cars' weights and mileage was collected and are in Example 10.1.14 ("Passenger car mileage," 2013). Create a scatter plot and find a regression equation between weight of cars and mileage. Then use the regression equation to find the mileage on a car that weighs 3800 pounds and on a car that weighs 2000 pounds. Which mileage that you calculated do you think is closer to the true mileage? Why?

Table 10.1.14: Weights and Mileages of Cars

Weight (100 pounds)	Mileage (mpg)
22.5	53.3
22.5	41.1
22.5	38.9
25.0	40.9
27.5	46.9
27.5	36.3
30.0	32.2

Weight (100 pounds)	Mileage (mpg)
30.0	32.2
30.0	31.5
30.0	31.4
30.0	31.4
35.0	32.6
35.0	31.3
35.0	31.3
35.0	28.0
35.0	28.0
35.0	28.0
40.0	23.6
40.0	23.6
40.0	23.4
40.0	23.1
45.0	19.5
45.0	17.2
45.0	17.0
55.0	13.2

Answer

For regression, only the equation is given. See solutions for the entire answer.

1. $\hat{y} = 1.719x + 93.709$
3. $\hat{y} = -4.706x + 84.873$
5. $\hat{y} = 5.859x - 230.942$
7. $\hat{y} = 4.0133x - 228.3313$
9. $\hat{y} = 0.12182x + 1.08608$

This page titled 10.1: Regression is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

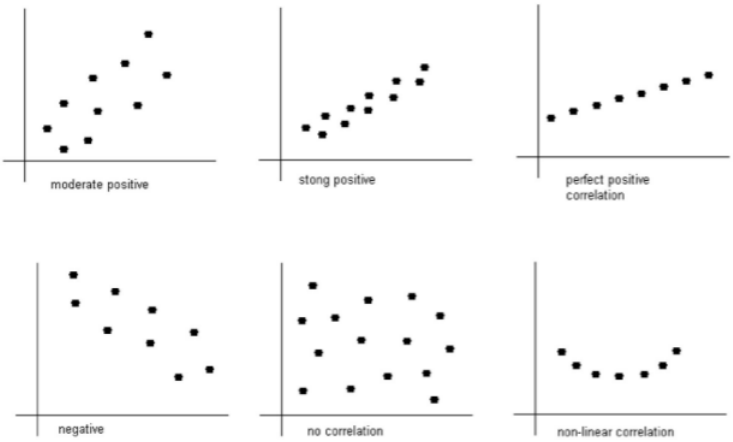
10.2: Correlation

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

When you see a pattern in the data you say there is a correlation in the data. Though this book is only dealing with linear patterns, patterns can be exponential, logarithmic, or periodic. To see this pattern, you can draw a scatter plot of the data.

Remember to read graphs from left to right, the same as you read words. If the graph goes up the correlation is positive and if the graph goes down the correlation is negative.

The words " weak", "moderate", and "strong" are used to describe the strength of the relationship between the two variables.



Figures

The **linear correlation coefficient** is a number that describes the strength of the linear relationship between the two variables. It is also called the Pearson correlation coefficient after Karl Pearson who developed it. The symbol for the sample linear correlation coefficient is *r*. The symbol for the population correlation coefficient is ρ (Greek letter rho).

The formula for *r* is

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Where

$SS_x = \sum (x - \overline{x})^2$
$SS_y = \sum (y - \overline{y})^2$
$SS_{xy} = \sum (x - \overline{x})(y - \overline{y})$

Assumptions of linear correlation are the same as the assumptions for the regression line:

a. The set (*x*, *y*) of ordered pairs is a random sample from the population of all such possible (*x*, *y*) pairs.
b. For each fixed value of *x*, the *y* -values have a normal distribution. All of the *y* -distributions have the same variance, and for a given *x*-value, the distribution of *y*-values has a mean that lies on the least squares line. You also assume that for a fixed *y*, each *x* has its own normal distribution. This is difficult to figure out, so you can use the following to determine if you have a normal distribution.
    i. Look to see if the scatter plot has a linear pattern.
    ii. Examine the residuals to see if there is randomness in the residuals. If there is a pattern to the residuals, then there is an issue in the data.

Note

Interpretation of the correlation coefficient

r is always between -1 and 1. *r* = -1 means there is a perfect negative linear correlation and *r* = 1 means there is a perfect positive correlation. The closer *r* is to 1 or -1, the stronger the correlation. The closer *r* is to 0, the weaker the correlation.

CAREFUL: $r = 0$ does not mean there is no correlation. It just means there is **no linear correlation**. There might be a very strong curved pattern.

r

How strong is the positive relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in *Table* 10.2.1. Find the correlation coefficient and interpret that value.

Table 10.2.1: Alcohol and Calorie Content in Beer without Outlier

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Assumptions check:

From Example 10.2.2, the assumptions have been met.

To compute the correlation coefficient using the TI-83/84 calculator, use the LinRegTTest in the STAT menu. The setup is in *Figure* 10.2.2. The reason that >0 was chosen is because the question was asked if there was a positive correlation. If you are asked if there is a negative correlation, then pick <0. If you are just asked if there is a correlation, then pick $\neq 0$. Right now the choice will not make a different, but it will be important later.

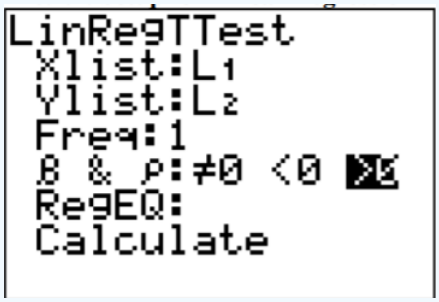


Figure 10.2.2: Setup for Linear Regression Test on TI-83/84

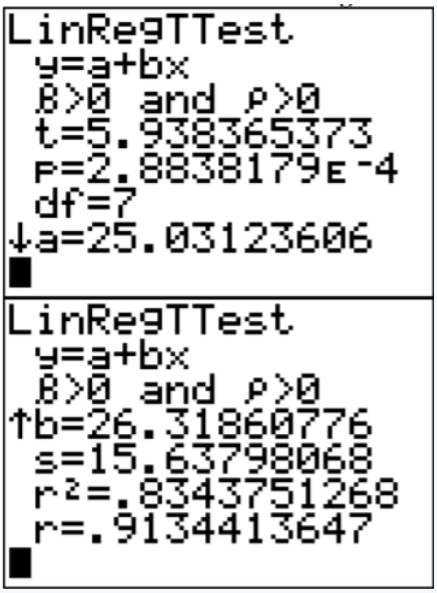


Figure 10.2.3: Results for Linear Regression Test on TI-83/84

To compute the correlation coefficient in R, the command is cor(independent variable, dependent variable). So for this example the command would be cor(alcohol, calories). The output is

[1] 0.9134414

The correlation coefficient is $r = 0.913$. This is close to 1, so it looks like there is a strong, positive correlation.

Causation

One common mistake people make is to assume that because there is a correlation, then one variable causes the other. This is usually not the case. That would be like saying the amount of alcohol in the beer causes it to have a certain number of calories. However, fermentation of sugars is what causes the alcohol content. The more sugars you have, the more alcohol can be made, and the more sugar, the higher the calories. It is actually the amount of sugar that causes both. Do not confuse the idea of correlation with the concept of causation. Just because two variables are correlated does not mean one causes the other to happen.

Example 10.2.2 correlation versus Causation

a. A study showed a strong linear correlation between per capita beer consumption and teacher's salaries. Does giving a teacher a raise cause people to buy more beer? Does buying more beer cause teachers to get a raise?
b. A study shows that there is a correlation between people who have had a root canal and those that have cancer. Does that mean having a root canal causes cancer?

Solution

a. There is probably some other factor causing both of them to increase at the same time. Think about this: In a town where people have little extra money, they won't have money for beer and they won't give teachers raises. In another town where people have more extra money to spend it will be easier for them to buy more beer and they would be more willing to give teachers raises.

b. Just because there is positive correlation doesn't mean that one caused the other. It turns out that there is a positive correlation between eating carrots and cancer, but that doesn't mean that eating carrots causes cancer. In other words, there are lots of relationships you can find between two variables, but that doesn't mean that one caused the other.

Remember a correlation only means a pattern exists. It does not mean that one variable causes the other variable to change.

Explained Variation

As stated before, there is some variability in the dependent variable values, such as calories. Some of the variation in calories is due to alcohol content and some is due to other factors. How much of the variation in the calories is due to alcohol content?

When considering this question, you want to look at how much of the variation in calories is explained by alcohol content and how much is explained by other variables. Realize that some of the changes in calories have to do with other ingredients. You can have two beers at the same alcohol content, but beer one has higher calories because of the other ingredients. Some variability is explained by the model and some variability is not explained. Together, both of these give the total variability. This is

$$\begin{array}{ccccc} (\text{total variation}) & = & (\text{explained variation}) & + & (\text{unexplained variation}) \\ \sum (y-\overline{y})^2 & = & \sum (\hat{y}-\overline{y})^2 & + & \sum (y-\hat{y})^2 \end{array}$$

Note

The proportion of the variation that is explained by the model is

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

This is known as the **coefficient of determination**.

To find the coefficient of determination, you square the correlation coefficient. In addition, r^2 is part of the calculator results.

Example 10.2.3 finding the coefficient of determination

Find the coefficient of variation in calories that is explained by the linear relationship between alcohol content and calories and interpret the value.

Solution

From the calculator results,

$r^2 = 0.8344$

Using R, you can do (cor(independent variable, dependent variable))^2. So that would be (cor(alcohol, calories))^2, and the output would be

[1] 0.8343751

Or you can just use a calculator and square the correlation value.

Thus, 83.44% of the variation in calories is explained to the linear relationship between alcohol content and calories. The other 16.56% of the variation is due to other factors. A really good coefficient of determination has a very small, unexplained part.

and r^2

How strong is the relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.2.1. Find the correlation coefficient and the coefficient of determination using the formula.

Solution

From Example 10.2.2, $SS_x = 12.45, SS_y = 10335.5556, SS_{xy} = 327.6667$

Correlation coefficient:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{327.6667}{\sqrt{12.45 * 10335.5556}} \approx 0.913$$

Coefficient of determination:

$r^2 = (r)^2 = (0.913)^2 \approx 0.834$

Now that you have a correlation coefficient, how can you tell if it is significant or not? This will be answered in the next section.

Homework

Exercise 10.2.1

For each problem, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the correlation analysis with and without the suspected outlier points to determine if their removal affects the correlation. The data sets in this section are in section 10.1 and will be used in section 10.3.

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in Example 10.2.5 ("Prediction of height," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
2. Example 10.2.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
3. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example 10.2.7. Find the correlation coefficient and coefficient of determination and then interpret both.
4. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in Example 10.2.8. Find the correlation coefficient and coefficient of determination and then interpret both.
5. The height and weight of baseball players are in Example 10.2.9 ("MLB heightsweights," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
6. Different species have different body weights and brain weights are in Example 10.2.10. ("Brain2bodyweight," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
7. A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in Example 10.2.11. Find the correlation coefficient and coefficient of determination and then interpret both.
8. Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in Example 10.2.12 ("OECD economic development," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
9. Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in Example 10.2.13 ("Smoking and cancer," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
10. The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in Example 10.2.14 ("Passenger car mileage," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
11. There is a negative correlation between police expenditure and crime rate. Does this mean that spending more money on police causes the crime rate to decrease? Explain your answer.
12. There is a positive correlation between tobacco sales and alcohol sales. Does that mean that using tobacco causes a person to also drink alcohol? Explain your answer.
13. There is a positive correlation between the average temperature in a location and the morality rate from breast cancer. Does that mean that higher temperatures cause more women to die of breast cancer? Explain your answer.
14. There is a positive correlation between the length of time a tableware company polishes a dish and the price of the dish. Does that mean that the time a plate is polished determines the price of the dish? Explain your answer.

Answer

Only the correlation coefficient and coefficient of determination are given. See solutions for the entire answer.

