

Elementary Statistical Methods

Elementary Statistical Methods

TIM CONTRERAS, ODESSA COLLEGE



Elementary Statistical Methods Copyright © by Lumen Learning is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), except where otherwise noted.

Contents

Part I. Sampling and Data

1. Introduction	3
2. 1.1 Definitions of Statistics and Key Terms	5
3. 1.2 Data: Quantitative Data & Qualitative Data	17
4. 1.3 Sampling	25
5. 1.4 Levels of Measurement	33
6. 1.5 Frequency & Frequency Tables	39
7. 1.6 Experimental Design & Ethics	50

Part II. Descriptive Statistics

8. Introduction: Descriptive Statistics	59
9. 2.1 Stem-and-Leaf Graphs, Line Graphs, Bar Graphs, and Pie Charts	62
10. 2.2 Histograms, Frequency Polygons, and Time Series Graphs	81
11. 2.3 Measures of the Location of the Data	104
12. 2.4 Box Plots	135
13. 2.5 Measures of the Center of the Data	146
14. 2.6 Skewness and the Mean, Median, and Mode	161
15. 2.7 Measures of the Spread of Data	168
16. 2.8 When to use each measure of Central Tendency	195

Part III. Probability

17. Introduction: Probability Topics	199
18. 3.1 The Terminology of Probability	202
19. 3.2 Independent and Mutually Exclusive Events	218
20. 3.3 Two Basic Rules of Probability	244
21. 3.4 Contingency Tables	260
22. 3.5 Tree and Venn Diagrams	274

Part IV. Discrete Random Variables

23. Introduction: Discrete Random Variables	297
24. 4.1 Probability Distribution Function (PDF) for a Discrete Random Variable	301
25. 4.2 Mean / Expected Value and Standard Deviation of Discrete Random Variable	307
26. 4.3 Binomial Distribution	324
27. Geometric Distribution	341
28. Poisson Distribution	353

Part V. Normal Distribution

29. Introduction to the Normal Distribution (Need Pic)	361
30. 6.1 The Standard Normal Distribution	367
31. 6.2 Using the Normal Distribution	383

Part VI. The Central Limit Theorem

32. Introduction: The Central Limit Theorem	403
33. 7.1 The Central Limit Theorem for Sample Means (Averages)	406

34.	7.2 The Central Limit Theorem for Sums	419
35.	7.3 Using the Central Limit Theorem	430

Part VII. Confidence Intervals

36.	Introduction: Confidence Intervals	453
37.	8.1 A Single Population Mean using the Normal Distribution	459
38.	8.2 A Single Population Mean using the Student t Distribution	488
39.	8.3 Confidence Interval for Population Proportion	501

Part VIII. Hypothesis Testing With One Sample

40.	Introduction: Hypothesis Testing with One Sample	527
41.	9.1 Null and Alternative Hypotheses	535
42.	9.2 Outcomes, Type I and Type II Errors	542
43.	9.3 Distribution Needed for Hypothesis Testing	555
44.	9.4 Rare Events, the Sample, Decision and Conclusion	559
45.	9.5 Additional Information and Full Hypothesis Test Examples	568

Part IX. Linear Regression and Correlation

46.	Introduction: Linear Regression and Correlation	609
47.	12.1 Linear Equations	611
48.	12.2 Scatter Plots	620
49.	12.3 The Regression Equation	626
50.	12.4 Prediction	640
51.	12.5 Testing the Significance of the Correlation Coefficient	645

PART I

SAMPLING AND DATA

I. Introduction



We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news.
(credit: David Sim)

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

You are probably asking yourself the question, “When and where will I use statistics?” If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or “fact.” Statistical methods can help you make the “best educated guess.”

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what “good” data can be distinguished from “bad.”

2. 1.1 Definitions of Statistics and Key Terms

The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives.

Important Terms in Statistics

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of sampling is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Example:

1. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages.
Show the population and sample.

Population: All the students at your school

Sample: a sample of 50 students

2. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country.
Show the population and sample.

Population: All the people in the entire country

Sample: Sample of 1000–2000 people

3. City of Houston wants to know if the annual household income in the city is higher than national average. The statisticians collect data from 1500 families.
Show the population and sample.

Population: All the households in Houston

Sample: 1500 families who are being sampled.

4. An automobile manufacturer wanted to know if more than 50% of the US drivers own at least a domestic car. This company surveyed 10,000 drivers over US.
Show the population and sample.

Population: All the drivers in the US

Sample: 10,000 drivers who were being surveyed

From the sample data, we can calculate a statistic. A statistic is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A parameter is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

Population: all math classes

Sample: One of the math classes

Parameter: Average number of points earned per student over all math classes

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A variable, notated by capital letters such as X and Y , is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. Numerical variables take on values with equal units such as weight in pounds and time in hours. Categorical variables place the person or thing into a category.

Example:

1. If we assume that X is equal to the number of points earned by one math student at the end of a term, then X is a numerical variable.
 2. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable.
-

We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**.

If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE (We will learn the
meaning of mean in next
chapter!)

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean,” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

Population

all students who attended the college last year

Sample

a group of students who graduated from the college last year, randomly selected

Data

3.65, 2.80, 1.50, 3.90

Statistics

the cumulative GPA of one student who graduated from the college last year

Variable

the average cumulative GPA of students who graduated from the college last year

Parameter:

the average cumulative GPA of students in the study who graduated from the college last year

Try It

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Show Answer:

Population: All the first year college students at ABC College

Sample: 100 first year students who are surveyed at the college.

Parameter: Average amount of money a first year college student spent on school supplies that do not include books.

Statistics: Average amount of money these 100 first year college student spent on school supplies that do not include books.

Variable: The amount of money a first year ABC College student spend on school supplies that do not include books.

Data: The amount that we collected from these 100 students, like \$150, \$200, and \$225.

Example 2

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Solution:

Population: All families with children at Knoll Academy

Sample: 100 families with children who are surveyed in the school.

Parameter: average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

Statistics: average (mean) amount of money spent on school uniforms by families in the sample.

Variable: the amount of money spent by one family

Data: The amount that we collected from these 100 families, like \$65, \$75, and \$95.

Example 3

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

Population:

all cars containing dummies in the front seat.

Sample:

the 75 cars, selected by a simple random sample.

Parameter:

the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

Statistics:

proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

Variable:

the number of driver dummies (if they had been real people) who would have suffered head injuries.

Data:

yes, had head injury, or no, did not.

Try It

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Population:

all medical doctors listed in the professional directory.

Sample:

the 500 doctors selected at random from the professional directory.

Parameter

the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

Statistics

the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

Variable:

the number of medical doctors who have been involved in one or more malpractice suits.

Data

Yes, was involved in one or more malpractice lawsuits;
or no, was not.

Watch the following video for a brief introduction to statistics.



One or more interactive elements has been excluded
from this version of the text. You can view them online
here: [https://library.achievingthedream.org/
odessastatistics/?p=19#oembed-1](https://library.achievingthedream.org/odessastatistics/?p=19#oembed-1)

References

The Data and Story Library, [http://lib.stat.cmu.edu/DASL/Stories/
CrashTestDummies.html](http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html) (accessed May 1, 2013).

Concept Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Glossary

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are

being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

3. 1.2 Data: Quantitative Data & Qualitative Data

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

	Quantitative Data	Qualitative Data
Definition	Quantitative data are the result of counting or measuring attributes of a population.	Qualitative data are the result of categorizing or describing attributes of a population.
Data that you will see	Quantitative data are always numbers.	Qualitative data are generally described by words or letters.
Examples	Amount of money you have Height Weight Number of people living in your town Number of students who take statistics	Hair color Blood type Ethnic group The car a person drives The street a person lives on

Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are **quantitative**

continuous data assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π , $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the **quantitative discrete data**.

Try It

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Show Answer

It is quantitative discrete data

Example of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are **quantitative continuous data** because weights are measured.

Try It

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Show Answer

It is quantitative continuous data.

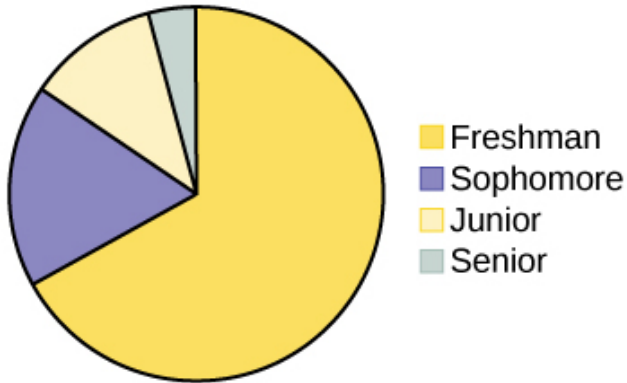
Examples of Qualitative Data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are **qualitative data**.

Try It

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart.

Classification of Statistics Students



What type of data does this graph show?

Show Answer

This pie chart shows the students in each year, which is qualitative data.

Example 1

Determine the correct data type (quantitative or qualitative).

Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words “the number of.”

1. The number of pairs of shoes you own
2. The type of car you drive
3. The place where you go on vacation
4. The distance it is from your home to the nearest grocery store
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use
8. Movie ratings
9. Political party preferences
10. Weights of sumo wrestlers
11. Amount of money (in dollars) won playing poker
12. Number of correct answers on a quiz
13. Peoples’ attitudes toward the government
14. IQ scores

Show Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Omitting Categories and Missing Data

The table displays **Ethnicity of Students** but is missing the “Other/Unknown” category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

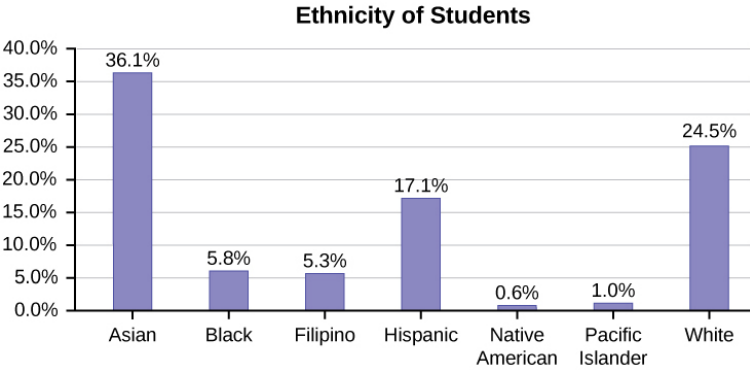


Figure 1. Ethnicity of Students

The following graph is the same as the previous graph but the “Other/Unknown” percent (9.6%) has been included. The “Other/Unknown” category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 2 can be difficult to understand visually.

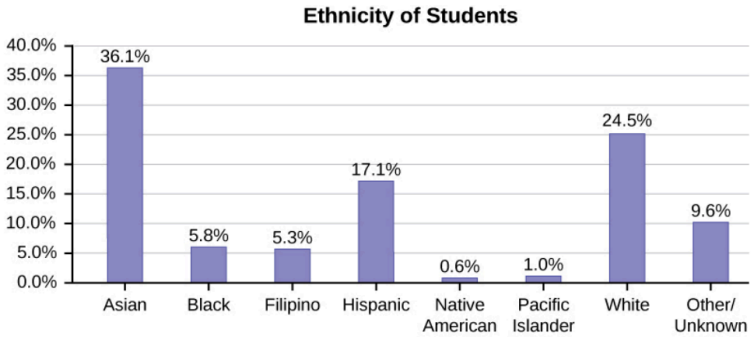


Figure 2. Bar Graph with Other/Unknown Category

The graph in Figure 3 is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

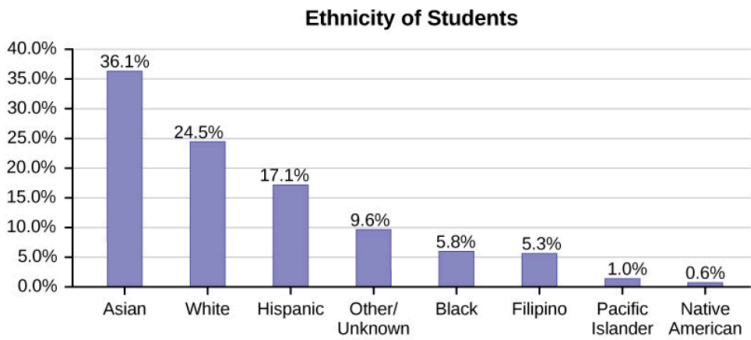


Figure 3. Pareto Chart with Bars Sorted by Size

4. 1.3 Sampling

Sampling

The following video introduces the different methods that statisticians use collect samples of data.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=21#oembed-1>

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons.

The 5 different types of random sampling methods are the simple random sample, the stratified sample, the cluster sample, and the systematic sample.

	<u>Type</u>
Random Sampling	1. Simple Random Sample
	2. Stratified Sample
	3. Cluster Sample
	4. Systematic Sample
	5. Convenient Sample

Simple Random Sample

The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in the following table.

Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cunningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

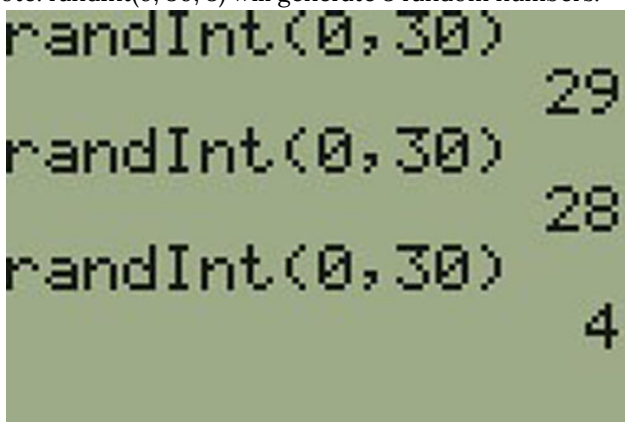
Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cunningham, and Cuarismo.

Generating Random Numbers

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers.
If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.



```
randInt(0, 30) → 29
randInt(0, 30) → 28
randInt(0, 30) → 4
```

Stratified Sample

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second

department, and so on represent the members who make up the stratified sample.

Cluster Sample

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

Systematic Sample

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

Convenience Sample

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample

is 1000 out of 10,000 (0.1000);

- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions $\frac{999}{10,000}$ and $\frac{999}{9,999}$. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions $\frac{9}{25}$ and $\frac{9}{24}$. To four decimal places, $\frac{9}{25} = 0.3600$ and $\frac{9}{24} = 0.3750$. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Watch the following video to learn more about sources of sampling bias.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=21#oembed-2>

5. 1.4 Levels of Measurement

Levels of Measurement

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Nominal Scale Level

Data that is measured using a nominal scale is **qualitative**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. Nominal scale data cannot be used in calculations.

Example:

1. To classify people according to their favorite food, like pizza, spaghetti, and sushi. Putting pizza first and sushi second is not meaningful.
 2. Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion.
-

Ordinal Scale Level

Data that is measured using an ordinal scale is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Example:

1. A list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.
 2. A cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured.
-

Interval Scale Level

Data that is measured using the interval scale is similar to ordinal

level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60° . Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Example:

1. Monthly income of 2000 part-time students in Texas
 2. Highest daily temperature in Odessa
-

Ratio Scale Level

Data that is measured using the ratio scale takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. You will not have a negative value in ratio scale data.

For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points) (given that the exams are machine-graded.) The data can be put in order from lowest to highest: 20, 68, 80, 92. There is no negative point in the final exam scores as the lowest score is 0 point.

The differences between the data have meaning. The score 92 is

more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. If one student scores 80 points and another student scores 20 points, the student who scores higher is 4 times better than the student who scores lower.

Example:

1. Weight of 200 cancer patients in the past 5 months
 2. Height of 549 newborn babies
 3. Diameter of 150 donuts
-



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=22#oembed-1>

References

“State & County QuickFacts,” U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

“State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. <http://quickfacts.census.gov/qfd/index.html> (accessed May 1, 2013).

“Table 5: Direct hits by mainland United States Hurricanes

(1851-2004),” National Hurricane Center, <http://www.nhc.noaa.gov/gifs/table5.gif> (accessed May 1, 2013).

“Levels of Measurement,” http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, “Levels of Measurement,” about.com, <http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm> (accessed May 1, 2013).

David Lane. “Levels of Measurement,” Connexions, <http://cnx.org/content/m10809/latest/> (accessed May 1, 2013).

Concept Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

6. 1.5 Frequency & Frequency Tables

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.

The following table lists the different data values in ascending order and their frequencies.

Frequency Table of Student Work Hours	
DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

In this research, 3 students studied for 2 hours. 5 students studies for 3 hours.

A frequency is the number of times a value of the data occurs. According to the table, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A relative frequency is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample—in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

$$\text{Relative frequency} = \frac{\text{frequency of the class}}{\text{total}}$$

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the table below.

$$\text{Cumulative relative frequency} = \text{sum of previous relative frequencies} + \text{current class frequency}$$

Example 1

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies			
DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Example 2

We sample the height of 100 soccer players. The result is shown below.

Height (inches)	Frequency
59.95 – 61.95	5
61.95 – 63.95	3
63.95 – 65.95	15
65.95 – 67.95	40
67.95 – 69.95	17
69.95 – 71.95	12
71.95 – 73.95	7
73.95 – 75.95	1
Total = 100	

Find:
a. the relative frequency for each class.

Show Answer

Height (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
59.95 – 61.95	5	$\frac{5}{100}$ or 0.05	0.05
61.95 – 63.95	3	$\frac{3}{100}$ or 0.03	$0.05 + 0.03 = 0.08$
63.95 – 65.95	15	$\frac{15}{100}$ or 0.15	$0.08 + 0.15 = 0.23$
65.95 – 67.95	40	$\frac{4}{100}$ or 0.04	$0.23 + 0.40 = 0.63$
67.95 – 69.95	17	$\frac{17}{100}$ or 0.17	$0.63 + 0.17 = 0.80$
69.95 – 71.95	12	$\frac{12}{100}$ or 0.12	$0.80 + 0.12 = 0.92$
71.95 – 73.95	7	$\frac{7}{100}$ or 0.07	$0.92 + 0.07 = 0.99$
73.95 – 75.95	1	$\frac{1}{100}$ or 0.01	$0.99 + 0.01 = 1.00$
Total = 100		Total = 1	

b. the percentage for height that is less than 63.95 inches.

Show Answer

$$\frac{5 + 3}{100} = 0.08 = 8\%$$

c. the percentage for height that is between 69.95 inches and 73.95 inches.

Show Answer

$$\frac{12}{100} + \frac{9}{100} = 0.12 + 0.07 = 0.19$$

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example 3

The table shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (inches)	Frequency
2.95 – 4.97	6
4.97 – 6.99	7
6.99 – 9.01	15
9.01 – 11.03	8
11.03 – 13.05	9
13.05 – 15.07	5

Find

1. the relative frequency and cumulative relative frequency for

each class.
Show Answer

Total = sum of all frequencies = 6 + 7 + 15 + 8 + 9 + 5 = 50

Rainfall (inches)	Frequency	Relative frequency	Cumulative relative frequency
2.95 – 4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97 – 6.99	7	$\frac{7}{50} = 0.14$	0.12 + 0.14 = 0.26
6.99 – 9.01	15	$\frac{15}{50} = 0.30$	0.26 + 0.30 = 0.56
9.01 – 11.03	8	$\frac{8}{50} = 0.16$	0.56 + 0.16 = 0.72
11.03 – 13.05	9	$\frac{9}{50} = 0.18$	0.72 + 0.18 = 0.90
13.05 – 15.07	5	$\frac{5}{50} = 0.10$	0.90 + 0.10 = 1.00

2. the percentage of rainfall that is less than 9.01 inches.
Show Answer

The percentage of rainfall that is less than 9.01 inches = 0.12 + 0.14 + 0.30 = 0.56

3. the percentage of heights that fall between 61.95 and 65.95 inches.
Show Answer

The percentage of heights that fall between 6.99 inches and 11.03 inches = $\frac{15}{50} + \frac{8}{50} = 0.26$

Try It

The table contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

- What is the frequency of deaths measured from 2006 through 2009?
Show Answer
97,118
- What percentage of deaths occurred after 2009?
Show Answer
41.6%
- What is the relative frequency of deaths that occurred in 2003 or earlier?
Show Answer

$$\frac{67,092}{823,356} = 0.081$$

4. What is the percentage of deaths that occurred in 2004?

Show Answer

27.8%

5. What kind of data are the numbers of deaths?

Show Answer

Quantitative discrete

6. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Show Answer

Quantitative continuous

Example 4

The table contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

1. What is the frequency of deaths measured from 2000 through 2004?

Show Answer

$$37,526 + 37,862 + 38,491 + 38,477 + 38,444 = 190,800$$

2. What percentage of deaths occurred after 2006?

Show Answer

$$\frac{37,435 + 34,172 + 30,862 + 30,296 + 29,757}{653,782}$$

or 24.9%

3. What is the relative frequency of deaths that occurred in 2000 or before?

Show Answer

$$\frac{260,086}{653,782} \text{ or } 39.8\%$$

4. What is the percentage of deaths that occurred in 2011?

Show Answer

$$\frac{29,757}{653,782} \text{ or } 4.6\%$$

5. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Show Answer

75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

7. 1.6 Experimental Design & Ethics

Experimental Design

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the explanatory variable. The affected variable is called the response variable. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called treatments. An experimental unit is a single object or individual to be measured.

The following video explains the difference between collecting data from observations and collecting data from experiments.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=24#oembed-1>

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called lurking variables. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the random assignment of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*¹

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a control group. This group is given a placebo treatment—a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. Blinding in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A double-blind experiment is one in which both the subjects and the researchers involved with the subjects are blinded.

Sampling

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."³ Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the

chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as Institutional Review Boards (IRB). All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give informed consent. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose

of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

References

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, <http://www.hsph.harvard.edu/nutritionsource/vitamin-e/> (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” [athleteinme.com](http://www.athleteinme.com), <http://www.athleteinme.com/ArticleView.aspx?id=1053> (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443> (accessed May 1, 2013).

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html> (accessed May 1, 2013).

M.L. Jaskon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013),

<http://www.ncbi.nlm.nih.gov/pubmed/22721550> (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. <http://earthquake.usgs.gov/earthquakes/eqarchives/year/> (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. <http://www-fars.nhtsa.dot.gov/Main/index.aspx> (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” <http://www.forbes.com/best-small-companies/list/> (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report> (accessed May 1, 2013).

Lori Alden, “Statistics can be Misleading,” econoclass.com, <http://www.econoclass.com/misleadingstats.html> (accessed May 1, 2013).

Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, <http://cnx.org/content/m15555/latest/> (accessed May 1, 2013).

Concept Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”⁴ Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

PART II

DESCRIPTIVE STATISTICS

8. Introduction: Descriptive Statistics



When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Try It

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **“Descriptive Statistics.”** You will learn how to calculate, and even

more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

NOTE:

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The [Texas Instruments \(TI\) website](#) provides additional instructions for using these calculators.

9. 2.1 Stem-and-Leaf Graphs, Line Graphs, Bar Graphs, and Pie Charts

Stem-and-leaf Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Form a stem-and-leaf graph.

Show Answer

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

The stem-and-leaf plot shows that most scores fell in the 60s, 70s, 80s, and 90s.

Eight out of the 31 scores or approximately 26% were in the 90s or 100, a fairly high number of As.

Try It

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48;
48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.[practice-area
rows="3"][/practice-area]

Show Answer

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

The stemplot is a quick way to graph data and gives an exact picture of the data.

Try It

The following data show the distances (in miles) from the homes of off-campus statistics students to the college.

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0;
2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7;
5.8; 8.0

Create a stem plot using the data and identify any outliers:

Show Answer

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

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Watch this video to see an example of how to create a stem plot.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://library.achievingthedream.org/odessastatistics/?p=27#oembed-1>

Example 2

A side-by-side stem-and-leaf plot allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. The two following tables show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Presidential Age at Death

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Show Answer

Ages at Inauguration

Ages at Inauguration	Ages at Death
9 9 8 7 7 7 6 3 2	4 6 9
8 7 7 7 7 6 6 6 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5 3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6 0 0 3 3 4 4 5 6 7 7 7 8
	7 0 0 1 1 1 4 7 8 8 9
	8 0 1 3 5 8
	9 0 0 3 3

Example 3

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Show Answer

Number of losses

Atalanta Hawks Wins and Leaves	
Number of Wins	Number of Losses
3	1 9
9 8 8 6 5	2 5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3 0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 7 6 6 6 3 3 3 2 2 1 1 0	4 0 0 1 1 2 4 5 6 6 7 7 8 9
7 7 6 3 2 0 0 0 0	5 1 1 1 2 3 4 4 6 7
	6 9

Line Graph

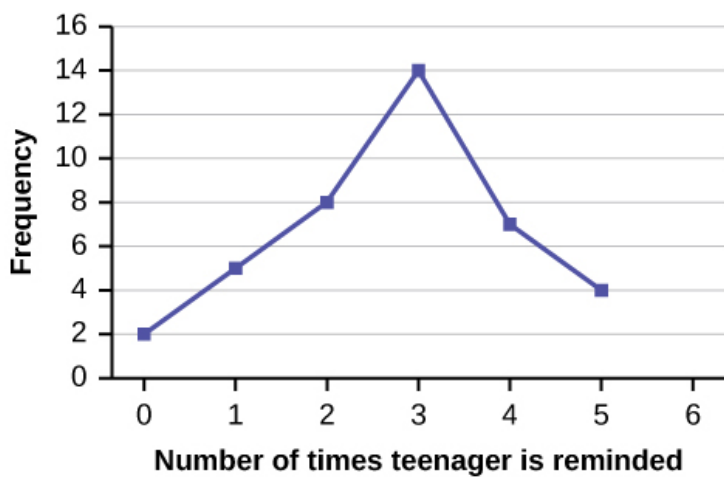
Another type of graph that is useful for specific data values is a **line graph**.

Example 1

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Show Answer



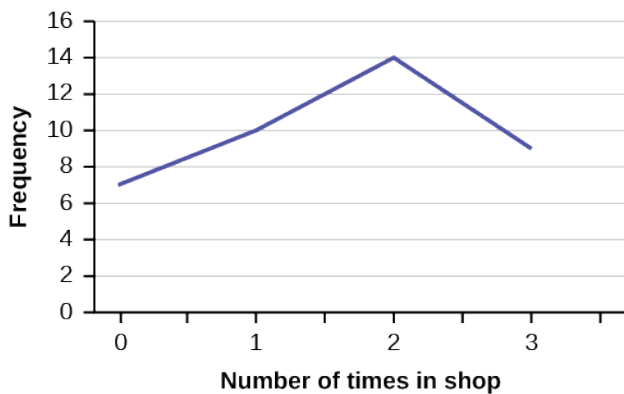
Try It

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Construct a line graph.

Show Answer



Bar Graphs & Pie Charts

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal.

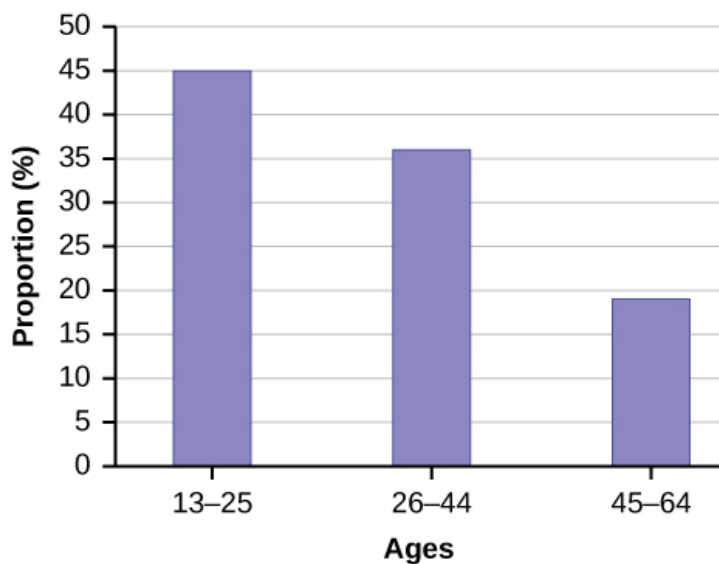
Pie Charts is another method to “visualize” the data. Each proportion on a pie chart represents the proportion of the corresponding class in the data set.

Example 1

By the end of 2011, Facebook had over 146 million users in the United States. The table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph and a pie chart using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

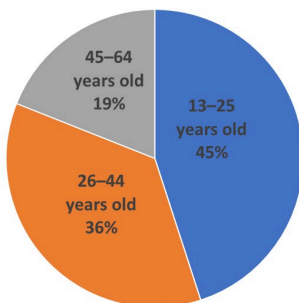
Bar Graph



The **bar graph** has age groups represented on the **x-axis** and proportions on the **y-axis**.

Pie Chart

Facebook users by age group



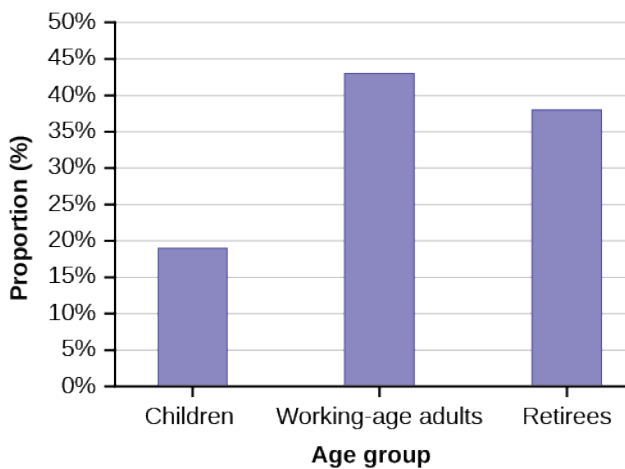
Try It

The population in Park City is made up of children, working-age adults, and retirees. The table shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

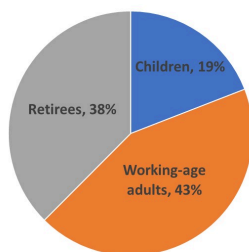
Construct a bar graph and a pie chart showing the proportions.

Bar Graph



Pie Chart

Population in Park City



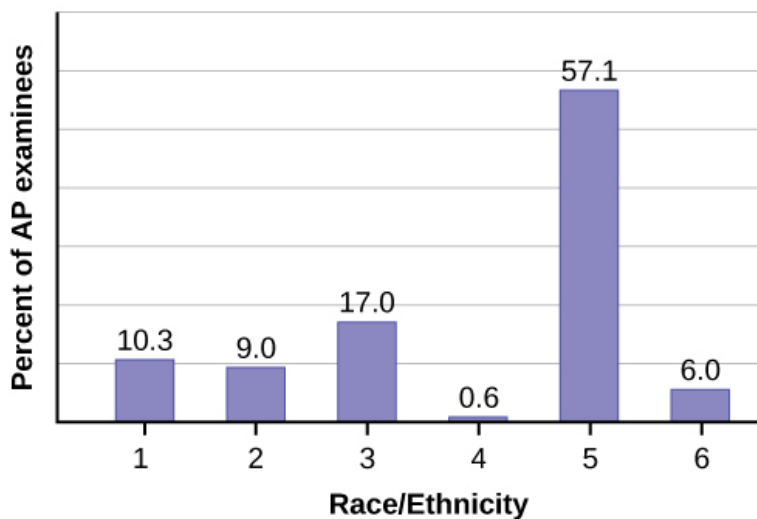
Example 2

The columns in the table contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examinee population for that class, and percentages for the overall student population.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

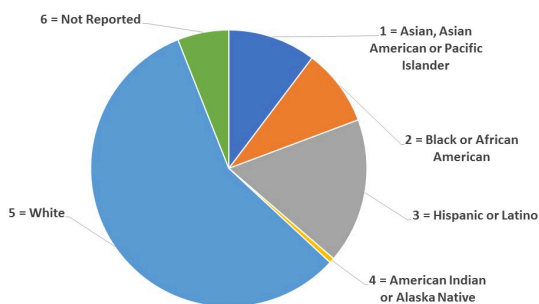
Create a bar graph and a pie chart with the student race or ethnicity (qualitative data) on the x -axis, and the Advanced Placement examinee population percentages on the y -axis.

Show Answer



Pie Chart

The race or ethnicity of students for the Advanced Placement examine population



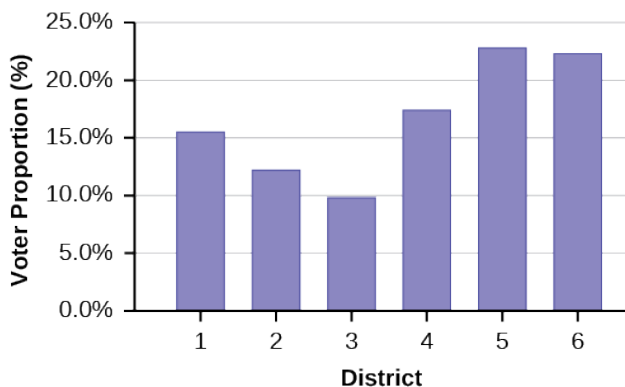
Try It

Park city is broken down into six voting districts. The table shows the percentage of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

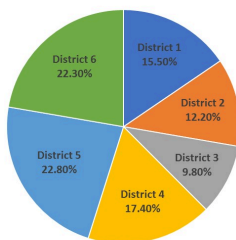
Construct a bar graph and a pie chart that shows the registered voter population by district.

Bar Graph



Pie Chart

The Percentage of Total Registered Voter Population
in Each District



10. 2.2 Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

Histogram

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value

carried out to one more decimal place than the value with the most decimal places.

For example:

1. If the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 ($6.1 - 0.05 = 6.05$). We say that 6.05 has more precision.
2. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 ($1.5 - 0.005 = 1.495$).
3. If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 ($1.0 - 0.0005 = 0.9995$).
4. If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 ($2 - 0.5 = 1.5$).

Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Watch the following video for an example of how to draw a histogram.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=28#oembed-1>

Example 1

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are continuous data, since height is measured.

60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5;
64.566; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5;
66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67;
67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5;
69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5; 74

Construct a relative frequency table and histogram.

Show Answer

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

$60 - 0.05 = 59.95$ which is more precise than, say, 61.5 by one decimal place.

The starting point is, then, 59.95. The largest value is 74, so $74 + 0.05 = 74.05$ is the ending value.

Next, calculate the width of each bar or class interval.

To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire).

Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76$$

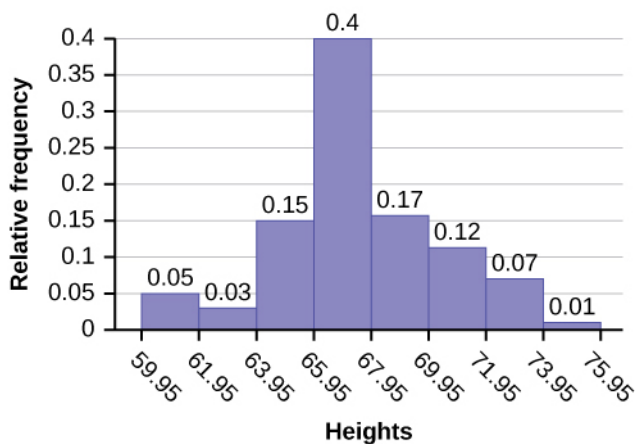
The boundaries are:

- 59.95
- $59.95 + 2 = 61.95$
- $61.95 + 2 = 63.95$
- $63.95 + 2 = 65.95$
- $65.95 + 2 = 67.95$
- $67.95 + 2 = 69.95$
- $69.95 + 2 = 71.95$
- $71.95 + 2 = 73.95$
- $73.95 + 2 = 75.95$

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

Class Interval	Frequency	Relative Frequency
59.95 – 61.95	5	0.05
61.95 – 63.95	3	0.03
63.95 – 65.95	15	0.15
65.95 – 67.95	4	0.04
67.95 – 69.95	17	0.17
69.95 – 71.95	12	0.12
71.95 – 73.95	7	0.07
73.95 – 75.95	1	0.01

The following histogram displays the heights on the x -axis and relative frequency on the y -axis.



Note:

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

Example 2

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

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

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____.

Solution:

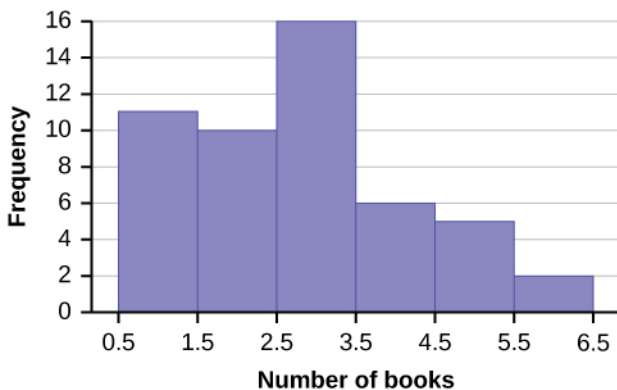
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



Create the histogram for Example 2 by using TI-Calculator:

- Press $Y=$. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xscl = $(6.5 - .5)/6$, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2nd $Y=$. Start by pressing 4:Plotsoff ENTER.
- Press 2nd $Y=$. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Example 3

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Show Answer



Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Practice Problems 1:

The following data are the shoe sizes of 50 male students. The sizes are continuous data and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5

11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5

12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Option 1:

Smallest value: 9

Largest value: 14

Convenient starting value: $9 - 0.05 = 8.95$

Convenient ending value: $14 + 0.05 = 14.05$

$$\frac{14.05 - 8.95}{6} = 0.85$$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Class Interval	Frequency
8.95 – 9.80	4
9.80 – 10.65	14
10.65 – 11.50	13
11.50 – 12.35	14
12.35 – 13.20	4
13.20 – 14.05	1

Option 2:

Smallest value: 9

Largest value: 14

Convenient starting value: $9 - 0.05 = 8.95$

Convenient ending value: $14 + 0.05 = 14.05$

$$\frac{14.05 - 8.95}{6} = 0.85$$

You can also use an interval with a width equal to one.

Frequency Polygons

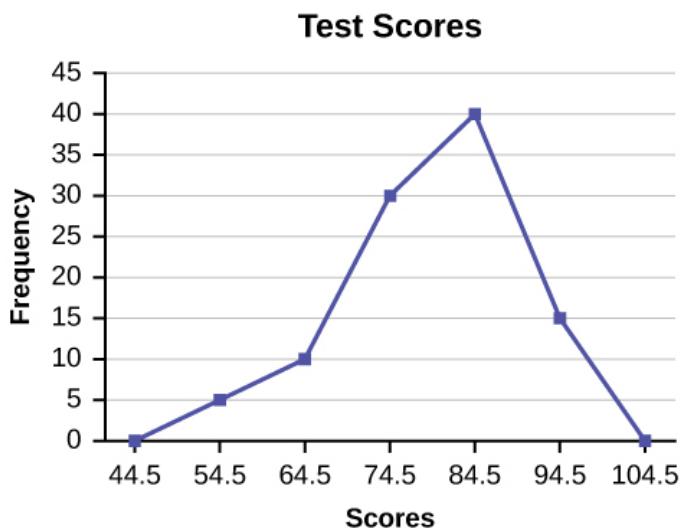
Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x -axis and y -axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 4

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100



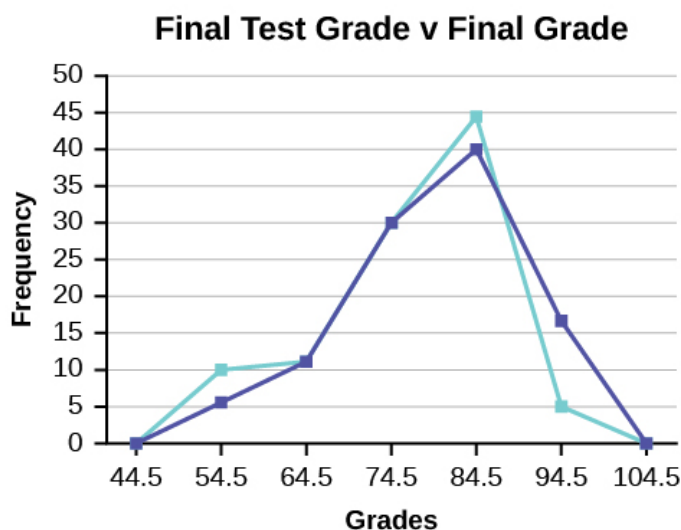
The first label on the x -axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the x -axis. The point labeled 54.5 represents the next interval, or the first “real” interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the x -axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Example 5

We will construct an overlay frequency polygon comparing the scores with the students’ final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades			
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100



Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

Practice Problem 3:

Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in the table.

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Show Answer

Frequency polygons are useful for comparing distributions. This is achieved by overlaying drawn for different data sets.

Constructing a Time Series Graph

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for

the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph. To construct a time series graph, we must look at both pieces of our paired data set. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example 6

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Show Answer

Try It

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO₂ emissions for the United States.

CO2 Emissions			
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Show Answer

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

References

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at <http://www.scholastic.com/teachers/article/timeline-guide-us-presidents> (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at <http://www.fao.org/economic/ess/ess-fs/en/> (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at <http://data.bls.gov/pdq/SurveyOutputServlet> (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at <http://databank.worldbank.org/data/home.aspx> (accessed April 3, 2013).

“Births Time Series Data.” General Register Office For Scotland, 2013. Available online at <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html> (accessed April 3, 2013).

“Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

“Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

Concept Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on y-axis with the frequency being graphed on the x-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

II. 2.3 Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**.

Quartiles are special percentiles.

- The first quartile, Q_1 , is the same as the 25th percentile.
25% of data will be less than 25th percentile; 75% of data will be more than 25th percentile.
- The second quartile, Q_2 , is the same as the 50th percentile / median.
50% of data will be less than 50th percentile; 50% of data will be more than 50th percentile.
- The third quartile, Q_3 , is the same as the 75th percentile.
75% of data will be less than 75th percentile; 25% of data will be more than 75th percentile.

The general form is :

**n % of data will be less than n^{th} percentile and
 $(100\% - n\%)$ of data will be more than n^{th}
percentile.**

The following video gives an introduction to Median, Quartiles and Interquartile Range, the topic you will learn in this section.



One or more interactive elements has been excluded

— from this version of the text. You can view them online here: <https://library.achievingthedream.org/odessastatistics/?p=29#oembed-1>

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220. To be admitted as Duke student, your SAT score has to be at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the “center” of the data. You can think of the median as the “middle value,” but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger.

For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven.

50% of the values are smaller than 7 and 50% of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data.

To find the quartiles,

- Find the median or second quartile (Q_2) first.
 - Find the first quartile (Q_1), the median of the lower half of the data.
 - Find the third quartile (Q_3), the median, of the upper half of the data.
-

To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

- The median or **second quartile** fall between the 7th and 8th data. The median is the mean of 6.8 and 7.2.

$$\text{Hence, the median} = \frac{6.8 + 7.2}{2} = 7$$

- The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.
1; 1; 2; 2; 4; 6; 6.8

The number 2, which is part of the data, is the **first quartile**.

25% of the entire sets of values are the same as or less than

2. 75% of the values are more than 2.

- The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is 9.

The **third quartile**, Q_3 , is 9.

75% of the ordered data set are less than 9. 25% of the ordered data set are greater than 9.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The IQR can help determine outliers.

A data is a potential outlier if and only if the data is

$$\left\{ \begin{array}{l} \text{smaller than } Q_1 - 1.5 * IQR \\ \text{or} \\ \text{larger than } Q_3 + 1.5 * IQR \end{array} \right.$$

Example 1

For the following 11 salaries, calculate the IQR and determine if any salaries are outliers. The salaries are in dollars.

\$33,000 \$64,500 \$28,000 \$54,000 \$72,000 \$68,500
 \$69,000 \$42,000 \$54,000 \$120,000 \$40,500

1. Find the median, Q_1 , and Q_3 .

Show Answer

Order the data from smallest to largest.
 \$28,000
 \$33,000 \$40,500 \$42,000 \$54,000 \$54,000
 \$64,500 \$68,500 \$69,000 \$72,000 \$120,000
 Median
 = \$54,000 $Q_1 = \$40,500$ $Q_3 = \$69,000$

2. Find the interquartile range, IQR.

Show Answer

$$\text{IQR} = \$69,000 - \$40,500 = \$28,500$$

3. Find the potential outlier.

Show Answer

$$\begin{aligned} (1.5)(\text{IQR}) &= (1.5)(\$28,500) = \$42,750 \\ Q_1 - (1.5)(\text{IQR}) &= \$40,500 - \$42,750 = -\$2,250 \\ Q_3 + (1.5)(\text{IQR}) &= \$69,000 + \$42,750 = \$111,750 \end{aligned}$$

No salary is less than \$2,250.

However, \$120,000 is more than \$11,750, so
 \$120,000 is a potential outlier.