1. r = 0.9578, r^2 = 0.7357

3. r = -0.9313, r^2 = 0.8674

5. r = 0.6605, r^2 = 0.4362

7. r = 0.8871, r^2 = 0.7869

9. r = 0.7036, r^2 = 0.4951

11. No, see solutions.

13. No, see solutions.

This page titled 10.2: Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Inference for Regression and Correlation

How do you really say you have a correlation? Can you test to see if there really is a correlation? Of course, the answer is yes. The hypothesis test for correlation is as follows:

Hypothesis Test for Correlation:

1. State the random variables in words.
   x = independent variable
   y = dependent variable
2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$ (There is no correlation)
   $H_A : \rho \neq 0$ (There is a correlation)
   or
   $H_A : \rho < 0$ (There is a negative correlation)
   or
   $H_A : \rho > 0$ (There is a postive correlation)
   Also, state your α level here.
3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test are the same assumptions for regression and correlation.
4. Find the test statistic and p-value
   $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$
   with degrees of freedom = $df = n - 2$
   p-value: Using the TI-83/84: tcdf(lower limit, upper limit, *df*)

   Note

   If $H_A : \rho < 0$, then lower limit is -1E99 and upper limit is your test statistic. If $H_A : \rho > 0$, then lower limit is your test statistic and the upper limit is 1E99. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.

   Using R: pt(*t*, *df*)

   Note

   If $H_A : \rho < 0$, then use pt(*t*, *df*), If $H_A : \rho > 0$, then use $1 - \text{pt}(t, df)$. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.

5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. Interpretation
   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Note

The TI-83/84 calculator results give you the test statistic and the p-value. In R, the command for getting the test statistic and p-value is cor.test(independent variable, dependent variable, alternative = "less" or "greater"). Use less for $H_A : \rho < 0$, use greater for $H_A : \rho > 0$, and leave off this command for $H_A : \rho \neq 0$.

Example 10.3.1 Testing the claim of a linear correlation

Is there a positive correlation between beer's alcohol content and calories? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data is in Example 10.3.1. Test at the 5% level.

Solution

1. State the random variables in words.
   x = alcohol content in the beer
   y = calories in 12 ounce beer
2. State the null and alternative hypotheses and the level of significance.
   Since you are asked if there is a positive correlation, $\rho > 0$.
   $H_o : \rho = 0$
   $H_A : \rho > 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test.
   The assumptions for the hypothesis test were already checked in *Example* 10.3.2.
4. Find the test statistic and p-value.
   The results from the TI-83/84 calculator are in *Figure* 10.3.1.

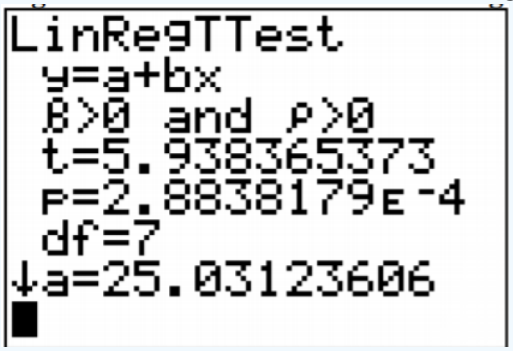


Figure 10.3.1*: Results for Linear Regression Test on TI-83/84*

Test statistic: $t \approx 5.938$ and p-value: $p \approx 2.884 \times 10^{-4}$
The results from R are
cor.test(alcohol, calories, alternative = "greater")
Pearson's product-moment correlation
data: alcohol and calories
t = 5.9384, df = 7, p-value = 0.0002884
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.7046161 1.0000000
sample estimates:
cor
0.9134414
Test statistic: $t \approx 5.9384$ and p-value: $p \approx 0.0002884$

5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a positive correlation between alcohol content and number of calories in a 12-ounce bottle of beer.

Prediction Interval

Using the regression equation you can predict the number of calories from the alcohol content. However, you only find one value. The problem is that beers vary a bit in calories even if they have the same alcohol content. It would be nice to have a range instead of a single value. The range is called a prediction interval. To find this, you need to figure out how much error is in the estimate from the regression equation. This is known as the **standard error of the estimate**.

Definition

Standard Error of the Estimate

This is the sum of squares of the residuals

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

This formula is hard to work with, so there is an easier formula. You can also find the value from technology, such as the calculator.

$$s_e = \sqrt{\frac{SS_y - b^* SS_{xy}}{n - 2}}$$

Example 10.3.2 finding the standard error of the estimate

Find the standard error of the estimate for the beer data. To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.3.1.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

Using the TI-83/84, the results are in *Figure* 10.3.2.

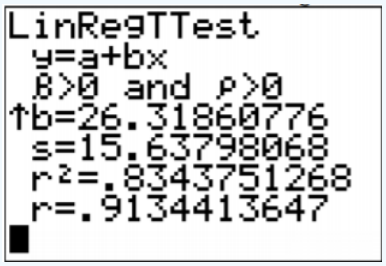


Figure 10.3.2: Results for Linear Regression Test on TI-83/84

The *s* in the results is the standard error of the estimate. So $s_e \approx 15.64$.

To find the standard error of the estimate in R, the commands are
lm.out = lm(dependent variable ~ independent variable) – this defines the linear model with a name so you can use it later. Then
summary(lm.out) – this will produce most of the information you need for a regression and correlation analysis. In fact, the only thing R doesn't produce with this command is the correlation coefficient. Otherwise, you can use the command to find the regression equation, coefficient of determination, test statistic, p-value for a two-tailed test, and standard error of the estimate.

The results from R are
lm.out=lm(calories~alcohol)
summary(lm.out)

Call:
lm(formula = calories ~ alcohol)

Residuals:

Min	1Q	Median	3Q	Max
−30.253	−1.624	2.744	9.271	14.271

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.031	24.999	1.001	0.350038
alcohol	26.319	4.432	5.938	0.000577

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.64 on 7 degrees of freedom
Multiple R-squared: 0.8344, Adjusted R-squared: 0.8107
F-statistic: 35.26 on 1 and 7 DF, p-value: 0.0005768

From this output, you can find the y-intercept is 25.031, the slope is 26.319, the test statistic is t = 5.938, the p-value for the two-tailed test is 0.000577. If you want the p-value for a one-tailed test, divide this number by 2. The standard error of the estimate is the residual standard error and is 15.64. There is some information in this output that you do not need.

If you want to know how to calculate the standard error of the estimate from the formula, refer to Example 10.3.3.

Example 10.3.3 finding the standard error of the estimate from the formula

Find the standard error of the estimate for the beer data using the formula. To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.3.1.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

From Example 10.3.3 :

$$SS_y = \sum (y - \overline{y})^2 = 10335.56$$
$$SS_{xy} = \sum (x - \overline{x})(y - \overline{y}) = 327.6666$$
$$n = 9$$
$$b = 26.3$$

The standard error of the estimate is

$$s_e = \sqrt{\frac{SS_y - b^* SS_{xy}}{n-2}}$$
$$= \sqrt{\frac{10335.56 - 26.3(327.6666)}{9-2}}$$
$$= 15.67$$

Prediction Interval for an Individual *y*

Given the fixed value x_0, the prediction interval for an individual *y* is

$$\hat{y} - E < y < \hat{y} + E$$

where

$$\hat{y} = a + bx$$
$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \overline{x})^2}{SS_x}}$$
$$df = n - 2$$

Note

To find $SS_x = \sum (x - \overline{x})^2$ remember, the standard derivation formula from chapter 3 $s_x = \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}}$

So, $s_x = \sqrt{\frac{SS_x}{n-1}}$

Now solve for SS_x

$SS_x = s_x^2 (n-1)$

You can get the standard deviation from technology.
R will produce the prediction interval for you. The commands are (Note you probably already did the lm.out command. You do not

need to do it again.)
lm.out = lm(dependent variable ~ independent variable) – calculates the linear model
predict(lm.out, newdata=list(independent variable = value), interval="prediction", level=C) – will compute a prediction interval for the independent variable set to a particular value (put that value in place of the word value), at a particular C level (given as a decimal)

Example 10.3.4 find the prediction interval

Find a 95% prediction interval for the number of calories when the alcohol content is 6.5% using a random sample taken of beer's alcohol content and calories ("Calories in beer," 2011). The data are in Example 10.3.1.

Solution

x = alcohol content in the beer

y = calories in 12 ounce beer

Computing the prediction interval using the TI-83/84 calculator:
From Example 10.3.2

$$\hat{y} = 25.0 + 26.3x$$
$$x_o = 6.50$$
$$\hat{y} = 25.0 + 26.3(6.50) = 196 \text{ calories}$$

From Example #10.3.2

$$s_e \approx 15.64$$

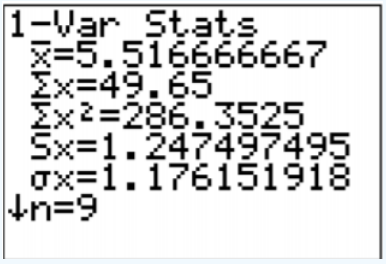


Figure 10.3.3: Results of 1-Var Stats on TI-83/84

$$\overline{x} = 5.517$$
$$s_x = 1.247497495$$
$$n = 9$$

Now you can find

$$\begin{aligned} SS_x &= s_x^2(n-1) \\ &= (1.247497495)^2(9-1) \\ &= 12.45 \\ df &= n - 2 = 9 - 2 = 7 \end{aligned}$$

Now look in table A.2. Go down the first column to 7, then over to the column headed with 95%.

$$t_c = 2.365$$

$$\begin{aligned} E &= t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \overline{x})^2}{SS_x}} \\ &= 2.365(15.64)\sqrt{1 + \frac{1}{9} + \frac{(6.50 - 5.517)^2}{12.45}} \\ &= 40.3 \end{aligned}$$

Prediction interval is

$$\hat{y} - E < y < \hat{y} + E$$
$$196 - 40.3 < y < 196 + 40.3$$
$$155.7 < y < 236.3$$

Computing the prediction interval using R:

```
predict(lm.out, newdata=list(alcohol=6.5), interval = "prediction", level=0.95)
       fit      lwr      upr
1 196.1022 155.7847 236.4196
```

fit = $\hat{y}$ when x = 6.5%. lwr = lower limit of prediction interval. upr = upper limit of prediction interval. So the prediction interval is $155.8 < y < 236.4$.

Statistical interpretation: There is a 95% chance that the interval $155.8 < y < 236.4$ contains the true value for the calories when the alcohol content is 6.5%.

Real world interpretation: If a beer has an alcohol content of 6.50% then it has between 156 and 236 calories.

Example 10.3.5 Doing a correlation and regression analysis using the ti-83/84

Example 10.3.1 contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

Table 10.3.1: Temperatures and Elevation of Cities on a Given Day

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

a. State the random variables.
b. Find a regression equation for elevation and high temperature on a given day.
c. Find the residuals and create a residual plot.
d. Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.
e. Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.
f. Between the answers to parts d and e, which estimate is probably more accurate and why?
g. Find the correlation coefficient and coefficient of determination and interpret both.
h. Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.
i. Find the standard error of the estimate.
j. Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution

a. x = elevation

y = high temperature

b.

a. A random sample was taken as stated in the problem.
b. The distribution for each high temperature value is normally distributed for every value of elevation.
   i. Look at the scatter plot of high temperature versus elevation.

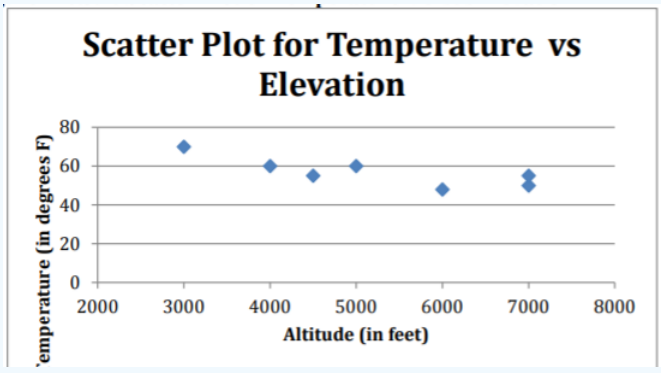


Figure 10.3.4*: Scatter Plot of Temperature Versus Elevation*

The scatter plot looks fairly linear.

ii. There are no points that appear to be outliers.

iii. The residual plot for temperature versus elevation appears to be fairly random. (See *Figure* 10.3.7.) It appears that the high temperature is normally distributed.

All calculations computed using the TI-83/84 calculator.

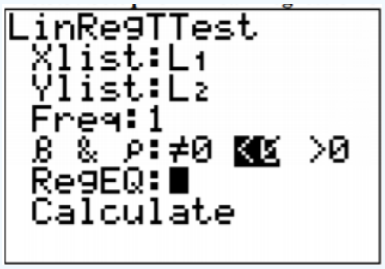


Figure 10.3.5: Setup for Linear Regression on TI-83/84 Calculator

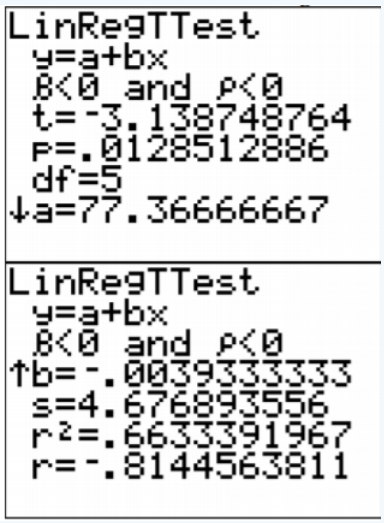


Figure 10.3.6: Results for Linear Regression on TI-83/84 Calculator

$$\hat{y} = 77.4 - 0.0039x$$

c.