Try It

Test Scores for Class A

69; 96; 81; 79; 34; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66;
91; 77; 69; 80; 94

Test Scores for Class B

90; 72; 80; 92; 90; 97; 92; 75; 79; 39; 70; 80; 129; 95; 78;
73; 71; 68; 95; 134

1. Find the interquartile range (IQR) for the following two data sets and compare them.

Show Answer

<div> <div>Class A</div> <div> <p>Order the data from smallest to largest.</p> <p>34; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99</p> <p>$Median = \frac{80 + 81}{2} = 80.5$</p> <p>$Q_1 = \frac{69 + 76}{2} = 72.5$</p> <p>$Q_3 = \frac{90 + 91}{2} = 90.5$</p> <p>$IQR = 90.5 - 72.5 = 18$</p> </div> </div>	<div> <div>Class B</div> <div> <p>Order the data from smallest to largest.</p> <p>39; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92; 95; 95; 97; 129; 134</p> <p>$Median = \frac{80 + 80}{2} = 80$</p> <p>$Q_1 = \frac{72 + 73}{2} = 72.5$</p> <p>$Q_3 = \frac{92 + 95}{2} = 93.5$</p> <p>$IQR = 93.5 - 72.5 = 21$</p> </div> </div>
<div> <p>The data for Class B has a larger IQR, so the scores between Q3 and Q1 (middle 50%) for the data for Class B are more spread out and not clustered about the median.</p> <p>2. Is there any outlier in class A?</p> <p>Show Answer</p> <p>$Q_1 - (1.5)(IQR) = 72.5 - (1.5)(18) = 45.5$</p> <p>$Q_3 + (1.5)(IQR) = 90.5 + (1.5)(18) = 117.5$</p> <p>In class A, we have these data:</p> <p>34; 66; 67; 69; 69; 76; 77; 77; 79; 80; 81; 83; 85; 89; 90; 91; 94; 96; 98; 99.</p> <p>34 is less than $Q_1 - (1.5)(IQR)$, 45.5.</p> </div>	

No data is greater than $Q_3 + (1.5)(IQR) = 117.5$.

So, 34 is the only potential outlier in class A.

3. Is there any outlier in class B?

Show Answer

$$Q_1 - (1.5)(IQR) = 72.5 - (1.5)(21) = 41$$

$$Q_3 + (1.5)(IQR) = 93.5 + (1.5)(21) = 125$$

In class B, we have these data:

39; 68; 70; 71; 72; 73; 75; 78; 79; 80; 80; 90; 90; 92; 92;
95; 95; 97; 129; 134.

39 is less than $Q_1 - (1.5)(IQR)$, 41.

129 and 134 are greater than $Q_3 + (1.5)(IQR)$, 125.

So 39, 129 and 134 are the potential outliers in class B.

Example 2

The following 13 real estate prices. (Prices are in dollars.)

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000;
387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Calculate the IQR and determine if any prices are potential outliers.

Show Answer

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800;
529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

Median = 488,800

Q₁ =

= 308,750

Q₃ =

= 649,000

IQR = 649,000 - 308,750 = 340,250

(1.5)(IQR) = (1.5)(340,250) = 510,375

Q₁ - (1.5)(IQR) = 308,750 - 510,375 = -201,625

Q₃ + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375

No house price is less than -201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential outlier.

A Formula for Finding the k^{th} Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n + 1)$
- If i is an integer, then the k^{th} percentile is the data value in the

i th position in the ordered set of data.

- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 3

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

Amount of Sleep per School Night (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

(a) Find the 28th percentile.

Show Answer

Notice the 0.28 in the “cumulative relative frequency”

column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. The 28th percentile is 6.5.

(b) Find median.

Show Answer

Look again at the “cumulative relative frequency” column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. The median is seven.

(c) Find the third quartile.

Show Answer

The third quartile is the same as the 75th percentile. You can “eyeball” this answer. If you look at the “cumulative relative frequency” column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. The 75th percentile, then, must be an eight. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3 , is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.

Try It

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour).

Amount of time spent on route (hours)	Frequ ncy	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Find the 65th percentile.

[practice-area rows="2"][/practice-area]

Show Answer

The 65th percentile is 3.5. The 65th percentile is between the last three and the first four.

Example 4

Amount of Sleep per School Night (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

1. Find the 80th percentile.

Show Answer

The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th

$$\text{percentile} = \frac{8 + 9}{2} = 8.5$$

2. Find the 90th percentile.

Show Answer

The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.

3. Find the first quartile. What is another name for

the first quartile?

Show Answer

Q_1 is also the 25th percentile. The 25th percentile location calculation:

25th percentile = $0.25(50) = 12.5 \approx 13$ the 13th data value.

Thus, the 25th percentile is six.

Try It

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Find the third quartile. What is another name for the third quartile?

Show Answer

[practice-area rows="2"][/practice-area]

Show Answer

The third quartile is the 75th percentile, which is four.

The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

Example 5

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

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

1. **Find the 70th percentile.**

Show Answer

$k = 70$ th percentile, $i =$ the index, $n = 29$

$$i = \frac{k}{100}(n + 1) = \left(\frac{70}{100}\right)(29 + 1) = 21$$

.

Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64.

The 70th percentile is 64 years.

2. **Find the 83rd percentile.**

Show Answer

k = 83rd percentile, i = the index, n = 29

$$i = \frac{k}{100}(n + 1) = \left(\frac{83}{100}\right)(29 + 1) = 24.9$$

, which is NOT an integer. Round it down to 24 and up to 25.

The age in the 24th position is 71 and the age in the 25th position is 72.

We will find the average of 71 and 72.

The 83rd percentile is 71.5 years.

Try It

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

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

Calculate

- **the 20th percentile**

Show Answer

k = 20. Index =

$$i = \frac{k}{100}(n + 1) = \frac{20}{100}(29 + 1) = 6$$

The age in the sixth position is 27. The 20th percentile is 27 years.

- **the 55th percentile.**

Show Answer

k = 55. Index, i =

$$\frac{k}{100}(n + 1) = \frac{55}{100}(29 + 1) = 16.5.$$

Round down to 16 and up to 17. The age in the 16th position is 52 and the age in the 17th position is 55.

The average of 52 and 55 is 53.5. The 55th percentile is 53.5 years.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.

- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

Example 6

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

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

1. Find the percentile for 58.
2. Find the percentile for 25.

Solution:

1. Counting from the bottom of the list, there are 18 data values less than 58.
There is one value of 58.
Number of data values counting from the bottom of the data list up to but not including the data value 58, $x = 18$

Number of data values equal to the data value 58, $y = 1$

$$\frac{\{x\} + \{0.5\}\{y\}}{\{n\}} \cdot \{100\} = \frac{\{18\} + \{0.5\}\{1\}}{\{29\}} \cdot \{100\} = \{63.80\}.$$

58 is the 64th percentile.

2. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

Number of data values counting from the bottom of the data list up to but not including the data value 25, $x = 18$

Number of data values equal to the data value 25, $y = 1$

$$\frac{\{x\} + \{0.5\}\{y\}}{\{n\}} \cdot \{100\} = \frac{\{3\} + \{0.5\}\{1\}}{\{29\}} \cdot \{100\} = \{12.07\}.$$

25 is the 12th percentile.

Try It

Listed are 30 ages for Academy Award winning best

actors

in order from smallest to largest.

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

Find the percentiles for 47.

Show Answer

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

$$.matrixx = 15 \quad \text{and} \quad y = 1 \frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(1)}{29}(100) = 53.45.$$

47 is the 53rd percentile.

Find the percentiles for 31.

Show Answer

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

$$.matrixx = 15 \quad \text{and} \quad y = 2 \frac{x + 0.5y}{n}(100) = \frac{15 + 0.5(2)}{29}(100) = 31.03.$$

31 is the 31st percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when

data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the p th percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is “good” or “bad.” The interpretation of whether a certain percentile is “good” or “bad” depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered “good;” in other contexts a high percentile might be considered “good”. In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Guideline

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
 - the data value (value of the variable) that represents the percentile
 - the percent of individuals or items with data values below the percentile
 - the percent of individuals or items with data values above the percentile.
-
-

Example 7

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes.

Interpret the first quartile in the context of this situation.

Solution:

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable.
(If you take too long, you might not be able to finish.)

Try It

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Show Answer

25% of runners finished the race in 11.5 seconds or more. 75% of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

Example 8

On a 20 question math test, the 70th percentile for

number of correct answers was 16.

Interpret the 70th percentile in the context of this situation.

Solution:

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It

On a 60 point written assignment, the 80th percentile

for the number of points earned was 49.

Interpret the 80th percentile in the context of this situation.

Show Answer

80% of students earned 49 points or fewer. 20% of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example 9

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units.

Interpret the 30th percentile in the context of this situation.

Solution:

- 30% of students are enrolled in 7 or fewer credit units.
- 70% of students are enrolled in 7 or more credit units.
- In this example, there is no “good” or “bad” value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Try It

During a season, the 40th percentile for points scored

per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Show Answer

40% of players scored eight points or fewer. 60% of players scored eight points or more.

A higher percentile is good because getting more points in a basketball game is desirable.

Example 10

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q1 = 20
- Med = 40
- Q₃ = 60
- Max = 300

Solution:

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the IQR is 40 minutes ($60 - 20 = 40$), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful.

The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120.$$

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q1 = 20
- Q3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

Concept Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile, the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the 75th percentile. The interquartile range, or IQR, is the range of the middle 50 percent of the data values. The IQR is found by subtracting Q_1 from Q_3 , and can help determine outliers by using the following two expressions.

- $Q_3 + \text{IQR}(1.5)$
- $Q_1 - \text{IQR}(1.5)$

Formula Review

$$i = \left(\frac{k}{100} \right) (n + 1) \text{ where}$$

i = the ranking or position of a data value,

k = the k th percentile,

n = total number of data.

Expression for finding the percentile of a data value:

$$\left(\frac{x + 0.5y}{n} \right) (100)$$

where

x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data

References

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1> (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/> (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/main/www/cen1990.html> (accessed April 3, 2013).

Data from
San Jose Mercury News.

Data from
Time Magazine; survey by Yankelovich Partners, Inc.

12. 2.4 Box Plots

Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile (Q_1), the median (Q_2), the third quartile (Q_3), and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box**. The “whiskers” extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

Note:

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile, Q_1 is 2.

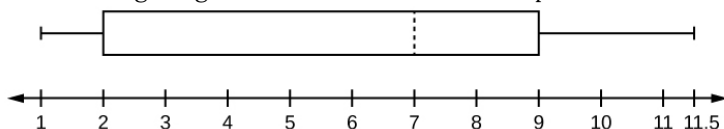
The median, Q_2 is 7.

The third quartile, Q_3 is 9.

The smallest value is 1.

The largest value is 11.5.

The following image shows the constructed box plot.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

Note:

It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

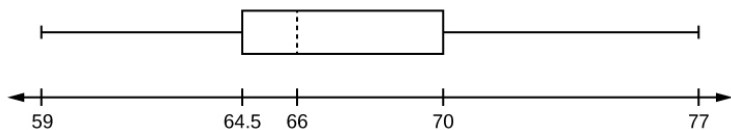
Example 1

The following data are the heights of 40 students in a statistics class.

59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67
67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q_1 , First quartile = 64.5
- Q_2 , Second quartile or median = 66
- Q_3 , Third quartile = 70



1. Each quarter has approximately 25% of the data.
2. The spreads of the four quarters are $64.5 - 59 = 5.5$ (first quarter), $66 - 64.5 = 1.5$ (second quarter), $70 - 66 = 4$ (third quarter), and $77 - 70 = 7$ (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
3. Range = maximum value - the minimum value = $77 - 59 = 18$
4. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
5. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
6. The middle 50% (middle half) of the data has a range of 5.5 inches.

TI-Calculator:

To find the minimum, maximum, and quartiles:

- Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.
- Put the data values into the list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
- Press ENTER.
- Use the down and up arrow keys to scroll.
Smallest value = 59.
Largest value = 77.
Q1: First quartile = 64.5.
Q2: Second quartile or median = 66.
Q3: Third quartile = 70.

To construct the box plot:

- Press 4:Plotsoff. Press ENTER.
- Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.
- Arrow down to Xlist: Press 2nd 1 for L1
- Arrow down to Freq: Press ALPHA. Press 1.
- Press Zoom. Press 9: ZoomStat.
- Press TRACE, and use the arrow keys to examine the box plot.

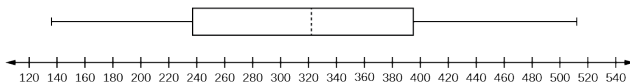
Try It

The following data are the number of pages in 40 books on a shelf.

136 140 178 190 205 215 217 218 232 234 240 255 270 275
290 301 303 315 317 318 326 333 343 349 360 369 377 388
391 392 398 400 402 405 408 422 429 450 475 512

Construct a box plot using a graphing calculator, and state the interquartile range.

Show Answer



IQR = 158

This video explains what descriptive statistics are needed to create a box and whisker plot.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://library.achievingthedream.org/odessastatistics/?p=30#oembed-1>

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



In this case, at least 25% of the values are equal to one. 25% of the values are between one and five, inclusive.

At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example 2

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

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

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

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

1. Find the smallest and largest values, the median, and the first and third quartile for the day class.

Show Answer

Min = 32

$Q_1 = 56$

$Q_2 = 74.5$

$Q_3 = 82.5$

Max = 99

2. Find the smallest and largest values, the median, and the first and third quartile for the night class.

Show Answer

Min = 25.5

$Q_1 = 78$

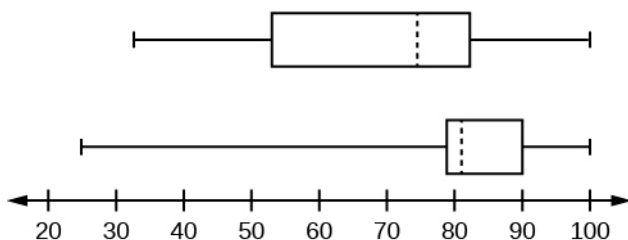
$Q_2 = 81$

$Q_3 = 89$

Max = 98

3. Create a box plot for each set of data. Use one number line for both box plots.

Show Answer



4. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Show Answer

The first data set has the wider spread for the middle 50% of the data. The IQR for the first data set is greater than the IQR for the second set. This means that there is more variability in the middle 50% of the first data set.

Try It

Try It

The following data set shows the heights in inches for the boys in a class of 40 students.

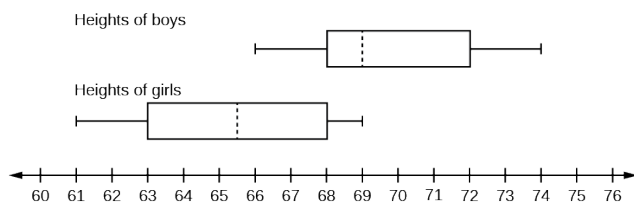
66; 66; 67; 67; 68; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72;
72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68;
68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Show Answer



IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

Example 3

Graph a box-and-whisker plot for the data values shown.

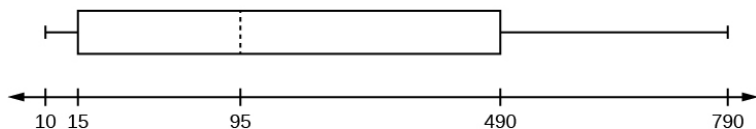
10 10 10 15 35 75 90 95 100 175 420 490 515 515 790

Show Answer

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q1: 15
- Med: 95
- Q3: 490
- Max: 790

The following graph shows the box-and-whisker plot.



Try It

Try It

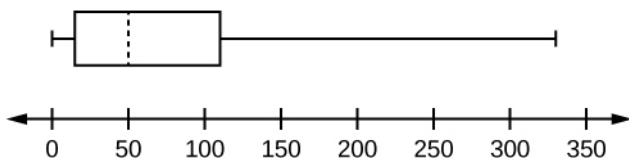
Graph a box-and-whisker plot for the data values shown.

0 5 5 15 30 30 45 50 50 60 75 110 140 240 330

Show Answer

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q1: 15
- Med: 50
- Q3: 110
- Max: 330



References

Data from *West Magazine*.

Concept Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Additional Resources

Use the [online imathAS box plot tool](#) to create box and whisker plots.

13. 2.5 Measures of the Center of the Data

Mean & Median

The “center” of a data set is also a way of describing location. The two most widely used measures of the “center” of the data are the mean (average) and the median. To calculate the mean weight of 50 people, add the 50 weights together and divide by 50. To find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (read “ x bar”): \bar{x} .

The Greek letter μ (pronounced “mew”) represents the population mean. One of the requirements for the sample mean to be a good

estimate of the population mean is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

$$\begin{aligned}\bar{x} &= \frac{1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4}{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4} = 2.7 \\ \bar{x} &= \frac{3(1) + 2(2) + 1(3) + 5(4)}{11} = 2.7\end{aligned}$$

In the second example, the frequencies are $3(1) + 2(2) + 1(3) + 5(4)$.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered.

For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are not the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25;
26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

Calculate the mean and the median.

Mean

The calculation for the mean is

$$\bar{x} = \frac{[3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + \dots + 35 + 37 + 40 + (44)(2) + 47]}{40} = 23.6$$

Median

To find the median, M, first use the formula for the location. The location is: $\frac{n + 1}{2} = \frac{40 + 1}{2} = 20.5$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25;
26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24 + 24}{2} = 24$$

Finding the Mean and the Median Using the TI-83, 83+, 84, 84+ Calculator

1. Clear list L1. Press STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.
2. Enter data into the list editor. Press STAT 1:EDIT.
3. Put the data values into list L1.
4. Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
5. Press the down and up arrow keys to scroll.
6. \bar{x} = 23.6, M = 24

Try It

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest.

3 4 5 7 7 7 7 8 8 9 9 10 10 10 10 11 12 12 13 14 14 15 15 17
17 18 19 19 19 21 21 22 22 23 24 24 24 24

Calculate the mean and median.

Show Answer

Mean: $3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 +$
 $10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 +$
 $18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 = 544$

$$\frac{544}{39} = 13.95$$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example 2

Suppose that in a small town of 50 people, one person earns

\$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the “center”: the mean or the median?

Show Answer

$$\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129400,$$

$$M = 30000$$

(There are 49 people who earn \$30,000 and one person who earns \$5,000,000.)

The median is a better measure of the “center” than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Try It

Try It

In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the “center”: the mean or the median?

Show Answer

The median is the better measure of the “center” than the mean because 59 of the values are \$280,000 and one is \$2,500,000. The \$2,500,000 is an outlier. Either \$280,000 or \$315,000 gives us a better sense of the middle of the data.

Mode

Another measure of the center is the mode. The mode is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 3

Statistics exam scores for 20 students are as follows:

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

Find the mode.

Show Answer

The most frequent score is 72, which occurs five times. Mode = 72.

Try It

The number of books checked out from the library from 25 students are as follows:

0, 0, 0, 1, 2, 3, 3, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 8, 9, 10, 10, 11, 11, 12, 12

Find the mode.

Show Answer

The most frequent number of books is 7, which occurs four times. Mode = 7.

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

6 credit scores are 590, 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice.

When is the mode the best measure of the “center”? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that

most people lose two pounds the first week, making the program less appealing.

Note:

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Example 4

Consider the annual earnings of workers at a factory. The mode is \$25,000 and occurs 150 times out of 301. The median is \$50,000 and the mean is \$47,500. What would be the best measure of the “center”?

Show Answer

Because \$25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

Watch the following video from Kahn Academy on finding the mean, median and mode of a set of data.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=31#oembed-1>

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean \bar{x} of the sample is very likely to get closer and closer to μ . This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a sampling distribution as a relative frequency distribution with a great many samples. Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the relative frequency table shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A statistic is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean is an example of a statistic which estimates the population mean μ .

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What

we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$\text{mean} = \frac{\text{data sum}}{\text{number of data values}}.$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary} + \text{upper boundary}}{2} \quad \text{where } f = \text{the}$$

frequency of the interval and m = the midpoint of the interval.

Example 5

A frequency table displaying professor Blount's last statistic test is shown.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Find the best estimate of the class mean.

Show Answer

Find the midpoints for all intervals.

Grade Interval	Midpoint
50.0–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Calculate the sum of the product of each interval frequency and midpoint.

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.2$$

$$\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$$

Try It

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Show Answer

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5, 17.5. Mean =
 $(1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9) = 409.75$

Review

The mean and the median can be calculated to help you find the “center” of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$$\mu = \frac{\sum fm}{\sum f}$$

Where f = interval frequencies and m = interval midpoints.

References

Data from The World Bank, available online at <http://www.worldbank.org> (accessed April 3, 2013).

“Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).

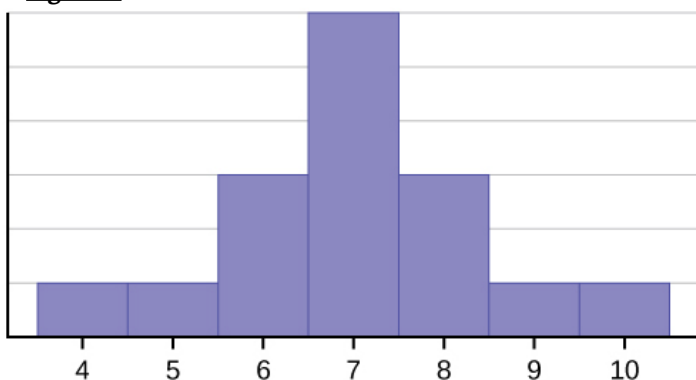
14. 2.6 Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

Figure 1.



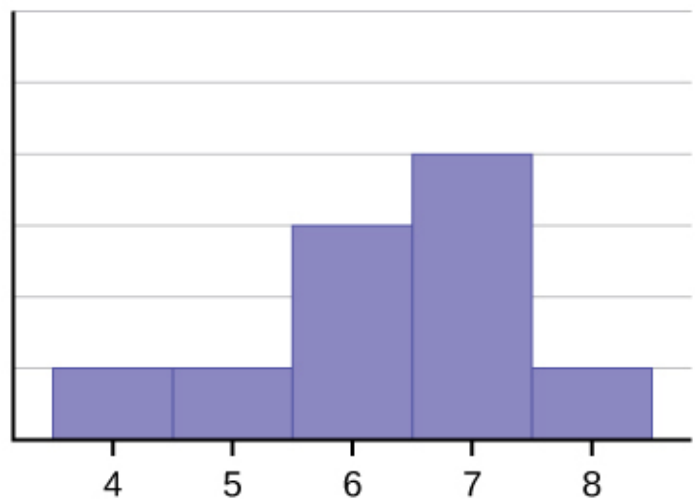
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

Consider the following data set: 4 5 6 6 6 7 7 7 7 8.