Table 10.3.2: Residuals for Elevation vs. Temperature Data

x	y	$\hat{y}$	$y - \hat{y}$
7000	50	50.1	-0.1
4000	60	61.8	-1.8
6000	48	54.0	-6.0
3000	70	65.7	4.3
7000	55	50.1	4.9
4500	55	59.85	-4.85
5000	60	57.9	2.1

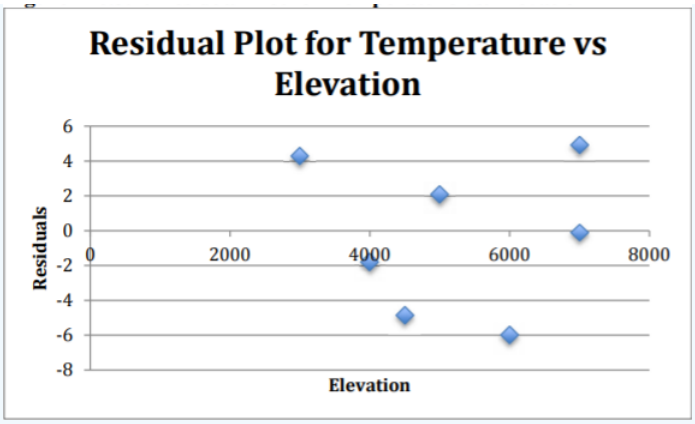


Figure 10.3.7: Residual Plot for Temperature vs. Election

The residuals appear to be fairly random.

d.

$x_o = 5500$
$\hat{y} = 77.4 - 0.0039(5500) = 55.95^\circ F$

e.

$x_o = 8000$
$\hat{y} = 77.4 - 0.0039(8000) = 46.2^\circ F$

f. Part d is more accurate, since it is interpolation and part e is extrapolation.

g. From *Figure* 10.3.6, the correlation coefficient is $r \approx$ -0.814, which is moderate to strong negative correlation.

From *Figure* 10.3.6, the coefficient of determination is $r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.

1. State the random variables in words.
   x = elevation
   y = high temperature
2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test The assumptions for the hypothesis test were already checked part b.
4. Find the test statistic and p-value
   From *Figure* 10.3.6,
   Test statistic:
   $t \approx -3.139$
   p-value:
   $p \approx 0.0129$
5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i. From *Figure* 10.3.6,

$s_e \approx 4.677$

j. $\hat{y} = 77.4 - 0.0039(6500) \approx 52.1^\circ F$

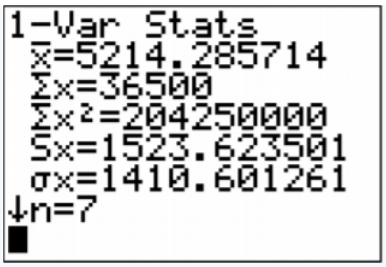


Figure 10.3.8: Results of 1-Var Stats on TI-83/84

$\overline{x} = 5214.29$
$s_x = 1523.624$
$n = 7$

Now you can find

$$\begin{aligned} SS_x &= s_x^2(n-1) \\ &= (1523.623501)^2(7-1) \\ &= 13928571.43 \end{aligned}$$
$df = n - 2 = 7 - 2 = 5$

Now look in table A.2. Go down the first column to 5, then over to the column headed with 95%.

$t_c = 2.571$

So

$$\begin{aligned} E &= t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \overline{x})^2}{SS_x}} \\ &= 2.571(4.677)\sqrt{1 + \frac{1}{7} + \frac{(6500 - 5214.29)^2}{13928571.43}} \\ &= 13.5 \end{aligned}$$

Prediction interval is

$$\begin{aligned} &\hat{y} - E < y < \hat{y} + E \\ &52.1 - 13.5 < y < 52.1 + 13.5 \\ &38.6 < y < 65.6 \end{aligned}$$

Statistical interpretation: There is a 95% chance that the interval $38.6 < y < 65.6$ contains the true value for the temperature at an elevation of 6500 feet.

Real world interpretation: A city of 6500 feet will have a high temperature between 38.6°F and 65.6°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Example 10.3.6 doing a correlation and regression analysis using r

Example 10.3.1 contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

a. State the random variables.
b. Find a regression equation for elevation and high temperature on a given day.
c. Find the residuals and create a residual plot.
d. Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.
e. Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.
f. Between the answers to parts d and e, which estimate is probably more accurate and why?
g. Find the correlation coefficient and coefficient of determination and interpret both.
h. Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.
i. Find the standard error of the estimate.
j. Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution

a. x = elevation

y = high temperature

b.

a. A random sample was taken as stated in the problem.
b. The distribution for each high temperature value is normally distributed for every value of elevation.
   i. Look at the scatter plot of high temperature versus elevation.
   R command: plot(elevation, temperature, main="Scatter Plot for Temperature vs Elevation", xlab="Elevation (feet)", ylab="Temperature (degrees F)", ylim=c(0,80))

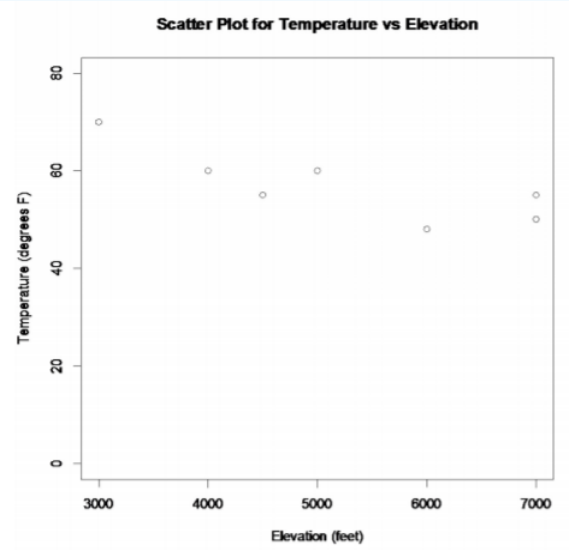


Figure 10.3.9*: Scatter Plot of Temperature Versus Elevation*

The scatter plot looks fairly linear.

   ii. The residual plot for temperature versus elevation appears to be fairly random. (See *Figure 10.3.10*.)
   It appears that the high temperature is normally distributed.

Using R:
Commands:
lm.out=lm(temperature ~ elevation)
summary(lm.out)

Output:
Call:
lm(formula = temperature ~ elevation)

Residuals:

1	2	3	4	5	6	7
0.1667	−1.6333	−5.7667	4.4333	5.1667	−4.6667	2.3000

Coefficients:

	Estimate Std.	Error	t value	Pr(>|t|)
(Intercept)	77.36667	6.769182	11.429	$8.98e-05$ **
elevation	−0.003933	0.001253	−3.139	0.0257*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.677 on 5 degrees of freedom
Multiple R-squared: 0.6633, Adjusted R-squared: 0.596
F-statistic: 9.852 on 1 and 5 DF, p-value: 0.0257

From the output you can see the slope = -0.0039 and the y-intercept = 77.4. So the regression equation is:

$\hat{y} = 77.4 - 0.0039x$

c. R command: (notice these are also in the summary(lm.out) output, but if you have too many data points, then R only gives a numerical summary of the residuals.)

residuals(lm.out)

Illegal pream-token (C)

So for the first x of 7000, the residual is approximately 0.1667. This means if you find the $\hat{y}$ for when x is 7000 and then subtract this answer from the y value of 50 that was measured, you would obtain 0.1667. Similar process is computed for the other residual values.

To plot the residuals, the R command is

plot(elevation, residuals(lm.out), main="Residual Plot for Temperautre vs Elevation", xlab="Elevation (feet)", ylab="Residuals")
abline(0,0)

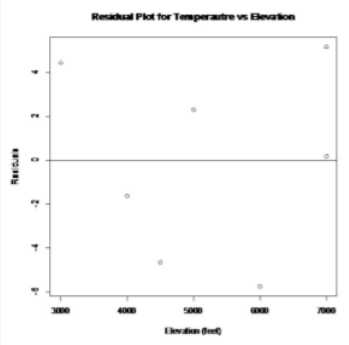


Figure 10.3.10: Residual Plot for Temperature vs. Elevation

The residuals appear to be fairly random.

d. $x_o = 5500$
$\hat{y} = 77.4 - 0.0039(5500) = 55.95^\circ F$

e. $x_o = 8000$
$\hat{y} = 77.4 - 0.0039(8000) = 46.2^\circ F$

f. Part d is more accurate, since it is interpolation and part e is extrapolation.

g. The R command for the correlation coefficient is
cor(elevation, temperature)
[1] -0.8144564

So, $r \approx -0.814$, which is moderate to strong negative correlation.

From summary(lm.out), the coefficient of determination is the Multiple R-squared.

So $r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.

1. State the random variables in words.
   x = elevation
   y = high temperature
2. . State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test.
   The assumptions for the hypothesis test were already checked part b.

4. Find the test statistic and p-value
   The R command is cor.test(elevation, temperature, alternative = "less")
   Pearson's product-moment correlation
   data: elevation and temperature
   t = -3.1387, df = 5, p-value = 0.01285
   alternative hypothesis: true correlation is less than 0
   95 percent confidence interval:
   -1.0000000 -0.3074247
   sample estimates:
   cor
   -0.8144564
   Test statistic: $t \approx -3.1387$ and p-value: $p \approx 0.01285$
5. Conclusion
   Reject H_o since the p-value is less than 0.05.
6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i. From summary(lm.out), Residual standard error: 4.677.

So, $s_e \approx 4.677$

j. R command is predict(lm.out, newdata=list(elevation = 6500), interval = "prediction", level=0.95)

	fit	lwr	upr
1	51.8	38.29672	65.30328

So when x = 6500 feet, $\hat{y} = 51.8^{\circ}F$ and $38.29672 < y < 65.30328$.

Statistical interpretation: There is a 95% chance that the interval $38.3 < y < 65.3$ contains the true value for the temperature at an elevation of 6500 feet.

Real world interpretation: A city of 6500 feet will have a high temperature between 38.3°F and 65.3°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Homework

Exercise 10.3.1

For each problem, state the random variables. The data sets in this section are in the homework for section 10.1 and were also used in section 10.2. If you removed any data points as outliers in the other sections, remove them in this sections homework too.

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone one (in cm) were collected and are in Example 10.3.5 ("Prediction of height," 2013).
   a. Test at the 1% level for a positive correlation between length of metacarpal bone one and height of a person.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for height of a person with a metacarpal length of 44 cm.
2. Example 10.3.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013).
   a. Test at the 5% level for a positive correlation between house value and rental amount.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the rental income on a house worth $230,000.
3. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example 10.3.7.
   a. Test at the 1% level for a negative correlation between fertility rate and life expectancy.

b. Find the standard error of the estimate.
c. Compute a 99% prediction interval for the life expectancy for a country that has a fertility rate of 2.7.

4. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in Example 10.3.8.
   a. Test at the 5% level for a correlation between percentage spent on health expenditure and the percentage of women receiving prenatal care.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the percentage of woman receiving prenatal care for a country that spends 5.0 % of GDP on health expenditure.
5. The height and weight of baseball players are in Example 10.3.9 ("MLB heightsweights," 2013).
   a. Test at the 5% level for a positive correlation between height and weight of baseball players.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the weight of a baseball player that is 75 inches tall.
6. Different species have different body weights and brain weights are in Example 10.3.10. ("Brain2bodyweight," 2013).
   a. Test at the 1% level for a positive correlation between body weights and brain weights.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for the brain weight for a species that has a body weight of 62 kg.
7. A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in Example 10.3.11.
   a. Test at the 5% level for a correlation between amount of calories and amount of sodium.
   b. Find the standard error of the estimate.
   c. Compute a 95% prediction interval for the amount of sodium a beef hotdog has if it is 170 calories.
8. Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in Example 10.3.12 ("OECD economic development," 2013).
   a. Test at the 5% level for a negative correlation between percent of labor force in agriculture and per capita income.
   b. Find the standard error of the estimate.
   c. Compute a 90% prediction interval for the per capita income in a country that has 21 percent of labor in agriculture.
9. Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in Example 10.3.13 ("Smoking and cancer," 2013).
   a. Test at the 1% level for a positive correlation between cigarette smoking and deaths of bladder cancer.
   b. Find the standard error of the estimate.
   c. Compute a 99% prediction interval for the number of deaths from bladder cancer when the cigarette sales were 20 per capita.
10. The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in Example 10.3.14 ("Passenger car mileage," 2013).
    a. Test at the 5% level for a negative correlation between the weight of cars and mileage.
    b. Find the standard error of the estimate.
    c. Compute a 95% prediction interval for the mileage on a car that weighs 3800 pounds.

Answer

For hypothesis test just the conclusion is given. See solutions for entire answer.

1. a. Reject Ho, b. $s_e \approx 4.559$, c. $151.3161\text{cm} < y < 187.3859\text{cm}$
3. a. Reject Ho, b. $s_e \approx 3.204$, c. $62.945 \text{ years } < y < 81.391\text{years}$
5. a. Reject Ho, b. $s_e \approx 15.33$, c. $176.02 \text{ inches } < y < 240.92\text{inches}$
7. a. Reject Ho, b. $s_e \approx 48.58$, c. $348.46\text{mg} < y < 559.38\text{mg}$
9. a. Reject Ho, b. $s_e \approx 0.6838$, c. $1.613 \text{ hundred thousand } < y < 5.432 \text{ hundred thousand}$

Data Source:

Brain2bodyweight. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...ain2BodyWeight

Calories in beer, beer alcohol, beer carbohydrates. (2011, October 25). Retrieved from www.beer100.com/beercalories.htm

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/rentcap.html

Data hotdogs. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...D_Data_HotDogs

Fertility rate. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Health expenditure. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

Life expectancy at birth. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

MLB heightsweights. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index...HeightsWeights

OECD economic development. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/oecdat.html

Passenger car mileage. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

Prediction of height from metacarpal bone length. (2013, September 26). Retrieved from http://www.statsci.org/data/general/stature.html

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

Smoking and cancer. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafil...cancerdat.html

This page titled 10.3: Inference for Regression and Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

11: Chi-Square and ANOVA Tests

This chapter presents material on three more hypothesis tests. One is used to determine significant relationship between two qualitative variables, the second is used to determine if the sample data has a particular distribution, and the last is used to determine significant relationships between means of 3 or more samples.

11.1: Chi-Square Test for Independence

11.2: Chi-Square Goodness of Fit

11.3: Analysis of Variance (ANOVA)

This page titled 11: Chi-Square and ANOVA Tests is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

11.1: Chi-Square Test for Independence

Remember, qualitative data is where you collect data on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. An example is that there is a theory that there is a relationship between breastfeeding and autism. To determine if there is a relationship, researchers could collect the time period that a mother breastfed her child and if that child was diagnosed with autism. Then you would have a table containing this information. Now you want to know if each cell is independent of each other cell. Remember, independence says that one event does not affect another event. Here it means that having autism is independent of being breastfed. What you really want is to see if they are not independent. In other words, does one affect the other? If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the **Chi-Square Test for Independence**. Technically it should be called the Chi-Square Test for Dependence, but for historical reasons it is known as the test for independence. Just as with previous hypothesis tests, all the steps are the same except for the assumptions and the test statistic.

Hypothesis Test for Chi-Square Test

1. State the null and alternative hypotheses and the level of significance
   H_o: the two variables are independent (this means that the one variable is not affected by the other)
   H_A: the two variables are dependent (this means that the one variable is affected by the other)
   Also, state your α level here.
2. State and check the assumptions for the hypothesis test
   a. A random sample is taken.
   b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later, and this assumption means $E \geq 5$).
3. Find the test statistic and p-value
   Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is *O*. Each table is made up of rows and columns. Then each row is totaled to give a row total and each column is totaled to give a column total.
   The null hypothesis is that the variables are independent. Using the multiplication rule for independent events you can calculate the probability of being one value of the first variable, *A*, and one value of the second variable, *B* (the probability of a particular cell $P(A \text{ and } B)$). Remember in a hypothesis test, you assume that H_o is true, the two variables are assumed to be independent.

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B) \text{ if } A \text{ and } Bare independent \\ &= \frac{\text{number of ways } A \text{ can happen}}{\text{total number of individuals}} \cdot \frac{\text{number of ways } B \text{ can happen}}{\text{total number of individuals}} \\ &= \frac{\text{row total}}{n} * \frac{\text{column total}}{n} \end{aligned}$$

Now you want to 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.

$$\begin{aligned} \text{Expected frequency } (\text{cell } A \text{ and } B) &= E(A \text{ and } B) \\ &= n\left(\frac{\text{row total}}{n} \cdot \frac{\text{column total}}{n}\right) \\ &= \frac{\text{row total -column total}}{n} \end{aligned}$$

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 do you find the test statistic? First find the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just

because the frequencies are large, you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This is the test statistic.

Test Statistic:

The symbol for Chi-Square is χ^2

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* is the observed frequency and *E* is the expected frequency

Distribution of Chi-Square

χ^2 has different curves depending on the degrees of freedom. It is skewed to the right for small degrees of freedom and gets more symmetric as the degrees of freedom increases (see *Figure* 11.1.1). Since the test statistic involves squaring the differences, the test statistics are all positive. A chi-squared test for independence is always right tailed.

Figure 11.1.1: Chi-Square Distribution

p-value:

Using the TI-83/84: χ cdf (lower limit, 1E99, df)

Using R: $1 -$ pchisq (x^2, df)

Where the degrees of freedom is $df = (\#$ of rows $-1) * (\#$ of columns $-1)$

4. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.

5. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Example 11.1.1 hypothesis test with chi-square test using formula

Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what time period they breastfed their children. The data is in table #11.1.1 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that breastfeeding and autism are independent? Test at the1% level.

Table 11.1.1: Autism Versus Breastfeeding

Autis261m	Breast Feeding Timelines				Row Total
	None	Less than 2 months	2 to 6 months	More than 6 months	
Yes	241	198	164	215	818
No	20	25	27	44	116
Column Total	261	223	191	259	934

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: Breastfeeding and autism are independent

H_A: Breastfeeding and autism are dependent

α = 0.01

2. State and check the assumptions for the hypothesis test

a. A random sample of breastfeeding time frames and autism incidence was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. All expected frequencies are more than 5.

3. Find the test statistic and p-value

Test statistic:

First find the expected frequencies for each cell

$$E(\text{ Autism and no breastfeeding }) = \frac{818^*261}{934} \approx 228.585$$

$$E(\text{ Autism and } < 2 \text{ months }) = \frac{818^*223}{934} \approx 195.304$$

$$E(\text{ Autism and 2 to 6 months }) = \frac{818^*191}{934} \approx 167.278$$

$$E(\text{ Autism and more than 6 months }) = \frac{818 * 259}{934} \approx 226.833$$

Others are done similarly. It is easier to do the calculations for the test statistic with a table, the others are in table #11.1.2 along with the calculation for the test statistic. (Note: the column of *O-E* should add to 0 or close to 0.)

Table 11.1.2: Calculations for Chi-Square Test Statistic

O	*E*	*O-E*	$(O-E)^2$	$(O-E)^2/E$
241	228.585	12.415	154.132225	0.674288448
198	195.304	2.696	7.268416	0.03721591
164	167.278	-3.278	10.745284	0.064236086
215	226.833	-11.833	140.019889	0.617281828
20	32.4154	-12.4154	154.1421572	4.755213792
25	27.6959	-2.6959	7.26787681	0.262417066
27	23.7216	3.2784	10.74790656	0.453085229
44	32.167	11.833	140.019889	4.352904809
Total		0.0001		11.2166432 = χ^2

The test statistic formula is $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is the total of the last column in Example 11.1.2.

p-value:

$df = (2-1)^*(4-1) = 3$

Using TI-83/84: $\chi\,\text{cdf}(11.2166432, 1\text{E}99, 3) \approx 0.01061$

Using R: $1 - \text{pchisq}(11.2166432, 3) \approx 0.01061566$

4. Conclusion

Fail to reject H_o since the p-value is more than 0.01.

5. Interpretation

There is not enough evidence to show that breastfeeding and autism are dependent. This means that you cannot say that the whether a child is breastfed or not will indicate if that the child will be diagnosed with autism.

Example 11.1.2 hypothesis test with chi-square test using technology

Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what time period they breastfed their children. The data is in Example 11.1.1 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that breastfeeding and autism are independent? Test at the1% level.

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: Breastfeeding and autism are independent

H_A: Breastfeeding and autism are dependent

$\alpha = 0.01$

2. State and check the assumptions for the hypothesis test

a. A random sample of breastfeeding time frames and autism incidence was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. All expected frequencies are more than 5.

3. Find the test statistic and p-value

Test statistic:

To use the TI-83/84 calculator to compute the test statistic, you must first put the data into the calculator. However, this process is different than for other hypothesis tests. You need to put the data in as a matrix instead of in the list. Go into the MATRX menu then move over to EDIT and choose 1:[A]. This will allow you to type the table into the calculator. *Figure* 11.1.2 shows what you will see on your calculator when you choose 1:[A] from the EDIT menu.

Figure 11.1.2: Matrix Edit Menu on TI-83/84

The table has 2 rows and 4 columns (don't include the row total column and the column total row in your count). You need to tell the calculator that you have a 2 by 4. The 1 X1 (you might have another size in your matrix, but it doesn't matter because you will change it) on the calculator is the size of the matrix. So type 2 ENTER and 4 ENTER and the calculator will make a matrix of the correct size. See *Figure* 11.1.3.

Figure 11.1.3: Matrix Setup for Table

Now type the table in by pressing ENTER after each cell value. *Figure* 11.1.4 contains the complete table typed in. Once you have the data in, press QUIT.

Figure 11.1.4: Data Typed into Matrix

To run the test on the calculator, go into STAT, then move over to TEST and choose χ^2-Test from the list. The setup for the test is in *Figure* 11.1.5.

Figure 11.1.5: Setup for Chi-Square Test on TI-83/84

Once you press ENTER on Calculate you will see the results in *Figure* 11.1.6.

Figure 11.1.6: Results for Chi-Square Test on TI-83/84

The test statistic is $\chi^2 \approx 11.2167$ and the p-value is $p \approx 0.01061$. Notice that the calculator calculates the expected values for you and places them in matrix B. To eview the expected values, go into MATRX and choose 2:[B]. *Figure* 11.1.7 shows the output. Press the right arrows to see the entire matrix.

Figure 11.1.7: Expected Frequency for Chi-Square Test on TI-83/84

To compute the test statistic and p-value with R,
row1 = c(data from row 1 separated by commas)
row2 = c(data from row 2 separated by commas)
keep going until you have all of your rows typed in.
data.table = rbind(row1, row2, …) – makes the data into a table. You can call it what ever you want. It does not have to be

data.table.
data.table – use if you want to look at the table
chisq.test(data.table) – calculates the chi-squared test for independence
chisq.test(data.table)$expected – let's you see the expected values

For this example, the commands would be
row1 = c(241, 198, 164, 215)
row2 = c(20, 25, 27, 44)
data.table = rbind(row1, row2)
data.table

Output:
[,1] [,2] [,3] [,4]
row1 241 198 164 215
row2 20 25 27 44

chisq.test(data.table)

Output:
Pearson's Chi-squared test

data: data.table
X-squared = 11.217, df = 3, p-value = 0.01061

chisq.test(data.table)$expected

Output: [,1] [,2] [,3] [,4]
row1 228.58458 195.30407 167.27837 226.83298
row2 32.41542 27.69593 23.72163 32.16702

The test statistic is $\chi^2 \approx 11.217$ and the p-value is $p \approx 0.01061$.

4. Conclusion

Fail to reject H_o since the p-value is more than 0.01.

5. Interpretation

There is not enough evidence to show that breastfeeding and autism are dependent. This means that you cannot say that the whether a child is breastfed or not will indicate if that the child will be diagnosed with autism.

Example 11.1.3 hypothesis test with chi-square test using formula

The World Health Organization (WHO) keeps track of how many incidents of leprosy there are in the world. Using the WHO regions and the World Banks income groups, one can ask if an income level and a WHO region are dependent on each other in terms of predicting where the disease is. Data on leprosy cases in different countries was collected for the year 2011 and a summary is presented in *Table* 11.1.3 ("Leprosy: Number of," 2013). Is there evidence to show that income level and WHO region are independent when dealing with the disease of leprosy? Test at the 5% level.