The histogram is not symmetrical.

The right-hand side seems “chopped off” compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

Figure 2.

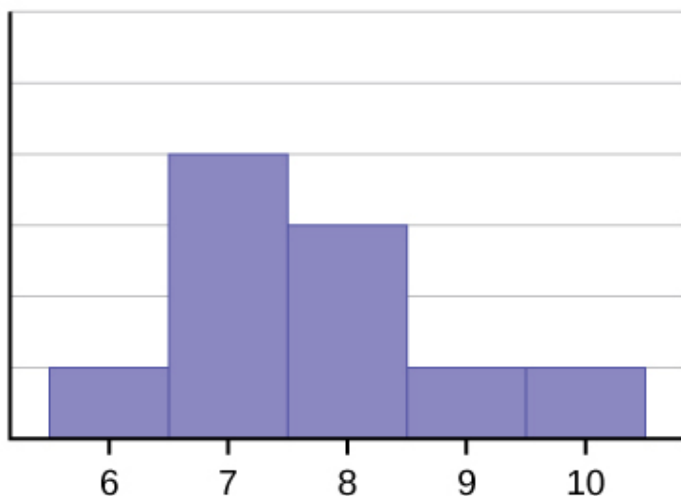


The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

Consider the following data set: 6 7 7 7 7 8 8 8 9 10.

The histogram is also not symmetrical. It is **skewed to the right**.

Figure 3.



The mean is 7.7, the median is 7.5, and the mode is 7.

Of the three statistics, **the mean is the largest, while the mode is the smallest.** Again, the mean reflects the skewing the most.

To summarize,

- if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode.
(mean < median < mode)
- If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.
(mode < median < mean)
- If the distribution of data is symmetric, the mode = the median = the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Here is a video that summarizes how the mean, median and mode can help us describe the skewness of a dataset. Don't worry about the terms leptokurtic and platykurtic for this course.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=32#oembed-1>

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2

Davi: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1

Mari: 2; 3; 4; 4; 4; 6; 6; 8; 3

1. Make a dot plot for the three authors and compare the shapes.
Show Answer

Terry's distribution is right-skewed.

Davi's distribution is slightly left-skewed.

Mari's distribution is symmetrical.

2. Calculate the mean for each.
Show Answer

Terry's mean is 3.7, Davi's mean is 2.7, Mari's mean is 4.6.

3. Calculate the median for each.
Show Answer

Terry's median is three, Davi's median is three. Mari's median is four.

4. Describe any pattern you notice between the shape and the measures of center.

Show Answer

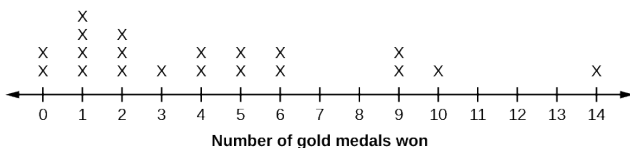
It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Try It

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

2010 Winter Olympics Gold Medal Wins by Top 20 Medal-Winning Countries



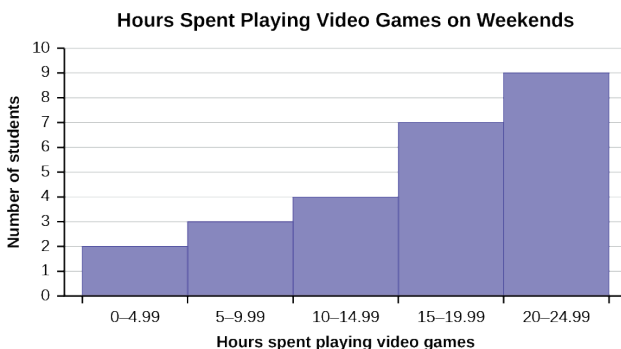
b.

The Ages Former U.S Presidents Died

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

Key: 8|0 means 80.

c.



Show Answer

- mean = 4.25, median = 3.5, mode = 1; The mean > median > mode which indicates skewness to the right. (data are 0, 1, 2, 3, 4, 5, 6, 9, 10, 14 and respective frequencies are 2, 4, 3, 1, 2, 2, 2, 2, 1, 1)
- mean = 70.1, median = 68, mode = 57, 67
bimodal; the mean and median are close but there is a little skewness to the right which is influenced

by the data being bimodal. (data are 46, 49, 53, 56, 57, 57, 57, 58, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 67, 68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 79, 80, 81, 83, 85, 88, 90, 90 93, 93).

3. These are estimates: mean =16.095, median = 17.495, mode = 22.495 (there may be no mode); The mean < median < mode which indicates skewness to the left. (data are the midpoints of the intervals: 2.495, 7.495, 12.495, 17.495, 22.495 and respective frequencies are 2, 3, 4, 7, 9).

Concept Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like Figure 2. A **left (or negative) skewed** distribution has a shape like Figure 3 . A **symmetrical** distribution looks like Figure 1.

15. 2.7 Measures of the Spread of Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation provides a numerical measure of the overall amount of variation in a data set, and can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set.

The standard deviation is always positive or zero.

- The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread.
- The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. the average wait time at both supermarkets is 5 minutes.

At supermarket A, the standard deviation for the wait time is 2 minutes.

At supermarket B, the standard deviation for the wait time is 4 minutes.

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket A. Rosa waits at the checkout counter for 7 minutes and Binh waits for 1 minute. At supermarket A, the mean waiting time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

- 7 minutes is 2 minutes longer than the average of 5 minutes; 2 minutes is equal to 1 standard deviation.
- Rosa's wait time of 7 minutes is **two minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 minute is 4 minutes less than the average of 5 minutes; 4 minutes is equal to 2 standard deviations.

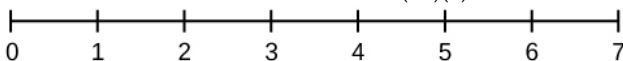
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
 - Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
-
-

A data value that is 2 standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate “rule of thumb” than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that

7 is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of 5 because $5 + (-2)(2) = 1$.



In general, **a value = mean + (# ofSTDEV)(standard deviation)**

- where # of STDEV = the number of standard deviations
- # of STDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation **a value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- Sample: $x = \bar{x} + (\# \text{ of STDEV})(s)$
- Population: $x = \mu + (\# \text{ of STDEV})(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol \bar{x} is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference “ $x - \text{mean}$ ” is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation.

If the numbers belong to a population, in symbols a deviation is $x - \mu$.

For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the**

deviations (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population).

The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n - 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad \text{or} \quad s = \sqrt{\frac{\sum (f)(x - \bar{x})^2}{n - 1}}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size - 1.

Formulas for the Population Standard Deviation

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

For the population standard deviation, the denominator is N , the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

In the following video an example of calculating the variance and standard deviation of a set of data is presented.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=33#oembed-1>

Sampling Variability of a Statistic

How much the statistic varies from one sample to another is known as the sampling variability of a statistic. You typically measure the sampling variability of a statistic by its standard error. The standard error of the mean is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean when you learn about The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note:

In practice, use a calculator or computer software to calculate the standard deviation. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 1

The following data are the ages for a sample of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

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

1. The teacher was interested in the average age and the sample standard deviation of the ages of her students.

Show Answer

The average age is 10.525 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s .

Data	Freq.	Deviations	Deviations ²	(Freq.)(Deviations ²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times 0.950625 = 2.851875$
The total is 9.7375				

The sample variance, s^2 , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1): $s^2 = \frac{9.7375}{20 - 1} = 0.5125$

The **sample standard deviation** s is equal to the square root of the sample variance: which is rounded to two decimal places, $s = \sqrt{0.5125} = 0.72$.

$$\bar{x} = 9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)20 = 10.525$$

2. Find the value that is one standard deviation above the mean.

(Find $\bar{x} + 1s$.)

Show Answer

$$\bar{x} + 1s = 10.53 + (1)(0.72) = 11.25$$

3. Find the value that is two standard deviations below the mean.

(Find $\bar{x} - 2s$.)

Show Answer

$$\bar{x} - 2s = 10.53 - (2)(0.72) = 9.09$$

4. Find the values that are 1.5 standard deviations from (below and above) the mean.

Show Answer

$$(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$$

$$(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

For the following problems, recall that **value = mean + (# of STDEVs)(standard deviation)**.

Verify the mean and standard deviation on a calculator or computer.

For a sample: $x = \bar{x} + (\# \text{ of STDEV})(s)$

For a population: $x = \mu + (\# \text{ of STDEV})(\sigma)$

For this example, use $x = \bar{x} + (\# \text{ of STDEV})(s)$ because the data is from a sample.

Verify the mean and standard deviation on your calculator or computer:

- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT.
If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2.
Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2).
Do not forget the comma. Press ENTER.
- $\bar{x} = 10.525$
- Use Sx because this is sample data (not a population):
 $Sx = 0.715891$.

Try It

On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Show Answer

$$\mu = 30.68$$

$$s = 6.09$$

$$\overline{x} + 2s = 30.68 + (2)(6.09) = 42.86.$$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value

11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is -1.525 for the data value nine. **If you add the deviations, the sum is always zero.** (For Example 1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

Note:

Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and

is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better “feel” for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example 2

Use the following data (first exam scores) from Susan Dean’s spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

1. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.

Show Answer

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

2. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:

- The sample mean

Show Answer

The sample mean = 73.5

- The sample standard deviation

Show Answer

The sample standard deviation = 17.9

- The median

Show Answer

The median = 73

- The first quartile

Show Answer

The first quartile = 61

- The third quartile

Show Answer

The third quartile = 90

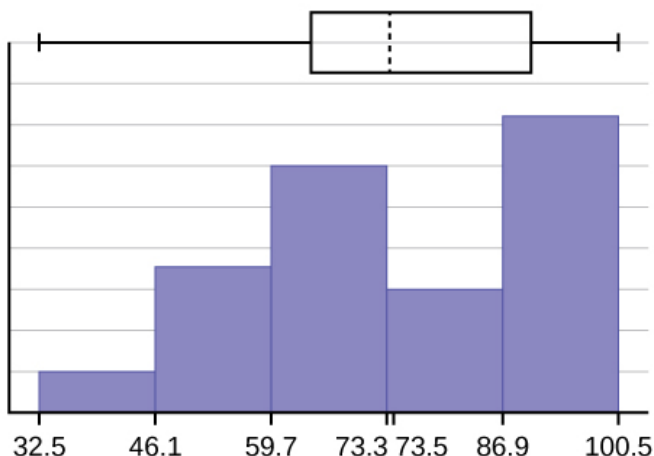
- IQR

Show Answer $IQR = 90 - 61 = 29$

3. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Show Answer

The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5$ = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ($IQR = 29$) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Try It

The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 8; 9; 9; 9; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Show Answer

$$\mu = 9.3$$

$$s = 2.2$$

Standard Deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

Mean of Frequency Table, $\bar{x} = \frac{\sum(fm)}{\sum(f)}$

Standard Deviation of Frequency Table,

$$s_x = \sqrt{\frac{f(m - \bar{x})^2}{n - 1}}$$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example 3

Class	Frequency
0 – 2	1
3 – 5	6
6 – 8	10
9 – 11	7
12 – 14	0
15 – 17	2

Find the standard deviation for the data in the table.

Show Answer

Steps 1: Find the mean.

Class	Frequency, f	Midpoint, m
0-2	1	1
3-5	6	4
6-8	10	7
9-11	7	10
12-14	0	13
15-17	2	16

For this data set, we have the mean, $\bar{x} = \frac{\sum fm}{n} = \frac{(1)(1) + (6)(4) + (10)(7) + (7)(10) + (0)(13) + (2)(16)}{26} = 7.58$.

Step 2: Use the mean to find the standard deviation.

Class	Frequency, f	Midpoint, m	$m - \bar{x}$	$(m - \bar{x})^2$	$(f)(m - \bar{x})^2$
0-2	1	1	$1 - 7.58 = -6.58$	$(-6.58)^2 = 43.2964$	$(1)(43.296) = 43.2964$
3-5	6	4	$4 - 7.58 = -3.58$	$(-3.58)^2 = 12.8164$	$(6)(12.816) = 76.8984$
6-8	10	7	$7 - 7.58 = -0.58$	$(-0.58)^2 = 0.3364$	$(10)(0.3364) = 3.364$
9-11	7	10	$10 - 7.58 = 2.42$	$(2.42)^2 = 5.8564$	$(7)(5.8564) = 40.9948$
12-14	0	13	$13 - 7.58 = 5.42$	$(5.42)^2 = 29.3764$	$(0)(29.376) = 0$
15-17	2	16	$16 - 7.58 = 8.42$	$(8.42)^2 = 70.8964$	$(2)(70.896) = 141.7928$

The standard deviation,

$$s_x = \sqrt{\frac{\sum (f)(m - \bar{x})^2}{n - 1}} = \sqrt{\frac{43.2964 + 76.8984 + 3.364 + 40.9948 + 0 + 141.7928}{26 - 1}} = \sqrt{\frac{306.3464}{25}} = \sqrt{12.2539} = 3.5$$

Mean of Frequency Table, $\bar{x} = \frac{\sum(fm)}{\sum(f)}$

Standard Deviation of Frequency Table, $s_x = \sqrt{\frac{f(m - \bar{x})^2}{n - 1}}$

where f = interval frequencies and m = interval midpoints.

The calculations are tedious. **It is usually best to use technology when performing the calculations.**

Try It

Find the standard deviation for the data from the previous example

Class	Frequency, f
0-2	1
3-5	6
6-8	10
9-11	7
12-14	0
15-17	2

Use TI-Calculators to find the standard deviation:

```

EDIT  CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

```

Input the midpoint values into **L1** and the frequencies into **L2**

L1	L2	L3	2
1	1	-----	
4	6		
7	10		
10	7		
13	0		
16	2		
-----	-----		
L2(7) =			

Select **STAT**, **CALC**, and **1: 1-Var Stats**

```

EDIT [2nd] [1] TESTS
1: 1-Var Stats
2: 2-Var Stats
3: Med-Med
4: LinReg(ax+b)
5: QuadReg
6: CubicReg
7: QuartReg

```

Select **2nd** then **1** then , **2nd** then **2** Enter

```

1-Var Stats
 $\bar{x}$  = 7.576923077
 $\Sigma x$  = 197
 $\Sigma x^2$  = 1799
 $Sx$  = 3.500549407
 $\sigma x$  = 3.432571103
↓ n = 26

```


You will see displayed both a population standard deviation, σ_x , and the sample standard deviation, s_x .

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: $\text{value} = \text{mean} + (\text{\#ofSTDEVs})(\text{standard deviation})$; solve for \#ofSTDEVs .
- $\text{\#ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

\#ofSTDEVs is often called a "z-score"; we can use the symbol z . In symbols, the formulas become:

	Data value, x	Z-Score of data value
Sample	$x = \bar{x} + (z)(s)$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + (z)(\sigma)$	$z = \frac{x - \mu}{\sigma}$

Example 4

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school.

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Which student had the highest GPA when compared to his school?
Show Answer

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \# \text{ of STDEVs} = \frac{x - \mu}{\sigma}$$

John	Ali
$z = \frac{2.85 - 3.00}{0.7} = -0.21$	$z = \frac{77 - 80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of -0.21 is higher than Ali's z-score of -0.3 .

For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Try It

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Show Answer

For Angie: $z = \frac{26.2 - 27.2}{0.8} = \frac{-1}{0.8} = -1.25$

$$\text{For Beth: } z = \frac{27.3 - 30.1}{1.4} = \frac{-2.8}{1.4} = -2$$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- 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.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

Concept Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.

- $s_x = \sqrt{\frac{f(m - \bar{x})^2}{n - 1}}$ is the formula for calculating the standard deviation of a sample.

- To calculate the standard deviation of a population, we would use the population mean, μ , and the formula

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

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2}$$

where s_x =sample standard deviation, \bar{x} = sample mean

16. 2.8 When to use each measure of Central Tendency

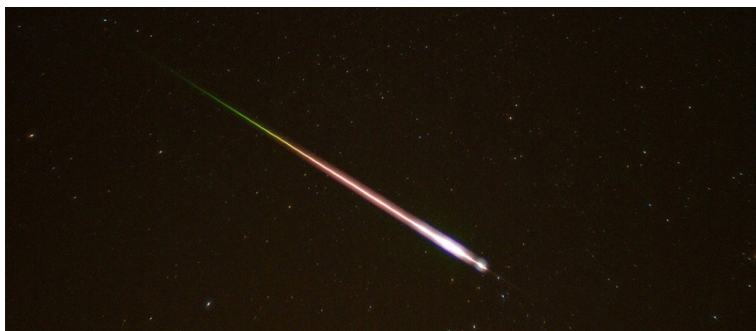
By now, everyone should know how to calculate mean, median and mode. They each give us a measure of Central Tendency (i.e. where the center of our data falls), but often give different answers. So how do we know when to use each? Here are some general rules:

1. Mean is the most frequently used measure of central tendency and generally considered the best measure of it. However, there are some situations where either median or mode are preferred.
2. Median is the preferred measure of central tendency when:
 1. There are a few extreme scores in the distribution of the data. (NOTE: Remember that a single outlier can have a great effect on the mean).
 2. There are some missing or undetermined values in your data.
 3. There is an open ended distribution (For example, if you have a data field which measures number of children and your options are 0, 1, 2, 3, 4, 5 or “6 or more,” then the “6 or more field” is open ended and makes calculating the mean impossible, since we do not know exact values for this field).
 4. You have data measured on an ordinal scale.
3. Mode is the preferred measure when data are measured in a nominal (and even sometimes ordinal) scale.

PART III

PROBABILITY

17. Introduction: Probability Topics



Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Learning Objectives

By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to “guess” about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

COLLABORATIVE EXERCISE

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered “yes” to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

18. 3.1 The Terminology of Probability

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance experiment**.

Example of an experiment: Flipping one fair coin twice.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space.

Example: if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head.

Example: The probability of an event A is probability of getting at most one head. It is also written as $P(A)$.

The **probability** of any outcome is the **long-term relative frequency**

of that outcome. **Probabilities are between zero and one, inclusive** (that is, $0 \leq \text{probability of an event} \leq 1$).

- $P(A) = 0$ means the event A can never happen.
- $P(A) = 1$ means the event A always happens.
- $P(A) = 0.5$ means the event A is equally likely to occur or not to occur.

Example: If you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability.

Example:

1. If you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face.
 2. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur.
 3. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.
-

To calculate the probability of an event A when all outcomes in the sample space are equally likely, count the number of outcomes for

event A and divide by the total number of outcomes in the sample space.

For examples:

- If you toss a fair dime and a fair nickel, you will see four possible outcomes. These 4 outcomes will form a sample space. Therefore, the sample space is {HH, TH, HT, TT} where T = tails and H = heads.
If event A = getting one head, then there are two outcomes that meet this condition {HT, TH}.

$$\text{The probability of event A, } P(A) = \frac{\text{number of outcome with only one head}}{\text{total of possible outcomes}} = \frac{2}{4} = 0.5.$$

- Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event E = rolling a number that is at least five. There are two outcomes {5, 6}.

$$P(E) = \frac{\text{number of outcome that rolling a number that is at least five}}{\text{total of possible outcomes}} = \frac{2}{6} \text{ as the number of repetitions grows larger and larger.}$$

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term

observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

This video gives more examples of basic probabilities.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=37#oembed-1>

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**.

Examples:

1. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias.
2. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are

not equally likely.

“OR” Event

An outcome is in the event $A \text{ OR } B$ if the outcome is in A or is in B or is in both A and B .

For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$.

$A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. (Notice that 4 and 5 are NOT listed twice.)

“AND” Event

An outcome is in the event $A \text{ AND } B$ if the outcome is in both A and B at the same time.

For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$, respectively.

Then

$A \text{ AND } B = \{4, 5\}$.

Complimentary Event

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A.

$$P(A) + P(A') = 1.$$

For example:

We have sample space $S = \{1, 2, 3, 4, 5, 6\}$.

If event $A = \{1, 2, 3, 4\}$. Then, event $A' = \{5, 6\}$.

$$P(A) = \frac{4}{6} \text{ and } P(A') = \frac{2}{6}$$

$$P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$$

Conditional Probability of an Event

The **conditional probability** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space.** We calculate the probability of A from the reduced sample space B. The formula to calculate $P(A|B)$ is

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} \text{ where } P(B) \text{ is greater than zero.}$$

For example:

Suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$.

Let event A = face is 2 or 3 and B = event that face is even.

Event A = {2, 3}, Event B = {2, 4, 6}.

To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space B = {2, 4, 6}. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

A and B = {2} (as 2 appears in both event A and event B.)

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{\text{the number of outcomes that are in both event A and event B}}{\text{total outcomes}}}{\frac{\text{the number of outcomes in event B}}{\text{total outcomes}}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 1

The sample space S is the whole numbers starting at one and less than 20.

1. S = _____ Let event A = the even numbers and event B = numbers greater than 13.

Show Answer

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$$

2. $A = \text{-----}, B =$

Show Answer

$$A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, B = \{14, 15, 16, 17, 18, 19\}$$

3. $P(A) = \text{-----}, P(B) = \text{-----}$

Show Answer

$$P(A) = \frac{9}{19}, P(B) = \frac{6}{19}$$

4. $A \text{ AND } B = \text{-----}, A \text{ OR } B =$

Show Answer

$$A \text{ AND } B = \{14, 16, 18\}, A \text{ OR } B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$$

5. $P(A \text{ AND } B) = \text{-----}, P(A \text{ OR } B) = \text{-----}$

Show Answer

$$P(A \text{ AND } B) = \frac{3}{19}, P(A \text{ OR } B) = \frac{12}{19}$$

6. $A' = \text{-----}, P(A') = \text{-----}$

Show Answer

$$A' = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}; P(A' \text{ right}) = \frac{10}{19}$$

7. $P(A) + P(A') = \text{-----}$

Show Answer

$$P(A) + P(A') = 1 \left(\frac{9}{19} + \frac{10}{19} \right) = 1$$

8. $P(A|B) = \text{-----}, P(B|A) = \text{-----};$ Are the probabilities equal?

Show Answer

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{3}{6}, P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{3}{9}, \text{No}$$

Try It

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

1. $S =$ _____ Let
event
A = the sum is even and event B = the first number
is prime.
2. A = _____, B =

3. $P(A) =$ _____, $P(B) =$

4. A AND B = _____, A OR
B = _____
5. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$

6. $B' = \text{-----}, P(B') = \text{-----}$
7. $P(A) + P(A') = \text{-----}$
8. $P(A|B) = \text{-----}, P(B|A) = \text{-----};$ are the probabilities equal?