Table 11.1.3: Number of Leprosy Cases

WHO Region	World Bank Income Group				Row Total
	High Income	Upper Middle Income	Lower Middle Income	Low Income	
Americas	174	36028	615	0	36817
Eastern Mediterranean	54	6	1883	604	2547
Europe	10	0	0	0	10
Western Pacific	26	216	3689	1155	5086

Africa	0	39	1986	15928	17953
South-East Asia	0	0	149896	10236	160132
Column Total	264	36289	158069	27923	222545

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: WHO region and Income Level when dealing with the disease of leprosy are independent

H_A: WHO region and Income Level when dealing with the disease of leprosy are dependent

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample of incidence of leprosy was taken from different countries and the income level and WHO region was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. There are actually 4 expected frequencies that are less than 5, and the results of the test may not be valid. If you look at the expected frequencies you will notice that they are all in Europe. This is because Europe didn't have many cases in 2011.

3. Find the test statistic and p-value

Test statistic:

First find the expected frequencies for each cell.

$$E(\text{ Americas and High Income }) = \frac{36817 * 264}{222545} \approx 43.675$$

$$E(\text{ Americas and Upper Middle Income }) = \frac{36817 * 36289}{222545} \approx 6003.514$$

$$E(\text{ Americas and Lower Middle Income}) = \frac{36817 * 158069}{222545} \approx 26150.335$$

$$E(\text{ Americas and Lower Income }) = \frac{36817 * 27923}{222545} \approx 4619.475$$

Others are done similarly. It is easier to do the calculations for the test statistic with a table, and the others are in Example 11.1.4 along with the calculation for the test statistic.

Table 11.1.4: Calculations for Chi-Square Test Statistic

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
174	43.675	130.325	16984.564	388.8838719
54	3.021	50.979	2598.813	860.1218328
10	0.012	9.988	99.763	8409.746711
26	6.033	19.967	398.665	66.07628214
0	21.297	-21.297	453.572	21.29722977
0	189.961	-189.961	36085.143	189.9608978
36028	6003.514	30024.486	901469735.315	150157.0038
6	415.323	-409.323	167545.414	403.4097962
0	1.631	-1.631	2.659	1.6306365
216	829.342	-613.342	376188.071	453.5983897
39	2927.482	-2888.482	8343326.585	2850.001268
0	26111.708	-26111.708	681821316.065	26111.70841

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
615	26150.335	-25535.335	652053349.724	24934.7988
1883	1809.080	73.290	5464.144	3.020398811
0	7.103	-7.103	50.450	7.1027882
3689	3612.478	76.522	5855.604	1.620938405
1986	12751.636	-10765.636	115898911.071	9088.944681
149896	113738.368	36157.632	1307374351.380	11494.57632
0	4619.475	-4619.475	21339550.402	4619.475122
604	319.575	284.425	80897.421	253.1404187
0	1.255	-1.255	1.574	1.25471253
1155	638.147	516.853	267137.238	418.6140882
15928	2252.585	13675.415	187016964.340	83023.25138
10236	20091.963	-9855.963	97140000.472	4834.769106
Total		0.000		328594.008 = χ^2

The test statistic formula is $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is the total of the last column in Example 11.1.2.

p-value:

$df = (6-1) * (4-1) = 15$

Using the TI-83/84: χ cdf$(328594.008, 1E99, 15) \approx 0$

Using R: $1 -$ pchisq $(328594.008, 15) \approx 0$

4. Conclusion

Reject H_o since the p-value is less than 0.05.

5. Interpretation

There is enough evidence to show that WHO region and income level are dependent when dealing with the disease of leprosy. WHO can decide how to focus their efforts based on region and income level. Do remember though that the results may not be valid due to the expected frequencies not all be more than 5.

Example 11.1.4 hypothesis test with chi-square test using technology

The World Health Organization (WHO) keeps track of how many incidents of leprosy there are in the world. Using the WHO regions and the World Banks income groups, one can ask if an income level and a WHO region are dependent on each other in terms of predicting where the disease is. Data on leprosy cases in different countries was collected for the year 2011 and a summary is presented in *Table* 11.1.3 ("Leprosy: Number of," 2013). Is there evidence to show that income level and WHO region are independent when dealing with the disease of leprosy? Test at the 5% level.

Solution

1. State the null and alternative hypotheses and the level of significance
H_o: WHO region and Income Level when dealing with the disease of leprosy are independent

H_A: WHO region and Income Level when dealing with the disease of leprosy are dependent

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample of incidence of leprosy was taken from different countries and the income level and WHO region was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. There are actually 4 expected frequencies that are less than 5, and the results of the test may not be valid. If you look at the expected frequencies you will notice that they are all in Europe. This is because Europe didn't have many cases in 2011.

3. Find the test statistic and p-value
Test statistic:
Using the TI-83/84. See Example 11.1.2 for the process of doing the test on the calculator. Remember, you need to put the data in as a matrix instead of in the list.

Figure 11.1.8: Setup for Matrix on TI-83/84

Figure 11.1.9: Results for Chi-Square Test on TI-83/84

$$\chi^2 \approx 328594.0079$$

Figure 11.1.10: Expected Frequency for Chi-Square Test on TI-83/84

Press the right arrow to look at the other expected frequencies.

p-value:

$p - \text{value} \approx 0$

Using R:
row1=c(174, 36028, 615, 0)
row2=c(54, 6, 1883, 604)
row3=c(10, 0, 0, 0)
row4=c(26, 216, 3689, 1155)
row5=c(0, 39, 1986, 15928)
row6=c(0, 0, 149896, 10236)
chisq.test(data.table)

Pearson's Chi-squared test

data: data.table
X-squared = 328590, df = 15, p-value < 2.2e-16

Warning message:
In chisq.test(data.table) : Chi-squared approximation may be incorrect

chisq.test(data.table)$expected

	[, 1]	[, 2]	[, 3]	[, 4]
row1	43.67515783	6003.514404	$2.615034e+04$	4619.475122
row2	3.02144735	415.323117	$1.809080e+03$	319.575281
row3	0.01186277	1.630637	$7.102788e+00$	1.254713
row4	6.03340448	829.341724	$3.612478e+03$	638.146793
row5	21.29722977	2927.481709	$1.275164e+04$	2252.585405
row6	189.96089780	26111.708410	$1.137384e+05$	20091.962686

Warning message:
In chisq.test(data.table) : Chi-squared approximation may be incorrect

$\chi^2 = 328590$ and p-value = 2.2×10^{-16}

4. Conclusion

Reject H_o since the p-value is less than 0.05.

5. Interpretation

There is enough evidence to show that WHO region and income level are dependent when dealing with the disease of leprosy. WHO can decide how to focus their efforts based on region and income level. Do remember though that the results may not be

valid due to the expected frequencies not all be more than 5.

Homework

Exercise 11.1.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. The number of people who survived the Titanic based on class and sex is in Example 11.1.5 ("Encyclopedia Titanica," 2013). Is there enough evidence to show that the class and the sex of a person who survived the Titanic are independent? Test at the 5% level.

Table 11.1.5: Surviving the Titanic

Class	Sex		Total
	Female	Male	
1st	134	59	193
2nd	94	25	119
3rd	80	58	138
Total	308	142	450

2. Researchers watched groups of dolphins off the coast of Ireland in 1998 to determine what activities the dolphins partake in at certain times of the day ("Activities of dolphin," 2013). The numbers in Example 11.1.6 represent the number of groups of dolphins that were partaking in an activity at certain times of days. Is there enough evidence to show that the activity and the time period are independent for dolphins? Test at the 1% level.

Table 11.1.6: Dolphin Activity

Activity	Period				Row Total
	Morning	Noon	Afternoon	Evening	
Travel	6	6	14	13	39
Feed	28	4	0	56	88
Social	38	5	9	10	62
Column Total	72	15	23	79	189

3. Is there a relationship between autism and what an infant is fed? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what they fed their infant. The data is in Example 11.1.7 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that what an infant is fed and autism are independent? Test at the 1% level.

Table 11.1.7: Autism Versus Breastfeeding

Autism	Feeding			Row Total
	Breast feeding	Formula with DHA/ARA	Formula without DRA/ARA	
Yes	12	39	65	116
No	6	22	10	38
Column Total	18	61	75	164

4. A person's educational attainment and age group was collected by the U.S. Census Bureau in 1984 to see if age group and educational attainment are related. The counts in thousands are in Example 11.1.8 ("Education by age," 2013). Do the data show that educational attainment and age are independent? Test at the 5% level.

Table 11.1.8: Educational Attainment and Age Group

Education	Age Group					Row Total
	25-34	35-44	45-54	55-64	>64	
Did not complete HS	5416	5030	5777	7606	13746	37575
Completed HS	16431	1855	9435	8795	7558	44074
College 1-3 years	8555	5576	3124	2524	2503	22282
College 4 or more years	9771	7596	3904	3109	2483	26863
Column Total	40173	20057	22240	22034	26290	130794

5. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important good grades were to them (1 very important and 4 least important). The data is in Example 11.1.9 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of grades are independent? Test at the 5% level.

Table 11.1.9: Personal Goal and Importance of Grades

Goal	Grades Importance Rating				Row Total
	1	2	3	4	
Grades	70	66	55	56	247
Popular	14	33	45	49	141
Sports	10	24	33	23	90
Column Total	94	123	133	128	478

6. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important being good at sports were to them (1 very important and 4 least important). The data is in Example 11.1.10 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of sports are independent? Test at the 5% level.

Table 11.1.10: Personal Goal and Importance of Sports

Goal	Sports Importance Rating				Row Total
	1	2	3	4	
Grades	83	81	55	28	247
Popular	32	49	43	17	141
Sports	50	24	14	2	90
Column Total	165	154	112	47	478

7. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having good looks were to them (1 very important and 4 least important). The data is in Example 11.1.11 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of looks

are independent? Test at the 5% level.

Table 11.1.11: Personal Goal and Importance of Looks

Goal	Looks Importance Rating				Row Total
	1	2	3	4	
Grades	80	66	66	35	247
Popular	81	30	18	12	141
Sports	24	30	17	19	90
Column Total	185	126	101	66	478

8. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having money were to them (1 very important and 4 least important). The data is in Example 11.1.12 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of money are independent? Test at the 5% level.

Table 11.1.12: Personal Goal and Importance of Money

Goal	Money Importance Rating				Row Total
	1	2	3	4	
Grades	14	34	71	128	247
Popular	14	29	35	63	141
Sports	6	12	26	46	90
Column Total	34	75	132	237	478

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Reject Ho

3. Reject Ho

5. Reject Ho

7. Reject Ho

This page titled 11.1: Chi-Square Test for Independence is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

11.2: Chi-Square Goodness of Fit

In probability, you calculated probabilities using both experimental and theoretical methods. There are times when it is important to determine how well the experimental values match the theoretical values. An example of this is if you wish to verify if a die is fair. To determine if observed values fit the expected values, you want to see if the difference between observed values and expected values is large enough to say that the test statistic is unlikely to happen if you assume that the observed values fit the expected values. The test statistic in this case is also the chi-square. The process is the same as for the chi-square test for independence.

Hypothesis Test for Goodness of Fit Test

1. State the null and alternative hypotheses and the level of significance
   H_o: The data are consistent with a specific distribution
   H_A: The data are not consistent with a specific distribution
   Also, state your α level here.
2. State and check the assumptions for the hypothesis test
   a. A random sample is taken.
   b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, *E*, will be calculated later, and this assumption means $E \geq 5$).
3. Find the test statistic and p-value
   Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is *O*. The table is made up of k entries. The total number of observed frequencies is *n*. The expected frequencies are calculated by multiplying the probability of each entry, *p*, times *n*.

$$\text{Expected frequency}(\text{ entry } i) = E = n^* p$$

Test Statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* is the observed frequency and *E* is the expected frequency.

Again, the test statistic involves squaring the differences, so the test statistics are all positive. Thus a chi-squared test for goodness of fit is always right tailed.

p-value:

Using the TI-83/84: χ cdf (lower limit, 1E99, df)

Using R: $1 -$ pchisq (χ^2, df)

Where the degrees of freedom is *df* = *k* - 1

4. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o,

5. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Example 11.2.1 goodness of fit test using the formula

Suppose you have a die that you are curious if 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. The data is in Example 11.2.1. Do the data show that the die is fair? Test at the 5% level.

Table 11.2.1: 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_o: The observed frequencies are consistent with the distribution for fair die (the die is fair)

H_A: The observed frequencies are not consistent with the distribution for fair die (the die is not fair)

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample is taken since each throw of a die is a random event.
b. Expected frequencies for each cell are greater than or equal to 5. See step 3.

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, then each expected frequency is the same.

$$\text{Expected Frequency} = E = n^*p = 500 * \frac{1}{6} \approx 83.33$$

Test Statistic:

It is easier to calculate the test statistic using a table.

Table 11.2.2: Calculation of the Chi-Square Test Statistic

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		$\chi^2 \approx 1.504060962$

The test statistic is $\chi^2 \approx 1.504060962$

The degrees of freedom are *df = k - 1 = 6 - 1 = 5*

Using TI-83/84: $p - \text{value} = \chi^2 \text{cdf}(1.50406096, 1E99, 5) \approx 0.913$

Using R: $p - \text{value} = 1 - \text{ pchisq } (1.50406096, 5) \approx 0.9126007$

4. Conclusion

Fail to reject H_o since the p-value is greater than 0.05.

5. Interpretation

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.

Example 11.2.2 goodness of fit test using technology

Suppose you have a die that you are curious if 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. The data is in Example 11.2.1. Do the data show that the die is fair? Test at the 5% level.

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: The observed frequencies are consistent with the distribution for fair die (the die is fair)

H_A: The observed frequencies are not consistent with the distribution for fair die (the die is not fair)

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample is taken since each throw of a die is a random event.
b. Expected frequencies for each cell are greater than or equal to 5. See step 3.

3. Find the test statistic and p-value

Using the TI-83/84 calculator:

Using the TI-83:

To use the TI-83 calculator to compute the test statistic, you must first put the data into the calculator. Type the observed frequencies into L1 and the expected frequencies into L2. Then you will need to go to L3, arrow up onto the name, and type in $(L1 - L2)\text{^}2/L2$. Now you use 1-Var Stats L3 to find the total. See *Figure* 11.2.1 for the initial setup, *Figure 11.2.2* for the results of that calculation, and *Figure* 11.2.3 for the result of the 1-Var Stats L3.

Figure 11.2.1: Input into TI-83

Figure 11.2.2: Result for L3 on TI-83

Figure 11.2.3: 1-Var Stats L3 Result on TI-83

The total is the chi-square value, $\chi^2 = \sum x \approx 1.50406$.

The p-value is found using $p - \text{value} = \chi^2 \text{cdf}(1.50406096, 1E99, 5) \approx 0.913$, where the degrees of freedom is *df* = *k* - 1 = 6 - 1 = 5.

Using the TI-84:

To run the test on the TI-84, type the observed frequencies into L1 and the expected frequencies into L2, then go into STAT, move over to TEST and choose χ^2 GOF-Test from the list. The setup for the test is in *Figure* 11.2.4.

Figure 11.2.4: Setup for Chi-Square Goodness of Fit Test on TI-84

Once you press ENTER on Calculate you will see the results in *Figure* 11.2.5.

Figure 11.2.5: Results for Chi-Square Test on TI-83/84

The test statistic is $\chi^2 \approx 1.504060962$

The *p-value* ≈ 0.913

The CNTRB represent the $\frac{(O-E)^2}{E}$ for each die value. You can see the values by pressing the right arrow.

Using R:

Type in the observed frequencies. Call it something like observed.
observed<- c(type in data with commas in between)
Type in the probabilities that you are comparing to the observed frequencies.
Call it something like null.probs.
null.probs <- c(type in probabilities with commas in between)
chisq.test(observed, p=null.probs) – the command for the hypothesis test

For this example (Note since you are looking to see if the die is fair, then the probability of each side of a fair die coming up is 1/6.)
observed<-c(78, 87, 87, 76, 85, 87)
null.probs<-c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
chisq.test(observed, p=null.probs)

Output:
Chi-squared test for given probabilities

data: observed
X-squared = 1.504, df = 5, p-value = 0.9126

The test statistic is $\chi^2 = 1.504$ and the p-value = 0.9126.

4. Conclusion

Fail to reject H_o since the p-value is greater than 0.05.

5. Interpretation

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.

Homework

Exercise 11.2.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. According to the M&M candy company, the expected proportion can be found in Example 11.2.3. In addition, the table contains the number of M&M's of each color that were found in a case of candy (Madison, 2013). At the 5% level, do the observed frequencies support the claim of M&M?

Table 11.2.3: M&M Observed and Proportions

	Blue	Brown	Green	Orange	Red	Yellow	Total
Observed Frequencies	481	371	483	544	372	369	2620
Expected Proportion	0.24	0.13	0.16	0.20	0.13	0.14	

2. Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Example 11.2.4 gives the defect and the number of defects. Do the data support the notion that each defect type occurs in the same proportion? Test at the 10% level.

Table 11.2.4: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped - small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485

Defect type	Number of defects
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape - big	1105
Lost in lab	976
Spots/bubble - intern	976

3. On occasion, medical studies need to model the proportion of the population that has a disease and compare that to observed frequencies of the disease actually occurring. Suppose the end-stage renal failure in south-west Wales was collected for different age groups. Do the data in Example 11.2.5 show that the observed frequencies are in agreement with proportion of people in each age group (Boyle, Flowerdew & Williams, 1997)? Test at the 1% level.

Table 11.2.5: Renal Failure Frequencies

Age Group	16-29	30-44	45-59	60-75	75+	Total
Observed Frequency	32	66	132	218	91	539
Expected Proportion	0.23	0.25	0.22	0.21	0.09	

4. In Africa in 2011, the number of deaths of a female from cardiovascular disease for different age groups are in Example 11.2.6 ("Global health observatory," 2013). In addition, the proportion of deaths of females from all causes for the same age groups are also in Example 11.2.6. Do the data show that the death from cardiovascular disease are in the same proportion as all deaths for the different age groups? Test at the 5% level.

Table 11.2.6: Deaths of Females for Different Age Groups

Age	5-14	15-29	30-49	50-69	Total
Cardiovascular Frequency	9	16	56	433	513
All Cause Proportion	0.10	0.12	0.26	0.52	

5. In Australia in 1995, there was a question of whether indigenous people are more likely to die in prison than non-indigenous people. To figure out, the data in Example 11.2.7 was collected. ("Aboriginal deaths in," 2013). Do the data show that indigenous people die in the same proportion as non-indigenous people? Test at the 1% level.

Table 11.2.7: Death of Prisoners

	Prisoner Dies	Prisoner Did Not Die	Total
Indigenous Prisoner Frequency	17	2890	2907
Frequency of Non-Indigenous Prisoner	42	14459	14501

6. A project conducted by the Australian Federal Office of Road Safety asked people many questions about their cars. One question was the reason that a person chooses a given car, and that data is in Example 11.2.8 ("Car preferences," 2013).

Table 11.2.8: Reason for Choosing a Car

Safety	Reliability	Cost	Performance	Comfort	Looks

84	62	46	34	47	27

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Reject Ho

3. Reject Ho

5. Reject Ho

This page titled 11.2: Chi-Square Goodness of Fit is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

11.3: Analysis of Variance (ANOVA)

There are times where you want to compare three or more population means. One idea is to just test different combinations of two means. The problem with that is that your chance for a type I error increases. Instead you need a process for analyzing all of them at the same time. This process is known as **analysis of variance (ANOVA)**. The test statistic for the ANOVA is fairly complicated, you will want to use technology to find the test statistic and p-value. The test statistic is distributed as an F-distribution, which is skewed right and depends on degrees of freedom. Since you will use technology to find these, the distribution and the test statistic will not be presented. Remember, all hypothesis tests are the same process. Note that to obtain a statistically significant result there need only be a difference between any two of the *k* means.

Before conducting the hypothesis test, it is helpful to look at the means and standard deviations for each data set. If the sample means with consideration of the sample standard deviations are different, it may mean that some of the population means are different. However, do realize that if they are different, it doesn't provide enough evidence to show the population means are different. Calculating the sample statistics just gives you an idea that conducting the hypothesis test is a good idea.

Hypothesis test using ANOVA to compare *k* means

1. State the random variables and the parameters in words
   $x_1 =$ random variable 1
   $x_2 =$ random variable 2
   $\vdots$
   $x_k =$ random variable k
   $\mu_1 =$ mean of random variable 2
   $\mu_2 =$ mean of random variable 2
   $\vdots$
   $\mu_k =$ mean of random variable k
2. State the null and alternative hypotheses and the level of significance
   $H_o : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
   H_A : at least two of the means are not equal
   Also, state your α level here.
3. State and check the assumptions for the hypothesis test
   a. A random sample of size n_i is taken from each population.
   b. All the samples are independent of each other.
   c. Each population is normally distributed. The ANOVA test is fairly robust to the assumption especially if the sample sizes are fairly close to each other. Unless the populations are really not normally distributed and the sample sizes are close to each other, then this is a loose assumption.
   d. The population variances are all equal. If the sample sizes are close to each other, then this is a loose assumption.
4. . Find the test statistic and p-value
   The test statistic is $F = \frac{MS_B}{MS_W}$, where $MS_B = \frac{SS_B}{df_B}$ is the mean square between the groups (or factors), and $MS_W = \frac{SS_W}{df_W}$ is the mean square within the groups. The degrees of freedom between the groups is $df_B = k - 1$ and the degrees of freedom within the groups is $df_W = n_1 + n_2 + \cdots + n_k - k$. To find all of the values, use technology such as the TI-83/84 calculator or R.
   The test statistic, *F*, is distributed as an F-distribution, where both degrees of freedom are needed in this distribution. The p-value is also calculated by the calculator or R.
5. Conclusion
   This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.
6. Interpretation
   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

If you do in fact reject H_o, then you know that at least two of the means are different. The next question you might ask is which are different? You can look at the sample means, but realize that these only give a preliminary result. To actually determine which means are different, you need to conduct other tests. Some of these tests are the range test, multiple comparison tests, Duncan test,

Student-Newman-Keuls test, Tukey test, Scheffé test, Dunnett test, least significant different test, and the Bonferroni test. There is no consensus on which test to use. These tests are available in statistical computer packages such as Minitab and SPSS.

Example 11.3.1 hypothesis test involving several means

Cancer is a terrible disease. Surviving may depend on the type of cancer the person has. To see if the mean survival time for several types of cancer are different, data was collected on the survival time in days of patients with one of these cancer in advanced stage. The data is in Example 11.3.1 ("Cancer survival story," 2013). (Please realize that this data is from 1978. There have been many advances in cancer treatment, so do not use this data as an indication of survival rates from these cancers.) Do the data indicate that at least two of the mean survival time for these types of cancer are not all equal? Test at the 1% level.

Table 11.3.1: Survival Times in Days of Five Cancer Types

Stomach	Bronchus	Colon	Ovary	Breast
124	81	248	1234	1235
42	461	377	89	24
25	20	189	201	1581
45	450	1843	356	1166
412	246	180	2970	40
51	166	537	456	727
1112	63	519		3808
46	64	455		791
103	155	406		1804
876	859	365		3460
146	151	942		719
340	166	776		
396	37	372		
	223	163		
	138	101		
	72	20		
	245	283		

Solution

1. State the random variables and the parameters in words

$x_1 =$ survival time from stomach cancer
$x_2 =$ survival time from bronchus cancer
$x_3 =$ survival time from colon cancer
$x_4 =$ survival time from ovarian cancer
$x_5 =$ survival time from breast cancer
$\mu_1 =$ mean survival time from breast cancer
$\mu_1 =$ mean survival time from bronchus cancer
$\mu_3 =$ mean survival time from colon cancer
$\mu_4 =$ mean survival time from ovarian cancer
$\mu_5 =$ mean survival time from breast cancer

Now before conducting the hypothesis test, look at the means and standard deviations.

$\overline{x}_1 = 286$ $s_1 \approx 346.31$
$\overline{x}_2 \approx 211.59$ $s_2 \approx 209.86$
$\overline{x}_3 \approx 457.41$ $s_3 \approx 427.17$
$\overline{x}_4 \approx 884.33$ $s_4 \approx 1098.58$
$\overline{x}_5 \approx 1395.91$ $s_5 \approx 1238.97$

There appears to be a difference between at least two of the means, but realize that the standard deviations are very different. The difference you see may not be significant.