Show Answer

1. $S = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
2. $A = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$
 $B = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
3. $P(A) = \frac{1}{2}, P(B) = \frac{2}{3}$
4. $A \text{ AND } B = \{(2,2), (2,4), (3,1), (3,3)\}$
 $A \text{ OR } B = \{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
5. $P(A \text{ and } B) = \frac{1}{3}, \{P(A \text{ or } B) = \frac{5}{6}$
6. $B' = \{(1,1), (1,2), (1,3), (1,4)\}, P(B') = \frac{1}{3}$
7. $P(B) + P(B') = 1$
8. $P(A|B) = \frac{2}{3}, \text{No.}$

Example 2

A fair, six-sided die is rolled. Describe the sample space S , identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

1. Event T = the outcome is two.

Show Answer

$$T = \{2\}, P(T) = \frac{1}{6}$$

2. Event A = the outcome is an even number.

Show Answer

$$A = \{2, 4, 6\}, P(A) = \frac{1}{2}$$

3. Event B = the outcome is less than four.

Show Answer

$$B = \{1, 2, 3\}, P(B) = \frac{1}{2}$$

4. The complement of A .

Show Answer

$$A' = \{1, 3, 5\}, P(A') = \frac{1}{2}$$

5. A GIVEN B

Show Answer

$$(A | B) = \{2\}, P(A | B) = \frac{1}{3}$$

6. B GIVEN A

Show Answer

$$(B | A) = \{2\}, P(B|A) = \frac{1}{3}$$

7. A AND B

$$\text{Show Answer } (A \text{ and } B) = \{2\}, P(A \text{ and } B) = \frac{1}{6}$$

8. A OR B

$$\text{Show Answer } (A \text{ or } B) = \{1, 2, 3, 4, 6\}, P(A \text{ or } B) = \frac{5}{6}$$

9. A OR B'

$$\text{Show Answer } (A \text{ or } B') = \{2, 4, 5, 6\}, P(A \text{ or } B') = \frac{2}{3}$$

10. Event N = the outcome is a prime number.

Show Answer

$$N = \{2, 3, 5\}, P(N) = \frac{1}{2}$$

11. Event I = the outcome is seven.

Show Answer

A six-sided die does not have seven dots. $P(7) = 0$.

Example 3

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed	Total
Males	43	9	43 + 9 = 52
Females	44	4	44 + 4 = 48
Total	43 + 44 = 87	9 + 4 = 13	100

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

1. $P(M)$

Show Answer

$$P(M) = \frac{\text{amount of male}}{\text{total}} = \frac{52}{100} = 0.52$$

2. $P(F)$

Show Answer

$$P(F) = \frac{\text{amount of female}}{\text{total}} = \frac{48}{100} = 0.48$$

3. $P(R)$

Show Answer

$$P(R) = \frac{\text{amount of right-handed}}{\text{total}} = \frac{87}{100} = 0.87$$

4. $P(L)$

Show Answer

$$P(L) = \frac{\text{amount of left-handed}}{\text{total}} = \frac{13}{100} = 0.13$$

5. $P(M \text{ AND } R)$

Show Answer

$$P(M \text{ AND } R) = \frac{\text{amount of right-handed male}}{\text{total}} = \frac{43}{100} = 0.43$$

6. $P(F \text{ AND } L)$

Show Answer

$$P(F \text{ AND } L) = \frac{\text{amount of left-handed female}}{\text{total}} = \frac{4}{100} = 0.04$$

7. $P(M \text{ OR } F)$

Show Answer

$$P(M \text{ OR } F) = \frac{\text{amount of male} + \text{amount female}}{\text{total}} = \frac{52 + 48}{100} = 1$$

8. $P(M \text{ OR } R)$

Show Answer

$$P(M \text{ OR } R) = \frac{\text{amount of male} + \text{amount of right-handed}}{\text{total}} = \frac{43 + 9 + 44}{100} = 0.96$$

9. $P(F \text{ OR } L)$

Show Answer

$$P(F \text{ OR } L) = \frac{\text{amount of female} + \text{amount of left-handed}}{\text{total}} = \frac{9 + 4 + 44}{100} = 0.57$$

10. $P(M')$

Show Answer

$$P(M') = 1 - P(M) = 1 - 0.52 = 0.48$$

11. $P(R|M)$

Show Answer

$$P(R|M) = \frac{\text{number of right-handed male}}{\text{number of male}} = \frac{43}{52} =$$

0.8269 (rounded to four decimal places)

12. $P(F|L)$

Show Answer

$$P(F|L) = \frac{\text{number of left-handed female}}{\text{number of left-handed}} = \frac{4}{13}$$

0.3077 (rounded to four decimal places)

13. $P(L|F)$

Show Answer

$$P(L|F) = \frac{\text{number of left-handed female}}{\text{number of female}} = \frac{4}{48} =$$

0.0833

References

“Countries List by Continent.” Worldatlas, 2013. Available online at <http://www.worldatlas.com/cntycont.htm> (accessed May 2, 2013).

Concept Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq$

$P(A) \leq 1$

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

19. 3.2 Independent and Mutually Exclusive Events

Independent and mutually exclusive do not mean the same thing.

Independent Events

Two events are independent if the following are true:

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

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two rolls of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions.

If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without

replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether A and B are independent or dependent, **assume they are dependent until you can show otherwise.**

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit.

1. **Sampling with replacement:** Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.
2. **Sampling without replacement:** Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without

replacement, you cannot pick the same card twice.

Example 1

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

1. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?

Show Answer

Sampling with replacement

2. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Show Answer

No, we cannot tell if the sampling was with or without replacement.

Example 2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs.

1. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are QS, 1D, 1C, QD.
2. Suppose you pick four cards and put each card back before you pick the next card. Your cards are KH, 7D, 6D, KH.

Which of 1 or 2 did you sample with replacement and which did you sample without replacement?

Show Answer

1. Without replacement
2. With replacement

This video provides a brief lesson on finding the probability of independent events.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=38#oembed-1>

Try It

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs. Suppose that you sample four cards.

1. QS, 7D, 6D, KS
2. KH, 7D, 6D, KH
3. QS, 1D, 1C, QD

Which of the following outcomes are possible for sampling without replacement?

Show Answer

Without replacement:

1. QS, 7D, 6D, KS: Possible
2. KH, 7D, 6D, KH: Impossible
3. QS, 1D, 1C, QD: Possible

Which of the following outcomes are possible for sampling with replacement?

Show Answer

With replacement:

1. QS, 7D, 6D, KS: Possible
2. KH, 7D, 6D, KH: Possible

3. QS, 1D, 1C, QD: Possible

Mutually Exclusive Events

A and B are **mutually exclusive** events if they cannot occur at the same time. This means that A and B do not share any outcomes and $P(A \text{ AND } B) = 0$.

For example, suppose the sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

Let $A = \{1, 2, 3, 4, 5\}$, $B = \{4, 5, 6, 7, 8\}$, and $C = \{7, 9\}$.

1. $A \text{ AND } B = \{4, 5\}$. $P(A \text{ AND } B) = \frac{2}{10}$ and is not equal to zero. Therefore, A and B are not mutually exclusive.
2. A AND C do not have any numbers in common so $P(A \text{ AND } C) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, assume they are not until you can show otherwise. The following examples illustrate these definitions and terms.

Example 3

Flip two fair coins. (This is an experiment.)

The sample space is $\{HH, HT, TH, TT\}$ where T = tails and H = heads. The possible outcomes are HH, HT, TH , and TT . The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as $\{HH, HT, TH\}$. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A , so $B = A'$.
 $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for $A = P(\text{event } A) = \frac{3}{4}$.
- Let C = the event of getting all heads. $C = \{HH\}$. Are event C and event B mutually exclusive?

Show Answer

Since $B = \{TT\}$, $P(B \text{ AND } C) = 0$.

B and C are mutually exclusive. (B and C have no members in common because you cannot have all tails and all heads at the same time.)

- Let D = event of getting **more than one** tail. $D = \{TT\}$. Find $P(D)$.

Show Answer

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

- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) Find $P(E)$.

Show Answer

$$E = \{HT, HH\}.$$

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

- Find the probability of getting **at least one** (one or two) tail in two flips.

Show Answer

Let F = event of getting at least one tail in two flips.

$$F = \{HT, TH, TT\}.$$

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

Try It

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Show Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event A = Getting at least one black card = $\{BB, BR, RB\}$

$$P(A) = \frac{3}{4} = 0.75$$

Example 4

Flip two fair coins. Find the probabilities of the events.

1. Let F = the event of getting at most one tail (zero or one tail).
2. Let G = the event of getting two faces that are the same.
3. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
4. Are F and G mutually exclusive?
5. Let J = the event of getting all tails. Are J and H mutually exclusive?

Show Answer

The sample space is $\{HH, HT, TH, TT\}$ where T = tails and H = heads.

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(F) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(G) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(H) = \frac{2}{4}$
4. F and G share HH so $P(F \text{ AND } G)$ is not equal to zero (0). F and G are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (TT). H 's outcomes are HH and HT .
 J and H have nothing in common so $P(J \text{ AND } H) = 0$. J and H are mutually exclusive.

This video provides two more examples of finding the probability of events that are mutually exclusive.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=38#oembed-2>

Try It

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement).

1. If event F = the event of getting the white ball twice, find $P(F)$.

Show Answer

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

2. If event G = the event of getting two balls of different colors, find $P(G)$.

Show Answer

$$P(G) = \frac{1}{2}$$

3. If event H = the event of getting white on the first pick, find $P(G)$.

Show Answer

$$P(H) = \frac{1}{2}$$

4. Are F and G mutually exclusive?

Show Answer

Yes

5. Are G and H mutually exclusive?

Show Answer

No

Example 5

Roll one fair, six-sided die. The sample space is $\{1, 2, 3, 4, 5, 6\}$. Let event

A = a face is odd. Then $A = \{1, 3, 5\}$. Let event B = a face is even. Then $B = \{2, 4, 6\}$.

- Find the complement of A , A' .

Show Answer

The complement of A , A' , is B because A and B together make up the sample space.

$$P(A) + P(B) = P(A) + P(A') = 1.$$

$$\text{Also, } P(A) = \frac{3}{6}.$$

- Let event C = odd faces larger than two. Let event D = all even faces smaller than five.

Are C and D mutually exclusive events? Why?

Show Answer

$$C = \{3, 5\}, D = \{2, 4\}.$$

$P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time.

Therefore, C and D are mutually exclusive events.

- Let event E = all faces less than five.

Are C and E mutually exclusive events? Why?

Show Answer

$$C = \{3, 5\} \text{ and } E = \{1, 2, 3, 4\}.$$

$$C \text{ and } E = \{3\}$$

$$P(C \text{ AND } E) = \frac{1}{6}.$$

No, event C and event E are not mutually exclusive events.

- Find $P(C|A)$.

Show Answer

This is a conditional probability.

Recall that the event C is $\{3, 5\}$ and event A is $\{1, 3, 5\}$.

To find $P(C|A)$, find the probability of C using the sample space A .

You have reduced the sample space from the original sample space $\{1, 2, 3, 4, 5, 6\}$ to $\{1, 3, 5\}$.

$$\text{So, } P(C|A) = \frac{2}{3}.$$

Try It

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent?

Hint:

You must show ONE of the following:

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

Show Answer

$P(A|B) =$

$$\frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A)$$

The events are independent because $P(A|B) = P(A)$.

Example 6

Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?

Hint: If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
 - $P(H|G) = P(H)$
 - $P(G \text{ AND } H) = P(G)P(H)$
-

- Method 1:

Show that $P(G|H) = P(G)$.

$$P(G|H) = \frac{P(G \text{ and } H)}{P(H)} = \frac{0.3}{0.5} = 0.6$$

$$P(G) = 0.6$$

Since $P(G|H) = P(G)$, G and H are independent.

- Method 2:

Show $P(G \text{ AND } H) = P(G)P(H)$.

$$P(G \text{ and } H) = 0.3$$

$$P(G) * P(H) = 0.6 * 0.5 = 0.3$$

Since $P(G \text{ AND } H) = P(G)P(H)$, G and H are independent.

- Method 3:

Show that $P(H|G) = P(H)$.

$$P(H|G) = \frac{P(H \text{ and } G)}{P(G)} = \frac{0.3}{0.6} = 0.5$$

$$P(H) = 0.5$$

Since $P(H|G) = P(H)$, G and H are independent.

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math.

Try It

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Show Answer

Event G and $O = \{G1, G3\}$

$$P(G \text{ and } O) = \frac{2}{10} = 0.2$$

Example 7

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

1. Are C and D independent?

Show Answer

Yes, because $P(C|D) = P(C)$.

2. Are C and D mutually exclusive?

Show Answer

No, because $P(C \text{ AND } D)$ is not equal to zero.

3. What is $P(D|C)$?

Show Answer

$$P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{0.75} = 0.3$$

Try It

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

1. Find $P(B|D)$.
2. Find $P(D|B)$.
3. Are B and D independent?
4. Are B and D mutually exclusive?

Show Answer

1. $P(B|D) = 0.6667$
2. $P(D|B) = 0.5$
3. No
4. No

Example 8

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards

are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$. S has eight outcomes.

- Find $P(R)$.

Show Answer

$$P(R) = \frac{3}{8}.$$

- Find $P(R \text{ and } B)$.

Show Answer

$P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)

- Find $P(E)$.

Show Answer

$$P(E) = \frac{3}{8}. \text{ (There are three even-numbered cards, } R2, B2, \text{ and } B4.)$$

- Find $P(E|B)$.

Show Answer

$$P(E|B) = \frac{P(E \text{ and } B)}{P(B)} = \frac{2}{5}$$

(There are five blue cards: $B1, B2, B3, B4$, and $B5$.)

Out of the blue cards, there are two even cards; $B2$ and $B4$.)

- Find $P(B|E)$.

Show Answer

$$P(B|E) = \frac{P(E \text{ and } B)}{P(E)} = \frac{2}{3}.$$

There are three even-numbered cards: R2, B2, and B4.
Out of the even-numbered cards, two are blue; B2 and B4.

- Are events R and B mutually exclusive?

Show Answer

The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.

- Let G = card with a number greater than 3. Find $P(G)$.

Show Answer

$G = \{B4, B5\}$. $P(G) = \frac{2}{8}$, $P(G) = P(G|H)$, which means that G and H are independent.

Try It

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team, B be the event that a fan is wearing blue.

1. Are the events of rooting for the away team and wearing blue independent?
2. Are they mutually exclusive?

Show Answer

1. $P(B|A) = 0.67$, $P(B) = 0.25$
So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent).
2. They are also not mutually exclusive, because $P(B \text{ AND } A) = 0.20 \neq 0$.

Example 9

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$

- $P(L|F) = 0.75$

Solution 1:

Check if $P(F \text{ AND } L) = P(F) * P(L)$.

We are given that $P(F \text{ AND } L) = 0.45$, but $P(F) * P(L) = (0.60)(0.50) = 0.30$.

$P(F \text{ AND } L) \neq P(F)P(L)$, the events of being female and having long hair are not independent.

Solution 2

Check whether $P(L|F)$ equals $P(L)$.

We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$.

Since $P(L|F) \neq P(L)$, the events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Try It

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.55$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Show Answer

$$P(I \text{ AND } F) = 0,$$

$$P(I \text{ OR } F) = P(I) + P(F) - P(I \text{ AND } F) = 0.44 + 0.56 - 0 = 1$$

Example 10

1. Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are _____ outcomes.

Show Answer

H and T ; 2

2. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ____ outcomes.

Show Answer

1, 2, 3, 4, 5, 6 ; 6

3. Multiply the two numbers of outcomes. The answer is

_____.

Show Answer

$$2(6) = 12$$

4. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are H1 and T6.)

Show Answer

T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6

5. Event A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.

$A = \{ \text{_____} \}$. Find $P(A)$.

Show Answer

$$A = \{H2, H4, H6\}; P(A) = \frac{3}{12}$$

6. Event B = heads on the coin followed by a three on the die. $B = \{ \text{_____} \}$. Find $P(B)$.

Show Answer

$$B = \{H3\}; P(B) = \frac{1}{12}$$

7. Are A and B mutually exclusive?

(Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.)

Show Answer

Yes, because $P(A \text{ AND } B) = 0$

8. Are A and B independent?

(Hint: Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then A and B are independent. If not, then they are dependent).

Show Answer

$$P(A \text{ AND } B) = 0.$$

$$P(A) * P(B) = \frac{3}{12} * \frac{1}{12} = \frac{3}{144}.$$

$P(A \text{ AND } B) \neq P(A)P(B)$, so A and B are dependent.

Try It

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let T be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

1. Compute $P(T)$.
2. Compute $P(T|F)$.
3. Are T and F independent?.
4. Are F and S mutually exclusive?
5. Are F and S independent?

Show Answer

1. $P(T) = \frac{1}{4}$
2. $P(T|F) = \frac{1}{2}$
3. No
4. No
5. Yes

References

Lopez, Shane, Preeti Sidhu. "U.S. Teachers Love Their Lives, but Struggle in the Workplace." Gallup Wellbeing, 2013. <http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx> (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Concept Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of

being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

20. 3.3 Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then: $P(A \text{ AND } B) = P(B) * P(A|B)$.

This rule may also be written as

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

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are **independent**, then $P(A|B) = P(A)$.
Then $P(A \text{ AND } B) = P(A|B) * P(B)$ becomes $P(A \text{ AND } B) = P(A) * P(B)$.

The Addition Rule

If A and B are defined on a sample space, then: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$.

If A and B are **mutually exclusive**, then $P(A \text{ AND } B) = 0$. Then $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$ becomes $P(A \text{ OR } B) = P(A) + P(B)$.

Example 1

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska.

Klaus can only afford one vacation.

The probability that he chooses *New Zealand* is 0.6 and the probability that he chooses *Alaska* is 0.35.

Klaus can only afford to take one vacation.

1. What is the probability that he chooses either New Zealand or Alaska?
2. What is the probability that he does not choose to go anywhere on vacation?

Show Answer

1. Let A be New Zealand, B be Alaska.
 $P(\text{Klaus chooses New Zealand}) = P(A) = 0.6$
 $P(\text{Klaus chooses Alaska}) = P(B) = 0.35$
 $P(\text{Klaus chooses both New Zealand and Alaska}) = P(A \text{ and } B) = 0$
as he can only afford one vacation.
 $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ and } B) = 0.6 + 0.35 - 0 = 0.95.$
Therefore, the probability that he chooses either New Zealand or Alaska is 0.95.
 2. The probability that he does not choose to go anywhere on vacation
 $= 1 - P(A \text{ or } B)$
 $= 1 - 0.95$
 $= 0.05.$
-
-

Example 2

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game.

A = the event Carlos is successful on his first attempt.

B = the event Carlos is successful on his second attempt.

Carlos tends to shoot in streaks. The probability that he makes the second goal GIVEN that he made the first goal is 0.90.

$$P(A) = 0.65, P(B) = 0.65, P(B|A) = 0.90.$$

1. What is the probability that he makes both goals?

Show Answer

The problem is asking you to find $P(A \text{ AND } B) = P(B \text{ AND } A)$.

$$\text{Since } P(B|A) = 0.90, P(B \text{ AND } A) = P(B|A) * P(A) = (0.90)(0.65) = 0.585$$

Carlos makes the first and second goals with probability 0.585.

2. What is the probability that Carlos makes either the first goal or the second goal?

Show Answer

The problem is asking you to find $P(A \text{ OR } B)$.

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) = 0.65 + 0.65 - 0.585 = 0.715$$

Carlos makes either the first goal or the second goal with probability 0.715.

3. Are A and B independent?

Show Answer

$$P(B \text{ AND } A) = 0.585. P(B)P(A) = (0.65)(0.65) = 0.423$$

Since $P(B \text{ AND } A) \neq P(B)P(A)$, A and B are not independent.

4. Are A and B mutually exclusive?

Show Answer

$$P(A \text{ and } B) = 0.585.$$

To be mutually exclusive, $P(A \text{ AND } B)$ must equal zero.
Therefore, A and are not mutually exclusive events.

Watch this video for another example about first determining whether a series of events are mutually exclusive, then finding the probability of a specific outcome.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=39#oembed-1>

Try It

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. The probability that Helen makes the first shot is 0.75. The probability that Helen makes the second shot is 0.75. The probability that Helen makes the second free throw given that she made the first is 0.85.

What is the probability that Helen makes both free throws?

Show Answer

Let C be the event that Helen makes the first shot.
Let D be the event that Helen makes the second shot.

$$P(C) = 0.75, P(D) = 0.75, P(D|C) = 0.85$$

$$P(C \text{ AND } D) = P(D \text{ AND } C) = P(D|C)P(C) = (0.85)(0.75) = 0.6375$$

Helen makes the first and second free throws with probability 0.6375.

Example 3

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

1. What is the probability that the member is a novice swimmer?

Show Answer

$$\frac{28}{150}$$

2. What is the probability that the member practices four times a week?

Show Answer

$$\frac{80}{150}$$

3. What is the probability that the member is an advanced swimmer and practices four times a week?

Show Answer

$$\frac{40}{150}$$

4. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive?

Why or why not?

Show Answer

A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time, so $P(\text{advanced AND intermediate}) = 0$. These are mutually exclusive events.

5. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Show Answer

No, these are not independent events.

$P(\text{novice AND practices four times per week}) = 0.0667$

$P(\text{novice}) * P(\text{practices four times per week}) = 0.09960$.

$P(\text{novice AND practices four times per week}) \neq P(\text{novice}) * P(\text{practices four times per week})$

Try It

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Show Answer

The probability that a senior is taking a gap year =

$$\frac{200 - 140 - 40}{200} = \frac{20}{200} = 0.1$$

Example 4

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let M = math class, S = speech class, M|S = math given speech

1. What is the probability that Felicity enrolls in math and speech?
2. What is the probability that Felicity enrolls in math or speech classes?
3. Are M and S independent?
4. Are M and S mutually exclusive?

Solution

1. Show Answer

$$P(M \text{ AND } S) = P(M|S)P(S) = 0.25 * 0.65 = 0.1625.$$

2. Show Answer

$$P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S) = 0.2 + 0.65 - 0.1625 = 0.6875$$

3. Show Answer

$$P(M \text{ AND } S) = 0.1625$$

$$P(M) * P(S) = 0.2 * 0.65 = 0.13$$

Since $P(M \text{ AND } S) \neq P(M) * P(S)$, M and S are not independent.

4. Show Answer

Since $P(M \text{ AND } S) = 0.1625 \neq 0$, M and S are not mutually exclusive.

Try It

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

1. Find $P(B \text{ AND } D)$.

Show Answer

$$P(B \text{ AND } D) = P(D|B)P(B) = (0.5)(0.4) = 0.20.$$

2. Find $P(B \text{ OR } D)$.

Show Answer

$$P(B \text{ OR } D) = P(B) + P(D) - P(B \text{ AND } D) = 0.40 + 0.30 - 0.20 = 0.50$$

Example 5

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time.

Let B = woman develops breast cancer and let N = tests negative.
Suppose one woman is selected at random.

1. What is the probability that the woman develops breast cancer?

Show Answer

$$P(B) = 0.143$$

2. What is the probability that woman tests negative?

Show Answer

$$P(N) = 0.85$$

3. Given that the woman has breast cancer, what is the probability that she tests negative?

Show Answer

$$P(N|B) = 0.02$$

4. What is the probability that the woman has breast cancer AND tests negative?

Show Answer

$$P(B \text{ AND } N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$$

5. What is the probability that the woman has breast cancer or tests negative?

Show Answer

$$P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = 0.143 + 0.85 - 0.0029 = 0.9901$$

6. Are having breast cancer and testing negative independent events?

Show Answer

No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

7. Are having breast cancer and testing negative mutually exclusive?

Show Answer

No. $P(B \text{ AND } N) = 0.0029$. For B and N to be mutually exclusive, $P(B \text{ AND } N)$ must be zero.

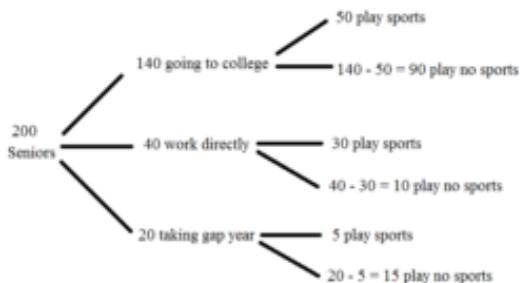
Try It

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports.

What is the probability that a senior is going to college and plays sports?

Show Answer

Let A = student is a senior going to college. Let B = student plays sports.



Show Answer

$$\begin{aligned}
 P(A \text{ and } B) &= \frac{\text{number of seniors going to college and play sports}}{\text{total amount of seniors}} \\
 &= \frac{50}{200} = \frac{1}{4}
 \end{aligned}$$

Example 6

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for

breast cancer is negative about 85% of the time.

Let B = woman develops breast cancer, let N = tests negative, and let P = tests positive.

Suppose one woman is selected at random.

1. Given that a woman develops breast cancer, what is the probability that she tests positive?

Show Answer

$$P(P|B) = 1 - P(N|B) = 1 - 0.02 = 0.98.$$

2. What is the probability that a woman develops breast cancer and tests positive?

Show Answer

$$P(B \text{ AND } P) = P(P|B)P(B) = 0.98 * 0.143 = 0.1401.$$

3. What is the probability that a woman does not develop breast cancer?

Show Answer

$$P(B') = 1 - P(B) = 1 - 0.143 = 0.857.$$

4. What is the probability that a woman tests positive for breast cancer?

Show Answer

$$P(P) = 1 - P(N) = 1 - 0.85 = 0.15.$$

Try It

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

1. Find $P(B')$.
2. Find $P(D \text{ AND } B)$.
3. Find $P(B|D)$.
4. Find $P(D \text{ AND } B')$.
5. Find $P(D|B')$.

Show Answer

1. $P(B') = 0.60$
2. $P(D \text{ AND } B) = P(D|B) * P(B) = 0.20$
3. $P(B|D) = \frac{P(B \text{ AND } D)}{P(D)} = \frac{0.20}{0.30} = 0.66$
4. $P(D \text{ AND } B') = P(D) - P(D \text{ AND } B) = 0.30 - 0.20 = 0.10$
5. $P(D|B') = P(D \text{ AND } B') * P(B') = (P(D) - P(D \text{ AND } B))(0.60) = (0.10)(0.60) = 0.06$

References

DiCamillo, Mark, Mervin Field. "The File Poll." Field Research Corporation. Available online at <http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf> (accessed May 2, 2013).

Rider, David, "Ford support plummeting, poll suggests," The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

"Mayor's Approval Down." News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/NewsArchives/NewsReleases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

"Roulette." Wikipedia. Available online at <http://en.wikipedia.org/wiki/Roulette> (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. "Language Use in the United States: 2007." United States Census Bureau. Available online at <http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf> (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at <http://www.ropercenter.uconn.edu/> (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2, 2013).

Concept Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \text{ AND } B) = P(A|B)P(B)$

The addition rule: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

2I. 3.4 Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

The following video shows an example of finding the probability of an event from a table.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=40#oembed-1>

Example 1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that $305 + 450 = 755$ and $70 + 685 = 755$.

Calculate the following probabilities using the table.

- Find $P(\text{Person is a car phone user})$.

Show Answer

$$\frac{\text{number of car phone users}}{\text{total number in study}} = \frac{305}{755}$$

- Find $P(\text{person had no violation in the last year})$.

Show Answer

$$\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$$

- Find $P(\text{Person had no violation in the last year AND was a car phone user})$.

Show Answer

$$\frac{280}{755}$$

- Find $P(\text{Person is a car phone user OR person had no violation in the last year})$.

Show Answer

$$\left(\frac{305}{755} + \frac{685}{755} \right) - \frac{280}{755} = \frac{710}{755}$$

- Find $P(\text{Person is a car phone user GIVEN person had a violation})$

in the last year).

Show Answer

$\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

6. Find $P(\text{Person had no violation last year GIVEN person was not a car phone user})$

Show Answer

$\frac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

This video shows an example of how to determine the probability of an AND event using a contingency table.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=40#oembed-2>

Try It

This table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is $P(\text{athlete stretches before exercising})$?

Show Answer

$$P(\text{athlete stretches before exercising}) = \frac{350}{800} =$$

0.4375

2. What is $P(\text{athlete stretches before exercising} | \text{no injury in the last year})$?

Show Answer

$$P(\text{athlete stretches before exercising} | \text{no injury in the last year}) = \frac{295}{514} = 0.5739$$

Example 2

This table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	---	45
Male	---	---	14	55
Total	---	41	---	---

1. Complete the table.

Show Answer

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

2. Are the events “being female” and “preferring the coastline” independent events?

Hint:

Let F = being female and let C = preferring the coastline.

Check if $P(F \text{ AND } C) = P(F) * P(C)$.

If $P(F \text{ AND } C) = P(F) * P(C)$, then F and C are independent.

If $P(F \text{ AND } C) \neq P(F) * P(C)$, then F and C are not independent.

Show Answer

$$P(F \text{ AND } C) = \frac{18}{100} = 0.18$$

$$P(F) * P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(F \text{ AND } C) \neq P(F) * P(C)$, so the events F and C are not independent.

3. Find the probability that a person is male given that the person prefers hiking near lakes and streams.
-

Hint:

Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
 2. Fill in the blanks and calculate the probability: $P(___|___) = \frac{___}{___}$.
 3. Is the sample space for this problem all 100 hikers? If not, what is it?
-

Show Answer

The word given tells you that this is a conditional.

$$P(M|L) = \frac{25}{41}$$

No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

4. Find the probability that a person is female or prefers hiking on mountain peaks.

Hint:

Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
 2. Find $P(P)$.
 3. Find $P(F \text{ AND } P)$.
 4. Find $P(F \text{ OR } P)$.
-

Show Answer

The probability that a person is female or prefers hiking on mountain peaks = $\frac{59}{100}$

- $P(F) = \frac{45}{100}$
 - $P(P) = \frac{25}{100}$
 - $P(F \text{ AND } P) = \frac{11}{100}$
 - $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$
-

Try It

This table shows a random sample of 200 cyclists and

the routes they prefer. Let M = males and H = hilly path.

	Gende r	Lake Path	Hilly Path	Wooded Path	Tot al
Femal e		45	38	27	110
Male		26	52	12	90
Total		71	90	39	200

1. Out of the males, what is the probability that the cyclist prefers a hilly path?

Show Answer

$$P(H|M) = \frac{52}{90} = 0.5778$$

2. Are the events “being male” and “preferring the hilly path” independent events?

Show Answer

For M and H to be independent, show $P(H|M) = P(H)$

$$P(H|M) = 0.5778, P(H) = \frac{90}{200} = 0.45$$

$P(H|M) \neq P(H)$, so M and H are not independent.

Example 3

Muddy Mouse lives in a cage with three doors.
If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$.
If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$.
The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$.
It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	----
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	----
Total	----	----	----	1

- The first entry $\frac{1}{15} = (\frac{1}{5})(\frac{1}{3})$ is P(Door One AND Caught)
- The entry $\frac{4}{15} = (\frac{4}{5})(\frac{1}{3})$ is P(Door One AND Not Caught)

Verify the remaining entries.

1. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry

is 1.
Show Answer

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{16}$	1

2. What is the probability that Alissa does not catch Muddy?
Show Answer

$$\frac{41}{60}$$

3. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?
Show Answer

$$\frac{9}{19}$$

Example 4

This table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants
2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

1. Find $P(2009 \text{ AND Robbery})$.

Show Answer

0.0294

2. Find $P(2010 \text{ AND Burglary})$.

Show Answer

0.1551

3. Find $P(2010 \text{ OR Burglary})$.

Show Answer

0.7165

4. Find $P(2011|Rape)$.

Show Answer

0.2365

5. Find $P(\text{Vehicle}|2008)$.

Show Answer

0.2575

This video gives an example of determining an “OR” probability given a table.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=40#oembed-3>

Try It

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

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column.

Show Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

2. Find the probability that a randomly chosen individual from this group is Tall.

Show Answer

$$P(\text{Tall}) = \frac{50}{205} = 0.244$$

3. Find the probability that a randomly chosen individual from this group is Obese and Tall.

Show Answer

$$P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$$

4. Find the probability that a randomly chosen individual from this group is Tall given that the individual is Obese.

Show Answer

$$P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$$

5. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.

Show Answer

$$P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$$

6. Find the probability a randomly chosen individual from this group is Tall and Underweight.

Show Answer

$$P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$$

7. Are the events Obese and Tall independent?

Show Answer

No. $P(\text{Tall}) \neq P(\text{Tall}|\text{Obese})$.

References

“Blood Types.” American Red Cross, 2013. Available online at <http://www.redcrossblood.org/learn-about-blood/blood-types> (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” *The New England Journal of Medicine*, 2013. Available online at <http://www.nejm.org/doi/full/10.1056/NEJMoa033250> (accessed May 2, 2013).

“Human Blood Types.” Unite Blood Services, 2011. Available online at <http://www.unitedbloodservices.org/learnMore.aspx> (accessed May 2, 2013).

Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at <http://www.disastercenter.com/crime/> (accessed May 2, 2013).

Concept Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

22. 3.5 Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

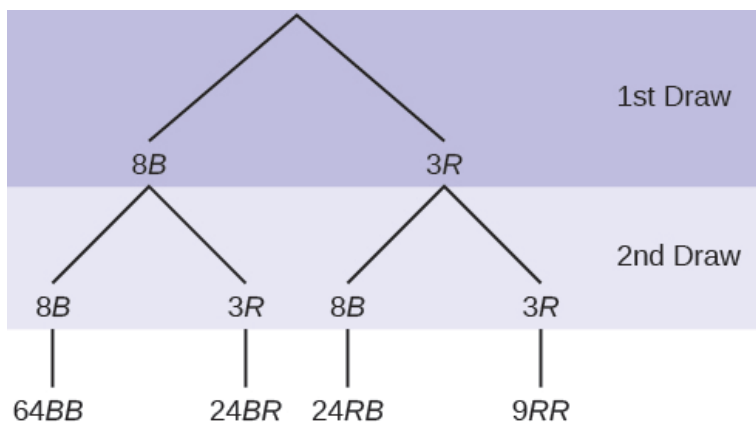
A tree diagram is a special type of graph used to determine the outcomes of an experiment. It consists of “branches” that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 1

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, **with replacement**. “With replacement” means that you put the first ball back in the urn before you select the second ball.

a. Use tree diagram to show all possible outcomes.

Show Answer



$$\text{Total} = 64 + 24 + 24 + 9 = 121$$

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct.

In fact, we can list each red ball as R1, R2, and R3 and each blue ball as B1, B2, B3, B4, B5, B6, B7, and B8.

Then the nine RR outcomes can be written as:

R1R1 R1R2 R1R3 R2R1 R2R2 R2R3 R3R1 R3R2 R3R3

The other outcomes are similar.

There are a total of 11 balls in the urn.

Draw two balls, one at a time, with replacement.

There are $11(11) = 121$ outcomes, the size of the sample space.

b. List the 24 BR outcomes.

Show Answer

B1R1, B1R2, B1R3, B2R1, B2R2, B2R3, B3R1, B3R2, B3R3, B4R1, B4R2, B4R3,
B5R1, B5R2, B5R3, B6R1, B6R2, B6R3, B7R1, B7R2, B7R3, B8R1, B8R2, B8R3

c. Using the tree diagram, calculate $P(RR)$.

Show Answer

$$P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$$

d. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

Show Answer

$$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$$

e. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

Show Answer

$$P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$$

f. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Show Answer

$$P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{24}{88} = \frac{3}{11}$$

g. Using the tree diagram, calculate $P(BB)$.

Show Answer

$$P(BB) = \frac{64}{121}$$

h. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.

Show Answer

$$P(B \text{ on 2nd} \mid R \text{ on 1st}) = \frac{8}{11}$$

There are 9 + 24 outcomes that have R on the first draw (9 RR and 24 RB).

The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B

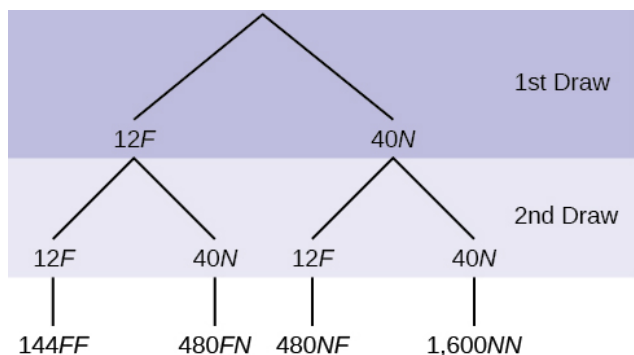
on the second draw.

$$\text{Therefore, } P(\text{B on the 2nd draw given R on the first draw}) = \frac{24}{33} = \frac{8}{11}$$

Try It

In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.

Show Answer



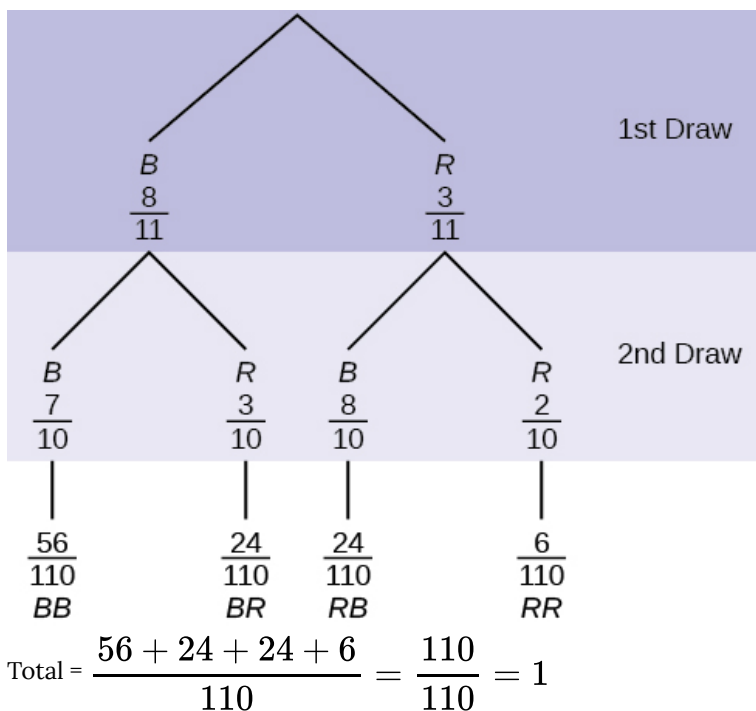
Total number of outcomes is $144 + 480 + 480 + 1600 = 2,704$.

$$P(\text{FF}) = \frac{144}{2704} = \frac{9}{169}$$

“Without replacement” means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches.

Example 2

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn.



Note: If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

a. $P(RR) = \text{-----}$

Show Answer

$$P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$$

b. Fill in the blanks:

$P(RB \text{ or } BR) = \text{-----}$

Show Answer

$$P(RB \text{ or } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$$

c. $P(\text{R on 2nd}|\text{B on 1st}) =$ _____

Show Answer

$$\frac{P(\text{R on 2nd and B on 1st})}{P(\text{B on 1st})} = \frac{\frac{24}{110}}{\frac{8}{11}} = \frac{3}{10}$$

d. Fill in the blanks.

$$P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = (---)(---) = \frac{24}{100}$$

Show Answer

$$P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$$

e. Find $P(\text{BB})$.

Show Answer

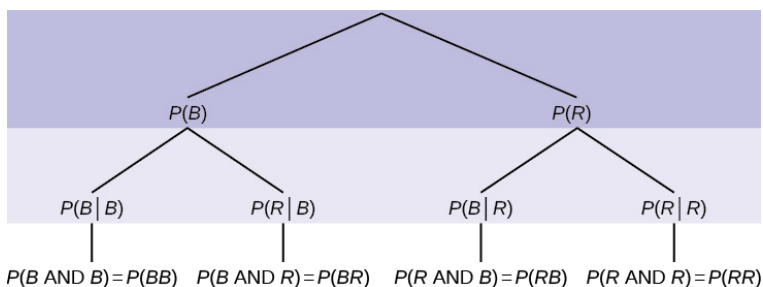
$$P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right) = \frac{56}{110}$$

f. Find $P(\text{B on 2nd}|\text{R on 1st})$.

Show Answer

$$\frac{P(\text{B on 2nd and R on 1st})}{P(\text{R on 1st})} = \frac{\frac{24}{110}}{\frac{3}{11}} = \frac{8}{10}$$

If we are using probabilities, we can label the tree in the following general way.



- $P(R|R)$ here means $P(R \text{ on 2nd} | R \text{ on 1st})$
- $P(B|R)$ here means $P(B \text{ on 2nd} | R \text{ on 1st})$
- $P(R|B)$ here means $P(R \text{ on 2nd} | B \text{ on 1st})$
- $P(B|B)$ here means $P(B \text{ on 2nd} | B \text{ on 1st})$

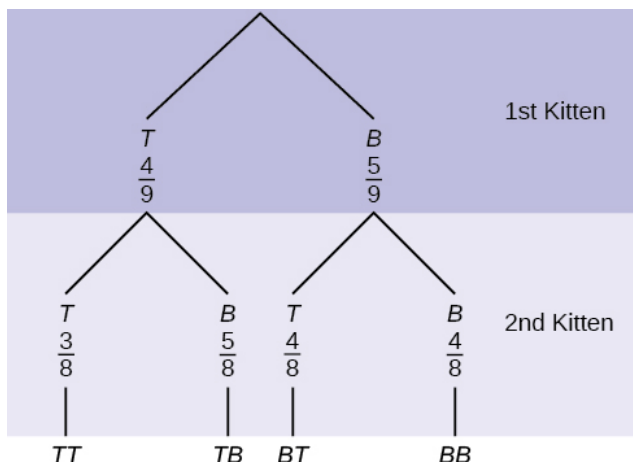


One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=41#oembed-1>

Try It

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



1. **What is the probability that both kittens are tabby?**

Show Answer

$P(\text{both kittens are tabby}) =$

$$\left(\frac{4}{9}\right)\left(\frac{3}{8}\right) = \frac{12}{72} = \frac{1}{6}$$

2. **What is the probability that one kitten of each coloring is selected?**

Show Answer

$P(\text{one kitten of each coloring is selected}) = P(TB)$

$+ P(BT) =$

$$\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right) = \frac{40}{72} = \frac{5}{9}$$

3. **What is the probability that a tabby is chosen as the second kitten when a black kitten was**

chosen as the first?

Show Answer

P(a tabby is chosen as the second kitten when a black kitten was chosen as the first) = $\frac{4}{8}$

4. **What is the probability of choosing two kittens of the same color?**

Show Answer

P(choosing two kittens of the same color) =

$P(TT) + P(BB) =$

$$\left(\frac{4}{9}\right)\left(\frac{3}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right) = \frac{32}{72} = \frac{4}{9}$$

Example 3

Suppose there are 4 red balls and 3 yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Show Answer

$$P(\text{one of each coloring is selected}) = P(RY) + P(YR) = \left(\frac{4}{7}\right)\left(\frac{3}{6}\right) + \left(\frac{3}{7}\right)\left(\frac{4}{6}\right) = \frac{24}{42} = \frac{4}{7}$$

Venn Diagram

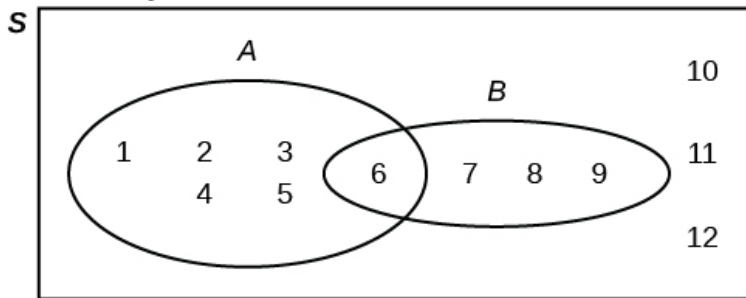
A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example 4

Suppose an experiment has the outcomes 1, 2, 3, ..., 12 where each outcome has an equal chance of occurring.

Let event $A = \{1, 2, 3, 4, 5, 6\}$ and event $B = \{6, 7, 8, 9\}$. Then $A \text{ AND } B = \{6\}$ and $A \text{ OR } B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

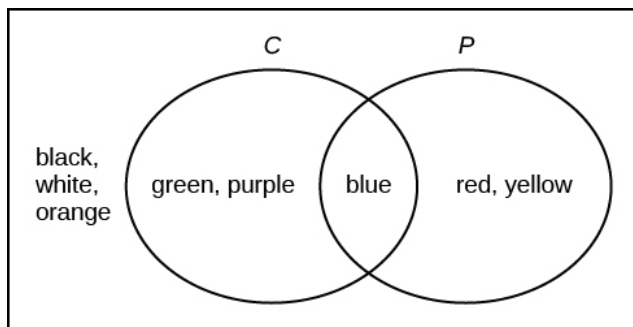
The Venn diagram is as follows:



Try It

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event $C = \{\text{green, blue, purple}\}$ and event $P = \{\text{red, yellow, blue}\}$. Then $C \text{ AND } P = \{\text{blue}\}$ and $C \text{ OR } P = \{\text{green, blue, purple, red, yellow}\}$. Draw a Venn diagram representing this situation.