Notice the sample sizes are not the same. The sample sizes are

$n_1 = 13, n_2 = 17, n_3 = 17, n_4 = 6, n_5 = 11$

2. State the null and alternative hypotheses and the level of significance

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_A : at least two of the means are not equal

α = 0.01

3. State and check the assumptions for the hypothesis test

a. A random sample of 13 survival times from stomach cancer was taken. A random sample of 17 survival times from bronchus cancer was taken. A random sample of 17 survival times from colon cancer was taken. A random sample of 6 survival times from ovarian cancer was taken. A random sample of 11 survival times from breast cancer was taken. These statements may not be true. This information was not shared as to whether the samples were random or not but it may be safe to assume that.
b. Since the individuals have different cancers, then the samples are independent.
c. Population of all survival times from stomach cancer is normally distributed.
Population of all survival times from bronchus cancer is normally distributed.
Population of all survival times from colon cancer is normally distributed.
Population of all survival times from ovarian cancer is normally distributed.
Population of all survival times from breast cancer is normally distributed.
Looking at the histograms, box plots and normal quantile plots for each sample, it appears that none of the populations are normally distributed. The sample sizes are somewhat different for the problem. This assumption may not be true.
d. The population variances are all equal. The sample standard deviations are approximately 346.3, 209.9, 427.2, 1098.6, and 1239.0 respectively. This assumption does not appear to be met, since the sample standard deviations are very different. The sample sizes are somewhat different for the problem. This assumption may not be true.

4. Find the test statistic and p-value

To find the test statistic and p-value using the TI-83/84, type each data set into L1 through L5. Then go into STAT and over to TESTS and choose ANOVA(. Then type in L1,L2,L3,L4,L5 and press enter. You will get the results of the ANOVA test.

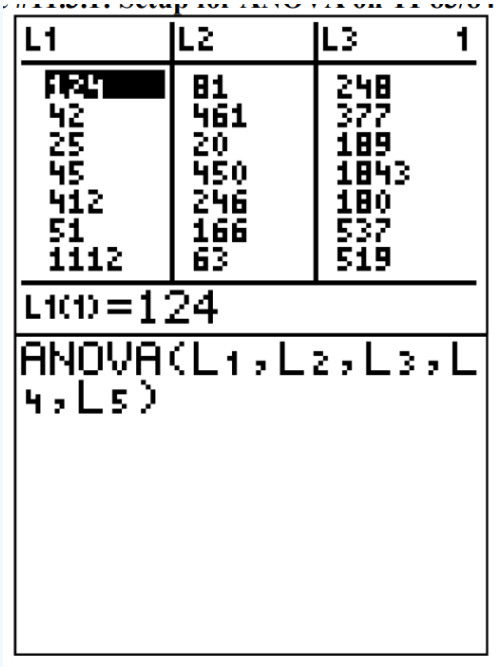


Figure 11.3.1: Setup for ANOVA on TI-83/84

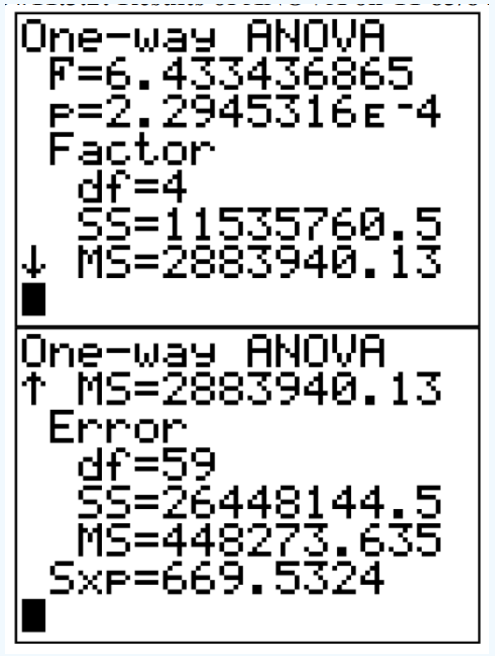


Figure 11.3.2: Results of ANOVA on TI-83/84

The test statistic is $F \approx 6.433$ and $p-$ value $\approx 2.29 \times 10^{-4}$

Just so you know, the Factor information is between the groups and the Error is within the groups. So

$MS_B \approx 2883940.13, SS_B \approx 11535760.5,$ and $df_B = 4$ and
$MS_W \approx 448273.635, SS_W \approx 448273.635,$ and $df_W = 59$

To find the test statistic and p-value on R:
The commands would be:
variable=c(type in all data values with commas in between) – this is the response variable
factor=c(rep("factor 1", number of data values for factor 1), rep("factor 2", number of data values for factor 2), etc) – this separates the data into the different factors that the measurements were based on.

data_name = data.frame(variable, factor) – this puts the data into one variable. data_name is the name you give this variable
aov(variable ~ factor, data = data name) – runs the ANOVA analysis

For this example, the commands would be:
time=c(124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340, 396, 81, 461, 20, 450, 246, 166, 63, 64, 155, 859, 151, 166, 37, 223, 138, 72, 245, 248, 377, 189, 1843, 180, 537, 519, 455, 406, 365, 942, 776, 372, 163, 101, 20, 283, 1234, 89, 201, 356, 2970, 456, 1235, 24, 1581, 1166, 40, 727, 3808, 791, 1804, 3460, 719)
factor=c(rep("Stomach", 13), rep("Bronchus", 17), rep("Colon", 17), rep("Ovary", 6), rep("Breast", 11))
survival=data.frame(time, factor)
results=aov(time~factor, data=survival)
summary(results)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor	4	11535761	2883940	6.4333	0.000229 * **
Residuals	59	26448144	448274		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The test statistic is F = 6.433 and the p-value = 0.000229.

5. Conclusion

Reject H_o since the p-value is less than 0.01.

6. Interpretation

There is evidence to show that at least two of the mean survival times from different cancers are not equal.

By examination of the means, it appears that the mean survival time for breast cancer is different from the mean survival times for both stomach and bronchus cancers. It may also be different for the mean survival time for colon cancer. The others may not be different enough to actually say for sure.

Homework

Exercise 11.3.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. Cuckoo birds are in the habit of laying their eggs in other birds’ nest. The other birds adopt and hatch the eggs. The lengths (in cm) of cuckoo birds’ eggs in the other species nests were measured and are in Example 11.3.2 ("Cuckoo eggs in," 2013). Do the data show that the mean length of cuckoo bird’s eggs is not all the same when put into different nests? Test at the 5% level.

Table 11.3.2: Lengths of Cuckoo Bird Eggs in Different Species Nests

Meadow Pipit		Tree Pipit	Hedge Sparrow	Robin	Pied Wagtail	Wren
19.65	22.25	21.05	20.85	21.05	21.05	19.85
20.05	22.25	21.85	21.65	21.85	21.85	20.05
20.65	22.25	22.05	22.05	22.05	21.85	20.25
20.85	22.25	22.45	22.85	22.05	21.85	20.85
21.65	22.65	22.65	23.05	22.05	22.05	20.85
21.65	22.65	23.25	23.05	22.25	22.45	20.85
21.65	22.85	23.25	23.05	22.45	22.65	21.05
21.85	22.85	23.25	23.05	22.45	23.05	21.05

Meadow Pipit		Tree Pipit	Hedge Sparrow	Robin	Pied Wagtail	Wren
21.85	22.85	23.45	23.45	22.65	23.05	21.05
21.85	22.85	23.45	23.85	23.05	23.25	21.25
22.05	23.05	23.65	23.85	23.05	23.45	21.45
22.05	23.25	23.85	23.85	23.05	24.05	22.05
22.05	23.25	24.05	24.05	23.05	24.05	22.05
22.05	23.45	24.05	25.05	23.05	24.05	22.05
22.05	23.65	24.05		23.25	24.85	22.25
22.05	23.85			23.85		
22.05	24.25					
22.05	24.45					
22.05	22.25					
22.05	22.25					
22.25	22.25					
22.25	22.25					
22.25						

2. Levi-Strauss Co manufactures clothing. The quality control department measures weekly values of different suppliers for the percentage difference of waste between the layout on the computer and the actual waste when the clothing is made (called run-up). The data is in Example 11.3.3, and there are some negative values because sometimes the supplier is able to layout the pattern better than the computer ("Waste run up," 2013). Do the data show that there is a difference between some of the suppliers? Test at the 1% level.

Table 11.3.3: Run-ups for Different Plants Making Levi Strauss Clothing

Plant 1	Plant 2	Plant 3	Plant 4	Plant 5
1.2	16.4	12.1	11.5	24
10.1	-6	9.7	10.2	-3.7
-2	-11.6	7.4	3.8	8.2
1.5	-1.3	-2.1	8.3	9.2
-3	4	10.1	6.6	-9.3
-0.7	17	4.7	10.2	8
3.2	3.8	4.6	8.8	15.8
2.7	4.3	3.9	2.7	22.3
-3.2	10.4	3.6	5.1	3.1
-1.7	4.2	9.6	11.2	16.8
2.4	8.5	9.8	5.9	11.3
0.3	6.3	6.5	13	12.3

Plant 1	Plant 2	Plant 3	Plant 4	Plant 5
3.5	9	5.7	6.8	16.9
-0.8	7.1	5.1	14.5	
19.4	4.3	3.4	5.2	
2.8	19.7	-0.8	7.3	
13	3	-3.9	7.1	
42.7	7.6	0.9	3.4	
1.4	70.2	1.5	0.7	
3	8.5			
2.4	6			
1.3	2.9			

3. Several magazines were grouped into three categories based on what level of education of their readers the magazines are geared towards: high, medium, or low level. Then random samples of the magazines were selected to determine the number of three-plus-syllable words were in the advertising copy, and the data is in Example 11.3.4 ("Magazine ads readability," 2013). Is there enough evidence to show that the mean number of three-plus-syllable words in advertising copy is different for at least two of the education levels? Test at the 5% level.

Table 11.3.4: Number of Three Plus Syllable Words in Advertising Copy

High Education	Medium Education	Low Education
34	13	7
21	22	7
37	25	7
31	3	7
10	5	7
24	2	7
39	9	8
10	3	8
17	0	8
18	4	8
32	29	8
17	26	8
3	5	9
10	5	9
6	24	9
5	15	9
6	3	9

High Education	Medium Education	Low Education
6	8	9

4. A study was undertaken to see how accurate food labeling for calories on food that is considered reduced calorie. The group measured the amount of calories for each item of food and then found the percent difference between measured and labeled food, $\frac{(\text{measured - labeled})}{\text{labeled}} * 100\%$. The group also looked at food that was nationally advertised, regionally distributed, or locally prepared. The data is in Example 11.3.5 ("Calories datafile," 2013). Do the data indicate that at least two of the mean percent differences between the three groups are different? Test at the 10% level.

Table 11.3.5: Percent Differences Between Measured and Labeled Food

National Advertised	Regionally Advertised	Locally Prepared
2	41	15
-28	46	60
-6	2	250
8	25	145
6	39	6
-1	16.5	8-
1-	17	95
13	28	3
15	-3	
-4	14	
-4	34	
-18	42	
10		
5		
3		
-7		
3		
-0.5		
-10		
6		

5. The amount of sodium (in mg) in different types of hotdogs is in Example 11.3.6 ("Hot dogs story," 2013). Is there sufficient evidence to show that the mean amount of sodium in the types of hotdogs are not all equal? Test at the 5% level.