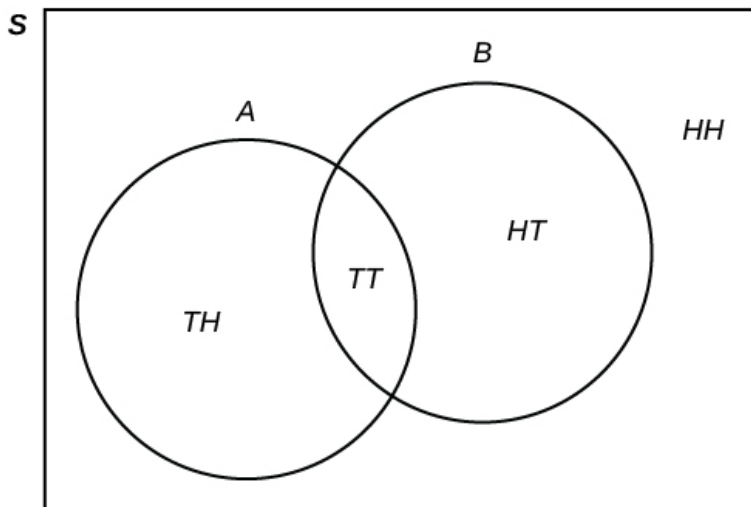
Show Answer



Example 5

Flip two fair coins. Let A = tails on the first coin. Let B = tails on the second coin. Then $A = \{TT, TH\}$ and $B = \{TT, HT\}$. Therefore, $A \text{ AND } B = \{TT\}$. $A \text{ OR } B = \{TH, TT, HT\}$.

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$. The outcome HH is in NEITHER A NOR B . The Venn diagram is as follows:

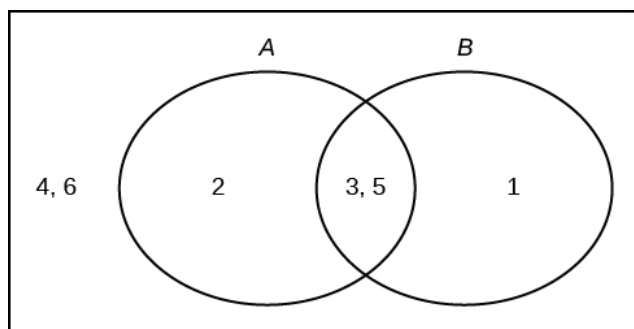


Try It

Roll a fair, six-sided die. Let A = a prime number of

dots is rolled. Let B = an odd number of dots is rolled. Then $A = \{2, 3, 5\}$ and $B = \{1, 3, 5\}$. Therefore, $A \text{ AND } B = \{3, 5\}$. $A \text{ OR } B = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

Show Answer



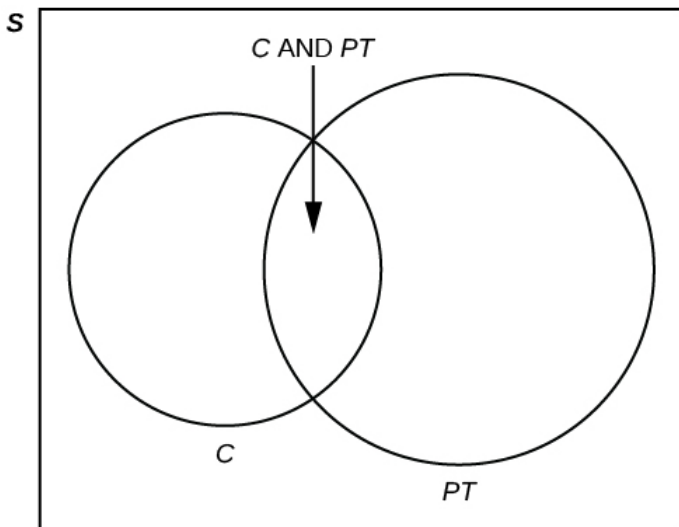
Example 6

40% of the students at a local college belong to a club and **50%** work part time. **5%** of the students work part time and belong to a club.

a. **Draw a Venn diagram showing the relationships.**

Show Answer

Let C = student belongs to a club and PT = student works part time.



b. If a student is selected at random, find the probability that the student belongs to a club.

Show Answer

$$P(\text{the student belongs to a club}) = 0.40$$

c. If a student is selected at random, find the probability that the student works part time.

Show Answer

$$P(\text{the student works part time}) = 0.50$$

d. If a student is selected at random, find the probability that the student belongs to a club AND works part time.

Show Answer

$$P(\text{the student belongs to a club AND works part time}) = 0.05$$

e. If a student is selected at random, find the probability that the student belongs to a club **given** that the student works part time.

Show Answer

$P(\text{the student belongs to a club} \mid \text{the student works part time}) =$

$$\frac{P(\text{the student belongs to a club AND the student works part time})}{P(\text{the student works part time})} = \frac{0.05}{0.50} = 0.1$$

f. If a student is selected at random, find the probability that the student belongs to a club **OR** works part time.

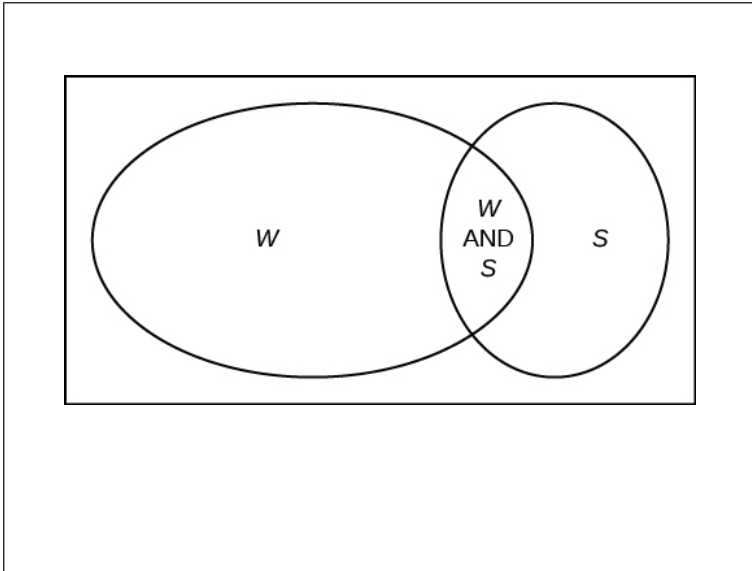
Show Answer

$P(\text{the student belongs to a club OR works part time}) = P(\text{the student belongs to a club}) + P(\text{the student works part time}) - P(\text{the student belongs to a club AND works part time}) = 0.40 + 0.50 - 0.05 = 0.85$

Try It

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let W = works a second job and S = spouse also works.

Show Answer



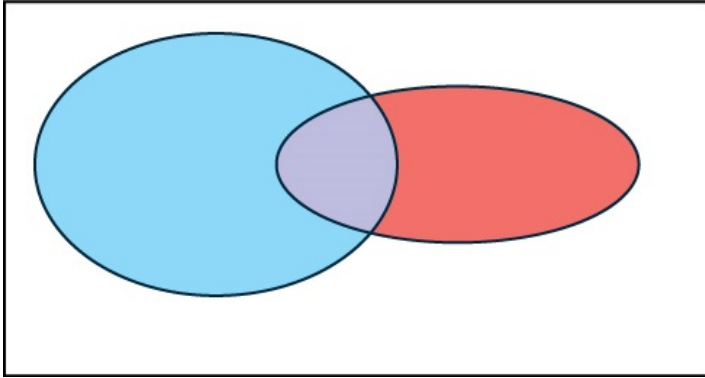
Example 7

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

1. Draw a Venn diagram representing the situation.

Show Answer

In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction.



2. Find the probability that the customer buys either a novel or anon-fiction book.

Show Answer

$$P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8.$$

3. In the Venn diagram, describe the overlapping area using a complete sentence.

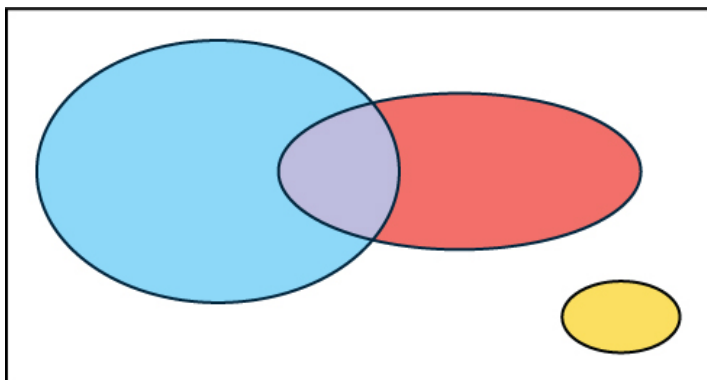
Show Answer

The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

4. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Show Answer

In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.



<https://youtu.be/MassxXy8iko>

Glossary

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

Solutions to Try These 1:

- a. B1R1 B1R2 B1R3 B2R1 B2R2 B2R3 B3R1 B3R2 B3R3 B4R1 B4R2 B4R3
B5R1 B5R2 B5R3 B6R1 B6R2 B6R3 B7R1B7R2 B7R3 B8R1 B8R2 B8R3
- b. $P(RR) = (311)(311) = 9121$
- c. $P(RB \text{ OR } BR) = (311)(811) + (811)(311) = 48121$

d. $P(\text{R on 1st draw AND B on 2nd draw}) = P(\text{RB}) = (311)(811) = 24121$

e. $P(\text{R on 2nd draw GIVEN B on 1st draw}) = P(\text{R on 2nd} | \text{B on 1st}) = 2488 = 311$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are $24 + 64 = 88$ possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $2488 = 311$.

f. $P(\text{BB}) = 64121$

g. $P(\text{B on 2nd draw} | \text{R on 1st draw}) = 811$

There are $9 + 24$ outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then $9 + 24 = 33$. 24 of the 33 outcomes have B on the second draw. The probability is then 2433 .

Solutions to Try These 2:

a. $P(\text{RR}) = (311)(210) = 6110$

b. $P(\text{RB OR BR}) = (311)(810) + (811)(310) = 48110$

c. $P(\text{R on 2nd} | \text{B on 1st}) = 310$

d. $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = (311)(810) = 24100$

e. $P(\text{BB}) = (811)(710)$

f. Using the tree diagram, $P(\text{B on 2nd} | \text{R on 1st}) = P(\text{R} | \text{B}) = 810$.

PART IV

DISCRETE RANDOM VARIABLES

23. Introduction: Discrete Random Variables



You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.
- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A random variable describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variable Notation

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable. If **X is a random variable, then X is written in words, and x is quantitative.**

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is TTT ; THH ; HTH ; HHT ; HTT ; THT ; TTH ; HHH . Then, $x = 0, 1, 2, 3$. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Glossary

Random Variable (RV)

a characteristic of interest in a population being studied;
common notation for variables are upper case Latin letters X , Y , Z ,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin

letters x , y , and z . For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,...

Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

24. 4.1 Probability Distribution Function (PDF) for a Discrete Random Variable

The idea of a random variable can be confusing. In this video we help you learn what a random variable is, and the difference between discrete and continuous random variables.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=44#oembed-1>

A discrete probability distribution function has two characteristics:

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

Example 1:

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$.

$P(x)$ = probability that X takes on a value x .

Probability distribution table for Example 1

x	$P(x)$
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

1. Each $P(x)$ is between zero and one, inclusive.

$$P(x = 0) = \frac{2}{50} > 0$$

$$P(x = 1) = \frac{11}{50} > 0$$

$$P(x = 2) = \frac{23}{50} > 0$$

$$P(x = 3) = \frac{9}{50} > 0$$

$$P(x = 4) = \frac{4}{50} > 0$$

$$P(x = 5) = \frac{1}{50} > 0$$

2. The sum of the probabilities is one, that is,

$$\begin{aligned} &P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5) \\ &= \frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} \\ &= 1 \end{aligned}$$

Try It

Suppose Nancy has classes **three days** a week. She attends classes three days a week **80%** of the time, **two days 15%** of the time, **one day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Let X = the number of days Nancy has classes in a week.

- a. X takes on what values?

Show Answer

0 day, 1 day, 2 days, 3 days

- b. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table). The table should have two columns labeled x and $P(x)$.

Show Answer

X	P(X)
0	1% = 0.01
1	4% = 0.04
2	15% = 0.15
3	80% = 0.80

c. What does the $P(x)$ column sum to?

Show Answer

$$P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) = 0.01 + 0.04 + 0.15 + 0.80 = 1.$$

Example 2

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Show Answer

X is the number of days Jeremiah attends basketball practice per week.

X takes on the values 0, 1, and 2.

Number of days Jeremiah attends basketball practice per week, X	P(X)
0	2% = 0.02
1	8% = 0.08
2	90% = 0.90

--	--

Concept Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Solutions to Try These:

- a. Let X = the number of days Nancy attends class per week.
- b. 0, 1, 2, and 3
- c.

x	$P(x)$
0	0.01
1	0.04
2	0.15
3	0.80

25. 4.2 Mean / Expected Value and Standard Deviation of Discrete Random Variable

The **expected value** is often referred to as the “**long-term**” **average or mean**. This means that over the long term of doing an experiment over and over, you would **expect** this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. Probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times.

In his experiment, Pearson illustrated the Law of Large Numbers.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (**the theoretical probability and the relative frequency get closer and closer together**). When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This “long-term average” is known as the **mean** or **expected value** of the experiment and is denoted by the Greek letter μ . In other words, after conducting many trials of an experiment, you would expect this average value.

$$\begin{aligned} \text{Expected Value, } \mu &= x_1 * P(x_1) + x_2 * P(x_2) + \\ & x_3 * P(x_3) + \dots = \text{sum of all } (x * P(x)) \\ \text{Variance, } \sigma^2 &= \text{sum of all } ((x - \mu)^2 * P(x)) \\ \text{Standard deviation, } \sigma &= \sqrt{\sigma^2} \end{aligned}$$

Example 1

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ , of the number of days per week the men's soccer team plays soccer.

Solution:

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. X takes on the values 0, 1, 2. Construct a PDF table adding a column $x \cdot P(x)$. In this column, you will multiply each x value by its probability.

Expected Value Table. This table is called an expected value table. The table helps you calculate the expected value or long-term average.

x	$P(x)$	$x \cdot P(x)$
0	0.2	$(0)(0.2) = 0$
1	0.5	$(1)(0.5) = 0.5$
2	0.3	$(2)(0.3) = 0.6$

Add the last column $x \cdot P(x)$ to find the long term average or expected value: $(0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1$.

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say $\mu = 1.1$.

Example 2

The probability that a newborn baby does not cry after midnight is

$$\frac{2}{50}.$$

The probability that a newborn baby cries once after midnight is

$$\frac{11}{50}.$$

The probability that a newborn baby cries twice after midnight is

$$\frac{23}{50}.$$

The probability that a newborn baby cries thrice after midnight is

$$\frac{9}{50}.$$

The probability that a newborn baby cries for 4 times after midnight

is $\frac{4}{50}.$

The probability that a newborn baby cries for 5 times after midnight

is $\frac{1}{50}.$

1. Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight.

(The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight.)

2. Calculate the standard deviation of the variable as well.

Solution:

1. You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

x (Number of times a newborn baby crying after midnight)	P(x)	x · P(x)	(x – μ)² · P(x)
0	$P(x = 0) = \frac{2}{50}$	$(0)(\frac{2}{50}) = \frac{0}{50}$	$(0 - 2.1)^2 \cdot 0.04 =$
1	$P(x = 1) = \frac{11}{50}$	$(1)(\frac{11}{50}) = \frac{11}{50}$	$(1 - 2.1)^2 \cdot 0.22 =$
2	$P(x = 2) = \frac{23}{50}$	$(2)(\frac{23}{50}) = \frac{46}{50}$	$(2 - 2.1)^2 \cdot 0.46 =$
3	$P(x = 3) = \frac{9}{50}$	$(3)(\frac{9}{50}) = \frac{27}{50}$	$(3 - 2.1)^2 \cdot 0.18 =$
4	$P(x = 4) = \frac{4}{50}$	$(4)(\frac{4}{50}) = \frac{16}{50}$	$(4 - 2.1)^2 \cdot 0.08 =$
5	$P(x = 5) = \frac{1}{50}$	$(5)(\frac{1}{50}) = \frac{5}{50}$	$(5 - 2.1)^2 \cdot 0.02 =$
		sum of all (x · P(x)) = 2.1	sum of all ((x – μ) ² · P(x)) =

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

$$\mu = \text{Expected Value} = \frac{0}{50} + \frac{11}{50} + \frac{46}{50} + \frac{27}{50} + \frac{16}{50} + \frac{5}{50} = \frac{105}{50} = 2.1$$

2. Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x , multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$).

Add the values in the fourth column of the table:

$$\text{variance of } X = 0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05$$

The standard deviation of X is the square root of this sum:

$$\sigma = \sqrt{1.05} \simeq 1.0247$$

Try It

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

[practice-area rows="1"][/practice-area]

Click here to show answer:

x	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

$$\left(0 \right) \frac{4}{50} + (1) \frac{4}{50} + (2) \frac{16}{50} + (3) \frac{14}{50} + (4) \frac{6}{50} + (5) \frac{2}{50} = 0 + \frac{8}{50} + \frac{32}{50} + \frac{42}{50} + \frac{24}{50} + \frac{10}{50} = \frac{116}{50} = 2.32$$

The expected value is 2.32.

Example 3

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

Solution:

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

If you purchase a ticket, you will either make a profit of \$100,000 or lose \$2.

Since you are interested in your profit, the values of x are 100,000 dollars and -2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there

are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)(10^{-5}) = 0.00001$$

Therefore, the probability of winning is 0.00001 and the probability of losing is $1 - 0.00001 = 0.99999$

The expected value table is as follows:

	x	$P(x)$	$x \cdot P(x)$
Loss	-2	0.99999	$(-2)(0.99999) = -1.99998$
Profit	100,000	0.00001	$(100000)(0.00001) = 1$

Add the value of $x \cdot P(x)$. Expected value, $\mu = -1.99998 + 1 = -0.99998 \simeq 1$

Since -0.99998 is closed to -1 , you would, on average, expect to lose approximately \$1 for each game you play.

However, each time you play, you either lose \$2 or profit \$100,000.

The \$1 is the average or expected LOSS per game after playing this game over and over.

Try It

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay \$1 to play. If you guess the right suit every time, you get your money back and \$256.

What is your expected profit of playing the game over the long term?

[Click here to show answer:](#)

Let X = the amount of money you profit. The x -values are $-\$1$ and $\$256$.

The probability of guessing the right suit each time is

$$\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{256} = 0.0039$$

The probability of losing is

$$1 - \frac{1}{256} = \frac{255}{256} = 0.9961$$

Expected profit of playing the game over the long term = $(0.0039)256 + (0.9961)(-1) = 0.9984 + (-0.9961) = 0.0023$ or 0.23 cents.

Example 4

Suppose you play a game with a biased coin. You play each game by tossing the coin once.

$$P(\text{heads}) = \frac{2}{3}$$

If you toss a head, you pay \$6. If you toss a tail, you win \$10.

1. Define a random variable X.
2. Complete the following expected value table.

<u>x</u>		<u>-----</u>	<u>-----</u>
WIN	10	$\frac{1}{3}$	<u>-----</u>
LOSE	<u>-----</u>	<u>-----</u>	$\frac{-12}{3}$

3. What is the expected value, μ ? Do you come out ahead?

Show Answer

1. X = amount of profit

	<u>x</u>	<u>P(x)</u>	<u>x · P(x)</u>
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	-6	$\frac{2}{3}$	$\frac{-12}{3}$

3. The expected value μ = sum of all $(x \cdot P(x)) = \frac{10}{3} + \frac{-12}{3} = \frac{-2}{3}$.

You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Example 5

Suppose you play a game with a spinner. You play each game by

spinning the spinner once.

$P(\text{red}) = \frac{2}{5}.$

If you land on red, you pay \$10. If you land on blue, you don't pay or win anything. If you land on green, you win \$10.

Complete the following expected value table.

Let X be your profit.

	x	P(x)	x · P(x)
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Show Answer

	x	P(x)	x · P(x)
Red	-10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$
Green	10	$\frac{1}{5}$	$\frac{10}{5}$

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution,

find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men’s soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x - \mu)^2 P(x)$ and take the square root.

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
0	0.2	$(0)(0.2) = 0$	$(0 - 1.1)^2 \cdot (0.2) = 0.242$
1	0.5	$(1)(0.5) = 0.5$	$(1 - 1.1)^2 \cdot (0.5) = 0.005$
2	0.3	$(2)(0.3) = 0.6$	$(2 - 1.1)^2 \cdot (0.3) = 0.243$

Variance, $\sigma^2 = \text{sum of all } ((x - \mu)^2 \cdot P(x)) = 0.242 + 0.005 + 0.243 = 0.490$.

The standard deviation, $\sigma = \sqrt{0.49} = 0.7$

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce rounding error. For some probability distributions, there are short-cut formulas for calculating μ and σ .

Example 6

Toss a fair, six-sided die twice.

Let X = the number of faces that show an even number.

Construct a table and calculate the mean μ and standard deviation σ of X .

[Click here to show answer:](#)

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(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)

Use the sample space to complete the following table:
Calculating μ and σ .

x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(2 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Add the values in the third column to find the Expected Value:
 $\mu = \frac{36}{36} = 1.$

Add the values in the fourth column and take the square root of the sum:

Standard deviation = $\sigma = \sqrt{\frac{18}{36}} \simeq 0.7071.$

Example 7

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20.

Let X = the amount of profit from a bet.

$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 21.42\%$

$P(\text{loss}) = P(\text{one moderate earthquake will not occur}) = 100\% - 21.42\%$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers.

What is the standard deviation of X ?

(Hint: Construct a table to help you answer these questions.)

Solution:

	x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 \cdot P(x)$
win	50	0.2142	10.71	$(50 - (-5.006))^2 (0.2142) = 648.0964$
loss	-20	0.7858	-15.716	$(-20 - (-5.006))^2 \cdot (0.7858) = 176.6636$

Mean = Expected Value = $10.71 + (-15.716) = -5.006$.