Table 11.3.6: Amount of Sodium (in mg) in Beef, Meat, and Poultry Hotdogs

Beef	Meat	Poultry
495	458	430

Beef	Meat	Poultry
477	506	375
425	473	396
322	545	383
482	496	387
587	360	542
370	387	359
322	386	357
479	507	528
375	393	513
330	405	426
300	372	513
386	144	358
401	511	581
645	405	588
440	428	522
317	339	545
319		
298		
253		

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Reject Ho

3. Reject Ho

5. Fail to reject Ho

Data Source:

Aboriginal deaths in custody. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/custody.html

Activities of dolphin groups. (2013, September 26). Retrieved from http://www.statsci.org/data/general/dolpacti.html

Boyle, P., Flowerdew, R., & Williams, A. (1997). Evaluating the goodness of fit in models of sparse medical data: A simulation approach. *International Journal of Epidemiology*, 26(3), 651-656. Retrieved from http://ije.oxfordjournals.org/conten...3/651.full.pdf html

Calories datafile. (2013, December 07). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/Calories.html

Cancer survival story. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Stories...rSurvival.html

Car preferences. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/carprefs.html

Cuckoo eggs in nest of other birds. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Stories/cuckoo.html

Education by age datafile. (2013, December 05). Retrieved from lib.stat.cmu.edu/DASL/Datafil...tionbyage.html

Encyclopedia Titanica. (2013, November 09). Retrieved from www.encyclopediatitanica.org/

Global health observatory data respository. (2013, October 09). Retrieved from http://apps.who.int/gho/athena/data/...t=GHO/MORT_400 &profile=excel&filter=AGEGROUP:YEARS05-14;AGEGROUP:YEARS15-29;AGEGROUP:YEARS30-49;AGEGROUP:YEARS50-69;AGEGROUP:YEARS70;MGHEREG:REG6_AFR;GHECAUSES:*;SEX:*

Hot dogs story. (2013, November 16). Retrieved from lib.stat.cmu.edu/DASL/Stories/Hotdogs.html

Leprosy: Number of reported cases by country. (2013, September 04). Retrieved from http://apps.who.int/gho/data/node.main.A1639

Magazine ads readability. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Datafiles/magadsdat.html

Popular kids datafile. (2013, December 05). Retrieved from lib.stat.cmu.edu/DASL/Datafil...pularKids.html

Schultz, S. T., Klonoff-Cohen, H. S., Wingard, D. L., Askhoomoff, N. A., Macera, C. A., Ji, M., & Bacher, C. (2006). Breastfeeding, infant formula supplementation, and autistic disorder: the results of a parent survey. *International Breastfeeding Journal*, 1(16), doi: 10.1186/1746-4358-1-16

Waste run up. (2013, December 04). Retrieved from lib.stat.cmu.edu/DASL/Stories/wasterunup.html

---

This page titled 11.3: Analysis of Variance (ANOVA) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

12: Appendix- Critical Value Tables

12.1: Critical Values for t-Interval
12.2: Normal Critical Values for Confidence Levels

This page titled 12: Appendix- Critical Value Tables is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

12.1: Critical Values for t-Interval

Table A.2: Critical Values for t-Interval

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
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	.1761	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
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738

Degrees of Freedom (*df*)	80%	90%	95%	98%	99%
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
36	1.306	1.688	2.028	2.434	2.719
37	1.305	1.687	2.026	2.431	2.715
38	1.304	1.686	2.024	2.429	2.712
39	1.304	1.685	2.023	2.426	2.712
40	1.303	1.684	2.021	2.423	2.704
41	1.303	1.683	2.020	2.421	2.701
42	1.302	1.682	2.018	2.418	2.698
43	1.302	1.681	2.017	2.416	2.695
44	1.301	1.680	2.015	2.414	2.692
45	1.301	1.679	2.014	2.412	2.690
46	1.300	1.679	2.013	2.410	2.687
47	1.300	1.678	2.012	2.408	2.685
48	1.299	1.677	2.011	2.407	2.682
49	1.299	1.677	2.010	2.405	2.680
50	1.299	1.676	2.009	2.403	2.678
51	1.298	1.675	2.008	2.402	2.676
52	1.298	1.675	2.007	2.400	2.674
53	1.298	1.674	2.006	2.399	2.672
54	1.297	1.674	2.005	2.397	2.670
55	1.297	1.673	2.004	2.396	2.668
56	1.297	1.673	2.003	2.395	2.667
57	1.297	1.672	2.002	2.394	2.665
58	1.296	1.672	2.002	2.392	2.663
59	1.296	1.671	2.001	2.391	2.662
60	1.296	1.671	2.000	2.390	2.660
61	1.296	1.670	2.000	2.389	2.659
62	1.295	1.670	1.999	2.388	2.657
63	1.295	1.669	1.998	2.387	2.656
64	1.295	1.669	1.998	2.386	2.655
65	1.295	1.669	1.997	2.385	2.654
66	1.295	1.668	1.997	2.384	2.652

Degrees of Freedom (*df*)	80%	90%	95%	98%	99%
67	1.294	1.668	1.996	2.383	2.651
68	1.294	1.668	1.995	2.382	2.650
69	1.294	1.667	1.995	2.382	2.649
70	1.294	1.667	1.994	2.381	2.648
71	1.294	1.667	1.994	2.380	2.647
72	1.293	1.666	1.993	2.379	2.646
73	1.293	1.666	1.993	2.379	2.645
74	1.293	1.666	1.993	2.378	2.644
75	1.293	1.665	1.992	2.377	2.643
76	1.293	1.665	1.992	2.376	2.642
77	1.293	1.665	1.991	2.376	2.641
78	1.292	1.665	1.991	2.375	2.640
79	1.292	1.664	1.990	2.374	2.640
80	1.292	1.664	1.990	2.374	2.639
81	1.292	1.664	1.990	2.373	2.638
82	1.292	1.664	1.989	2.373	2.637
83	1.292	1.663	1.989	2.372	2.636
84	1.292	1.663	1.989	2.372	2.636
85	1.292	1.663	1.988	2.371	2.635
86	1.291	1.663	1.988	2.370	2.634
87	1.291	1.663	1.988	2.370	2.634
88	1.291	1.662	1.987	2.369	2.633
89	1.291	1.662	1.987	2.369	2.632
90	1.291	1.662	1.987	2.368	2.632
91	1.291	1.662	1.986	2.368	2.631
92	1.291	1.662	1.986	2.368	2.630
93	1.291	1.661	1.986	2.367	2.630
94	1.291	1.661	1.986	2.367	2.629
95	1.291	1.661	1.985	2.366	2.629
96	1.290	1.661	1.985	2.366	2.628
97	1.290	1.661	1.985	2.365	2.627
98	1.290	1.661	1.984	2.365	2.627
99	1.290	1.660	1.984	2.365	2.626
100	1.290	1.660	1.984	2.364	2.626

Degrees of Freedom (*df*)	80%	90%	95%	98%	99%
101	1.290	1.660	1.984	2.364	2.625
102	1.290	1.660	1.983	2.363	2.625
103	1.290	1.660	1.983	2.363	2.624
104	1.290	1.660	1.983	2.363	2.624
105	1.290	1.659	1.983	2.362	2.623

This page titled 12.1: Critical Values for t-Interval is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

12.2: Normal Critical Values for Confidence Levels

Table A.1: Normal Critical Values for Confidence Levels

Confidence Level, C	Critical Value, Z_c
99%	2.575
98%	2.33
95%	1.96
90%	1.645
80%	1.28

Critical values for Z_c created using Microsoft Excel

This page titled 12.2: Normal Critical Values for Confidence Levels is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Kathryn Kozak via source content that was edited to the style and standards of the LibreTexts platform.

Index

A

anova
- 11.3: Analysis of Variance (ANOVA)

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 and Standard Deviation of Binomial Distribution

C

central limit theorem
- 6.5: Sampling Distribution and the Central Limit Theorem

Chebyshev's theorem
- 3.2: Measures of Spread

Chi Square
- 8: Estimation

Complementary events
- 4.2: Theoretical Probability

confidence intervals
- 8.1: Basics of Confidence Intervals

confounding variable
- 1.4: How Not to Do Statistics

continuous probability distribution
- 6: Continuous Probability Distributions

D

deviation
- 3.2: Measures of Spread

E

equally likely outcomes
- 4.2: Theoretical Probability

F

frequency distribution
- 2.2: Quantitative Data

H

hidden bias
- 1.4: How Not to Do Statistics

histogram
- 2.2: Quantitative Data

Hypothesis Testing
- 7.1: Basics of Hypothesis Testing

I

Inferential Theory
- 4: Probability

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
- 6.2: Graphs of the Normal Distribution

O

overgeneralization
- 1.4: How Not to Do Statistics

P

Pareto charts
- 2.1: Qualitative Data

percentiles
- 3.3: Ranking

pie charts
- 2.1: Qualitative Data

Q

quartiles
- 3.3: Ranking

R

range
- 3.2: Measures of Spread

S

sample variance
- 3.2: Measures of Spread

sampling distribution
- 6.5: Sampling Distribution and the Central Limit Theorem

T

Theoretical Probabilities
- 4.2: Theoretical Probability

U

uniform distribution
- 6.1: Uniform Distribution

Index

A

anova
- 11.3: Analysis of Variance (ANOVA)

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 and Standard Deviation of Binomial Distribution

C

central limit theorem
- 6.5: Sampling Distribution and the Central Limit Theorem

Chebyshev's theorem
- 3.2: Measures of Spread

Chi Square
- 8: Estimation

Complementary events
- 4.2: Theoretical Probability

confidence intervals
- 8.1: Basics of Confidence Intervals

confounding variable
- 1.4: How Not to Do Statistics

continuous probability distribution
- 6: Continuous Probability Distributions

D

deviation
- 3.2: Measures of Spread

E

equally likely outcomes
- 4.2: Theoretical Probability

F

frequency distribution
- 2.2: Quantitative Data

H

hidden bias
- 1.4: How Not to Do Statistics

histogram
- 2.2: Quantitative Data

Hypothesis Testing
- 7.1: Basics of Hypothesis Testing

I

Inferential Theory
- 4: Probability

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
- 6.2: Graphs of the Normal Distribution

O

overgeneralization
- 1.4: How Not to Do Statistics

P

Pareto charts
- 2.1: Qualitative Data

percentiles
- 3.3: Ranking

pie charts
- 2.1: Qualitative Data

Q

quartiles
- 3.3: Ranking

R

range
- 3.2: Measures of Spread

S

sample variance
- 3.2: Measures of Spread

sampling distribution
- 6.5: Sampling Distribution and the Central Limit Theorem

T

Theoretical Probabilities
- 4.2: Theoretical Probability

U

uniform distribution
- 6.1: Uniform Distribution

Detailed Licensing

Overview

Title: Statistics with Technology 2e (Kozak)

Webpages: 64

All licenses found:

- CC BY-SA 4.0: 82.8% (53 pages)
- Undeclared: 17.2% (11 pages)

By Page

- Statistics with Technology 2e (Kozak) - *CC BY-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Preface - *Undeclared*
  - 1: Statistical Basics - *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*
  - 2: Graphical Descriptions of Data - *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*
  - 3: Examining the Evidence Using Graphs and Statistics - *CC BY-SA 4.0*
    - 3.1: Measures of Center - *CC BY-SA 4.0*
    - 3.2: Measures of Spread - *CC BY-SA 4.0*
    - 3.3: Ranking - *CC BY-SA 4.0*
  - 4: Probability - *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: Conditional Probability - *CC BY-SA 4.0*
    - 4.4: Counting Techniques - *CC BY-SA 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 and Standard Deviation of Binomial Distribution - *CC BY-SA 4.0*
  - 6: Continuous Probability Distributions - *CC BY-SA 4.0*
    - 6.1: Uniform Distribution - *CC BY-SA 4.0*
    - 6.2: Graphs of the Normal Distribution - *CC BY-SA 4.0*
    - 6.3: Finding Probabilities for the Normal Distribution - *CC BY-SA 4.0*
    - 6.4: Assessing Normality - *CC BY-SA 4.0*
    - 6.5: Sampling Distribution and the Central Limit Theorem - *CC BY-SA 4.0*
  - 7: One-Sample Inference - *CC BY-SA 4.0*
    - 7.1: Basics of Hypothesis Testing - *CC BY-SA 4.0*
    - 7.2: One-Sample Proportion Test - *CC BY-SA 4.0*
    - 7.3: One-Sample Test for the Mean - *CC BY-SA 4.0*
  - 8: Estimation - *CC BY-SA 4.0*
    - 8.1: Basics of Confidence Intervals - *CC BY-SA 4.0*
    - 8.2: One-Sample Interval for the Proportion - *CC BY-SA 4.0*
    - 8.3: One-Sample Interval for the Mean - *CC BY-SA 4.0*
  - 9: Two-Sample Interference - *CC BY-SA 4.0*
    - 9.1: Two Proportions - *CC BY-SA 4.0*
    - 9.2: Paired Samples for Two Means - *CC BY-SA 4.0*
    - 9.3: Independent Samples for Two Means - *CC BY-SA 4.0*
    - 9.4: Which Analysis Should You Conduct? - *CC BY-SA 4.0*
  - 10: Regression and Correlation - *CC BY-SA 4.0*
    - 10.1: Regression - *CC BY-SA 4.0*
    - 10.2: Correlation - *CC BY-SA 4.0*
    - 10.3: Inference for Regression and Correlation - *CC BY-SA 4.0*
  - 11: Chi-Square and ANOVA Tests - *CC BY-SA 4.0*
    - 11.1: Chi-Square Test for Independence - *CC BY-SA 4.0*
    - 11.2: Chi-Square Goodness of Fit - *CC BY-SA 4.0*
    - 11.3: Analysis of Variance (ANOVA) - *CC BY-SA 4.0*
  - 12: Appendix- Critical Value Tables - *CC BY-SA 4.0*
    - 12.1: Critical Values for t-Interval - *CC BY-SA 4.0*

- 12.2: Normal Critical Values for Confidence Levels - *CC BY-SA 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*