If you make this bet many times under the same conditions, your long term outcome will be an **average loss of \$5.01 per bet**.

Standard Deviation = $\sqrt{648.0964 + 176.6636} \simeq 28.7186$

Try It

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. As in Example 6, you bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win \$100. If you lose the bet, you pay \$10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X .

[Click here to show answer:](#)

	x	$P(x)$	$x \cdot P(x)$	$(x - \mu)^2 P(x)$
win	100	0.0108	1.08	$[100 - (-8.812)]^2 \cdot 0.0108 = 127.8726$
loss	-10	0.9892	-9.892	$[-10 - (-8.812)]^2 \cdot 0.9892 = 1.3961$

$$\text{Mean} = \text{Expected Value} = \mu = 1.08 + (-9.892) = -8.812$$

If you make this bet many times under the same conditions, your long term outcome will be an average loss of \$8.81 per bet.

$$\text{Standard Deviation} = \sqrt{127.8726 + 1.3961} \approx 11.3696$$

Some of the more common discrete probability functions are [binomial](#), [geometric](#), hypergeometric, and [Poisson](#). Most elementary courses do not cover these distributions. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

References

Class Catalogue at the Florida State University. Available online at <https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy> (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

Concept Review

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

Mean or Expected Value:

$$\mu = \sum_{x \in X} xP(x)$$

Standard Deviation:

$$\sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$$



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=45#oembed-1>

26. 4.3 Binomial Distribution

There are three characteristics of a binomial experiment. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials. There are only two possible outcomes, called “success” and “failure,” for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p , of a success and probability, q , of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials.

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

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example

At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A “success” could be defined as an individual who withdrew. The random variable X = the number of students who withdraw from the randomly selected elementary physics class.

Try It

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a “success” be in this case?

Show Solution

A success would be a school that offers fruit in their lunch every day.

Example

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times.

Show Solution

Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = 0.45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$.

Try It

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Show Solution

$$P(x = 12)$$

Example

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, $p = 0.5$ and $q = 0.5$. The number of trials is $n = 15$. State the probability question mathematically.

Show Solution

$$P(x > 10)$$

Try It

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Show Solution

$$P(x > 3)$$

Example

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

- This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.
- If we are interested in the number of students who do their homework on time, then how do we define X ?
- What values does x take on?
- What is a “failure,” in words?
- If $p + q = 1$, then what is q ?
- The words “at least” translate as what kind of inequality for the probability question $P(x \geq 40)$.

Show Solution

- failure
- X = the number of statistics students who do their homework on time
- 0, 1, 2, ..., 50
- Failure is defined as a student who does not complete his or her homework on time. The probability of a success is $p = 0.70$. The number of trials is $n = 50$.

- e. $q = 0.30$
- f. greater than or equal to (\geq) The probability question is $P(x \geq 40)$.

Try It

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Show Solution

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

Notation for the Binomial: $B = \text{Binomial}$ Probability Distribution Function

$$X \sim B(n, p)$$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p ; n = number of trials, p = probability of a success on each trial.

Example

It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$,
 $p = 0.41$, and $q = 1 - 0.41 = 0.59$.
 $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$.
(calculator or computer)

Show Solution

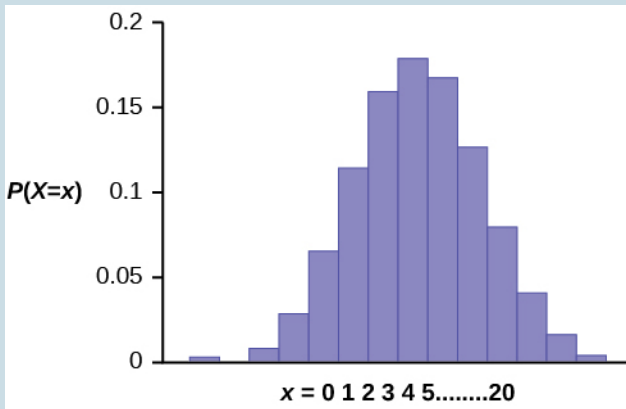
- Go into 2nd DISTR. The syntax for the instructions are as follows:
- To calculate (x = value): **binompdf(n, p, number)** if “number” is left out, the result is the binomial probability table.
- To calculate $P(x \leq \text{value})$: **binomcdf(n, p, number)** if “number” is left out, the result is the cumulative binomial probability table.
- For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,0.41,12). The result is $P(x \leq 12) = 0.9738$.

Note

If you want to find $P(x = 12)$, use the pdf (binompdf).
If you want to find $P(x > 12)$, use
 $1 - \text{binomcdf}(20, 0.41, 12)$.

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:



The y-axis contains the probability of x , where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(0.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is \sqrt{npq} .

$$\sigma = \sqrt{(20)(0.41)(0.59)} = 2.20$$

Try It

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Show Solution

$$P(x \leq 14) = 0.9695$$

Example

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

- What values does x take on?

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

Show Solution

- $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
- $X \sim B(100, \frac{8}{560})$
 - $P(x = 2) = \text{binompdf}(100, \frac{8}{560}, 2) = 0.2466$
 - $P(x \leq 6) = \text{binomcdf}(100, \frac{8}{560}, 6) = 0.9994$
 - $P(x > 3) = 1 - P(x \leq 3) = 1 - \text{binomcdf}(100, \frac{8}{560}, 3) = 1 - 0.9443 = 0.0557$
- Mean $= np = (100)(\frac{8}{560}) = \frac{800}{560} \approx 1.4286$
 - Standard Deviation $= \sqrt{npq} = \sqrt{(100)(\frac{8}{560})(\frac{552}{560})} \approx 1.1867$

Try It

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

- a. What is the probability distribution for X ?
- b. Use your calculator to find the following probabilities:
 - a. the probability that 25 adults in the sample prefer saving over spending
 - b. the probability that at most 20 adults prefer saving
 - c. the probability that more than 30 adults prefer saving
- c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X .

Show Solution

- a. $X \sim B(50, 0.6)$
- b. Using the TI-83, 83+, 84 calculator with instructions as provided earlier:
 - a. $P(x = 25) = \text{binompdf}(50, 0.6, 25) = 0.0405$
 - b. $P(x \leq 20) = \text{binomcdf}(50, 0.6, 20) = 0.0034$
 - c. $P(x > 30) = 1 - \text{binomcdf}(50, 0.6, 30) = 1 - 0.5535 = 0.4465$
- c.
 - a. **Mean** $= np = 50(0.6) = 30$
 - b. **Standard Deviation**

$$= \sqrt{npq} = \sqrt{50(0.6)(0.4)} \approx 3.4641$$

Example

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

- What is the probability distribution for X ?
- Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- Use your calculator to find the probability that at most eight people develop pancreatic cancer.
- Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Show Solution

- $X \sim B(200, 0.0128)$
- Mean = $np = 200(0.0128) = 2.56$
 - Standard Deviation = $\sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$
- Using the TI-83, 83+, 84 calculator:
 $P(x \leq 8) = \text{binomcdf}(200, 0.0128, 8) = 0.9988$
- $P(x = 5) = \text{binompdf}(200, 0.0128, 5) = 0.0707$

 $P(x = 6) = \text{binompdf}(200, 0.0128, 6) = 0.0298$

So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Try It

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

- What is the probability distribution for X ?
- Using the formulas, calculate the (i) mean and (ii) standard deviation of X .
- Use your calculator to find the probability that DeAndre scored with 60 of these shots.
- Find the probability that DeAndre scored with more than 50 of these shots.

Show Solution

- $X \sim B(80, 0.613)$
- Mean = $np = 80(0.613) = 49.04$
 - Standard Deviation =
$$\sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$$
- Using the TI-83, 83+, 84 calculator:

$$P(x = 60) = \text{binompdf}(80, 0.613, 60) = 0.0036$$

d. $P(x > 50) = 1 - P(x \leq 50) = 1 - \text{binomcdf}(80, 0.613, 50) = 1 - 0.6282 = 0.3718$

Example

The following example illustrates a problem that is not binomial. It violates the condition of independence. ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students?

Show Solution

The names of all committee members are put into a box, and two names are drawn without replacement. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Try It

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Show Solution

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

Concept Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n .
2. There are only two possible outcomes, called “success” and, “failure” for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be

calculated using the formula $\mu = np$, and the standard deviation is given by the formula

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

Formula Review

$X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p .

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x = 0, 1, 2, 3, \dots, n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$$p + q = 1$$

$$q = 1 - p$$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=46#oembed-1>

Show References

“Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

“Distance Education.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

“NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. “Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income,” GALLUP® Economy, 2013. Available online at <http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx> (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics> (accessed May 15, 2013).

27. Geometric Distribution

There are three main characteristics of a geometric experiment.

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bullseye until you hit the bullseye. The first time you hit the bullseye is a “success” so you stop throwing the dart. It might take six tries until you hit the bullseye. You can think of the trials as failure, failure, failure, failure, failure, success, STOP.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p , of a success and the probability, q , of a failure is the same for each trial. $p + q = 1$ and $q = 1 - p$. For example, the probability of rolling a three when you throw one fair die is $\frac{1}{6}$, the probability of a failure.

The probability of getting a three on the fifth roll is X = the number of independent trials until the first success.

Example

You play a game of chance that you can either win or lose (there are no other possibilities) **until** you lose. Your probability of losing is $p = 0.57$. What is the probability that it takes five games until you lose?

Show Solution

Let X = the number of games you play until you lose (includes the losing game). Then X takes on the values 1, 2, 3, ... (could go on indefinitely). The probability question is $P(x = 5)$.

Try It

You throw darts at a board until you hit the center area. Your probability of hitting the center area is $p = 0.17$. You want to find the probability that it takes eight throws until you hit the center. What values does X take on?

Show Solution

1, 2, 3, 4, ... n . It can go on indefinitely.

Example

A safety engineer feels that 35% of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) **until** she finds one that shows an accident caused by failure of employees to follow instructions. On average, how many reports would the safety engineer **expect** to look at until she finds a report showing an accident caused by employee

failure to follow instructions? What is the probability that the safety engineer will have to examine at least three reports until she finds a report showing an accident caused by employee failure to follow instructions?

Show Solution

Let X = the number of accidents the safety engineer must examine **until** she finds a report showing an accident caused by employee failure to follow instructions. X takes on the values 1, 2, 3, The first question asks you to find the **expected value** or the mean. The second question asks you to find $P(x \geq 3)$. (“At least” translates to a “greater than or equal to” symbol).

Try It

An instructor feels that 15% of students get below a C on their final exam. She decides to look at final exams (selected randomly and replaced in the pile after reading) until she finds one that shows a grade below a C. We want to know the probability that the instructor will have to examine at least ten exams until she finds one with a grade below a C. What is the probability question stated mathematically?

Show Solution

$$P(x \geq 10)$$

Example

Suppose that you are looking for a student at your college who lives within five miles of you. You know that 55% of the 25,000 students do live within five miles of you. You randomly contact students from the college until one says he or she lives within five miles of you. What is the probability that you need to contact four people?

This is a geometric problem because you may have a number of failures before you have the one success you desire. Also, the probability of a success stays the same each time you ask a student if he or she lives within five miles of you. There is no definite number of trials (number of times you ask a student).

- Let X = the number of _____ you must ask _____ one says yes.
- What values does X take on?
- What are p and q ?
- The probability question is $P(\text{_____})$.

Show Solution

- Let X = the number of **students** you must ask **until** one says yes.
- 1, 2, 3, ..., (total number of students)
- $p = 0.55$; $q = 0.45$
- $P(x = 4)$

Try It

You need to find a store that carries a special printer ink. You know that of the stores that carry printer ink, 10% of them carry the special ink. You randomly call each store until one has the ink you need. What are p and q ?

Show Solution

$$p = 0.1$$

$$q = 0.9$$

Notation for the Geometric: $G =$ Geometric Probability Distribution Function

$$X \sim G(p)$$

Read this as “ X is a random variable with a **geometric distribution**.” The parameter is p ; $p =$ the probability of a success for each trial.

Example

Assume that the probability of a defective computer component is 0.02. Components are randomly selected. Find the probability that the first defect is caused by the

seventh component tested. How many components do you expect to test until one is found to be defective?

Let X = the number of computer components tested until the first defect is found.

X takes on the values 1, 2, 3, ... where
 $p = 0.02$.

$$X \sim G(0.02)$$

Find $P(x = 7)$. $P(x = 7) = 0.0177$.

To find the probability that $x = 7$,

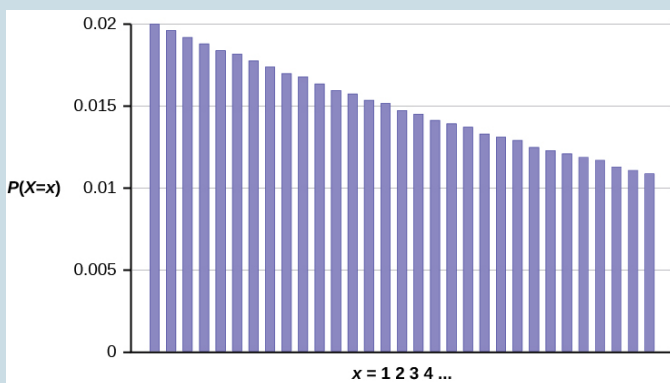
- Enter 2nd, DISTR
- Scroll down and select geometpdf(
- Press ENTER
- Enter 0.02, 7); press ENTER to see the result:

$$P(x = 7) = 0.0177$$

To find the probability that $x \leq 7$, follow the same instructions EXCEPT select E:geometcdf(as the distribution function.

The probability that the seventh component is the first defect is 0.0177.

The graph of $X \sim G(0.02)$ is:



The y-axis contains the probability of x , where X = the number of computer components tested.

The number of components that you would expect to test until you find the first defective one is the mean, $\mu = 50$.

The formula for the mean is $\mu = \frac{1}{p} = \frac{1}{0.02} = 50$

The formula for the variance is

$$\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right) = 2,450$$

The standard deviation is

$$\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{0.02}\right)\left(\frac{1}{0.02} - 1\right)} = 49.5$$

Try It

The probability of a defective steel rod is 0.01. Steel rods

are selected at random. Find the probability that the first defect occurs on the ninth steel rod. Use the TI-83+ or TI-84 calculator to find the answer.

Show Solution

$$P(x = 9) = 0.0092$$

Example

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Let X = the number of people you ask until one says he or she has pancreatic cancer. Then X is a discrete random variable with a geometric distribution:

$$X \sim G\left(\frac{1}{78}\right) \quad \text{or} \quad X \sim G(0.0128)$$

- What is the probability of that you ask ten people before one says he or she has pancreatic cancer?
- What is the probability that you must ask 20 people?
- Find the (i) mean and (ii) standard deviation of X .

Show Solution

- $P(x = 10) = \text{geompdf}(0.0128, 10) = 0.0114$
- $P(x = 20) = \text{geompdf}(0.0128, 20) = 0.01$
-

$$\text{i. Mean} = \mu = \frac{1}{p} = \frac{1}{0.0128} = 78$$

$$\text{ii. Standard Deviation} = \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.0128}{0.0128^2}} \approx 77.6234$$

Try It

The literacy rate for a nation measures the proportion of people age 15 and over who can read and write. The literacy rate for women in Afghanistan is 12%. Let X = the number of Afghani women you ask until one says that she is literate.

- What is the probability distribution of X ?
- What is the probability that you ask five women before one says she is literate?
- What is the probability that you must ask ten women?
- Find the (i) mean and (ii) standard deviation of X .

Show Solution

- $X \sim G(0.12)$
- $P(x = 5) = \text{geompdf}(0.12, 5) = 0.0720$
- $P(x = 10) = \text{geompdf}(0.12, 10) = 0.0380$
-

$$\text{i. Mean} = \mu = \frac{1}{p} = \frac{1}{0.12} \approx 3333$$

$$\text{ii. Standard Deviation} = \sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.12}{0.12^2}} \approx 7.8174$$

Concept Review

There are three characteristics of a geometric experiment:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p , of a success and the probability, q , of a failure are the same for each trial.

In a geometric experiment, define the discrete random variable X as the number of independent trials until the first success. We say that X has a geometric distribution and write $X \sim G(p)$ where p is the probability of success in a single trial.

The mean of the geometric distribution $X \sim G(p)$ is

$$\mu = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1}{p} \left(\frac{1}{p} - 1 \right)}.$$

Formula Review

$X \sim G(p)$ means that the discrete random variable X has a geometric probability distribution with probability of success in a single trial p .

X = the number of independent trials until the first success

X takes on the values $x=1, 2, 3, \dots$

p = the probability of a success for any trial

q = the probability of a failure for any trial $p + q = 1$

$q = 1-p$

The mean is $\mu = \frac{1}{p}$

The standard deviation is $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1}{p}(\frac{1}{p} - 1)}$

Show References

“Millennials: A Portrait of Generation Next,” PewResearchCenter. Available online at <http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf> (accessed May 15, 2013).

“Millennials: Confident. Connected. Open to Change.” Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at <http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-to-change/> (accessed May 15, 2013).

“Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran.

The American Freshman: National Norms Fall 2011. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan,” The European Union and ICON-Institute. Available online at http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“UNICEF reports on Female Literacy Centers in Afghanistan

established to teach women and girls basic reading [sic] and writing skills,” UNICEF Television. Video available online at <http://www.unicefusa.org/assets/video/afghan-female-literacy-centers.html> (accessed May 15, 2013).

28. Poisson Distribution

There are two main characteristics of a Poisson experiment.

1. The Poisson probability distribution gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial if the probability of success is “small” (such as 0.01) and the number of trials is “large” (such as 1,000). You will verify the relationship in the homework exercises. n is the number of trials, and p is the probability of a “success.”

The random variable X = the number of occurrences in the interval of interest.

Example 1:

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in five minutes is three?

Solution:

Let X = the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, **then the average number of loaves put on the shelf in five minutes is** $\left(\frac{5}{30}\right)(12) = 2$ loaves of bread.

The probability question asks you to find $P(x = 3)$.

Notation for the Poisson: $P = \text{Poisson}$ Probability Distribution Function

$$X \sim P(\mu)$$

Read this as “X is a random variable with a Poisson distribution.” The parameter is μ (or λ); μ (or λ) = the mean for the interval of interest.

Example 2:

Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call **in the next 15 minutes?**

Solution:

Let X = the number of calls Leah receives in 15 minutes. (The **interval of interest** is 15 minutes or 14 hour.)

$$x = 0, 1, 2, 3, \dots$$

If Leah receives, on the average, six telephone calls in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

calls in 15 minutes, on average. So, $\mu = 0.75$ for this problem.

$$X \sim P(0.75)$$

Find $P(x > 1)$. $P(x > 1) = 0.1734$ (calculator or computer)

- Press 1 - and then press 2nd DISTR.
- Arrow down to poissoncdf. Press ENTER.
- Enter (.75,1).
- The result is $P(x > 1) = 0.1734$.

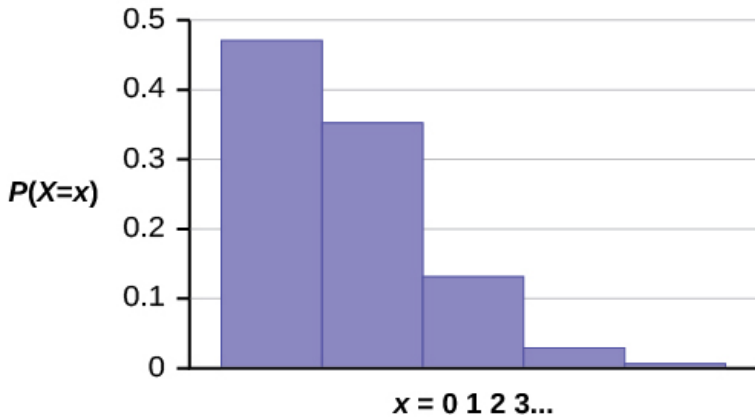
NOTE

The TI calculators use λ (lambda) for the mean.

The probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734:

$$P(x > 1) = 1 - \text{poissoncdf}(0.75, 1).$$

The graph of $X \sim P(0.75)$ is:



The y-axis contains the probability of x where X = the number of calls in 15 minutes.

Example 3:

According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let X = the number of emails an email user receives per day. The discrete random variable X takes on the values $x = 0, 1, 2, \dots$. The random variable X has a Poisson distribution: $X \sim P(147)$. The mean is 147 emails.

1. What is the probability that an email user receives exactly 160 emails per day?
2. What is the probability that an email user receives at most 160 emails per day?
3. What is the standard deviation?

Solution

1. $P(x = 160) = \text{poissonpdf}(147, 160) \approx 0.0180$
2. $P(x \leq 160) = \text{poissoncdf}(147, 160) \approx 0.8666$
3. Standard Deviation = $\sigma = \sqrt{\mu} = \sqrt{144} \approx 12.1244$

Review

A Poisson probability distribution of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is “small” (less than or equal to 0.05) and the number of trials is “large” (greater than or equal to 20).

Formula Review

$X \sim P(\mu)$ means that X has a Poisson probability distribution where X = the number of occurrences in the interval of interest.

X takes on the values $x = 0, 1, 2, 3, \dots$

The mean μ is typically given.

The variance is $\sigma^2 = \mu$, and the standard deviation is $\sigma = \sqrt{\mu}$.

When $P(\mu)$ is used to approximate a binomial distribution, $\mu = np$ where n represents the number of independent trials and p represents the probability of success in a single trial.





One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://library.achievingthedream.org/odessastatistics/?p=48#oembed-1>

PART V

NORMAL DISTRIBUTION

Apply the normal distribution and the Central Limit Theorem

Learning Outcomes

- Compute basic probabilities
- Apply discrete random variables
- Identify continuous random variables
- Apply the standard normal distribution and other normal distributions
- Interpret the Central Limit Theorem

29. Introduction to the Normal Distribution (Need Pic)



If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used

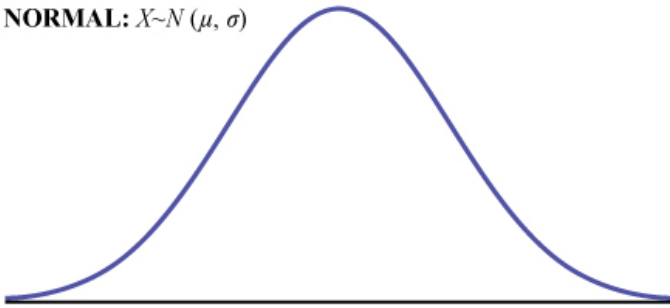
and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be

applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

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

NORMAL: $X \sim N(\mu, \sigma)$



The probability density function is a rather complicated function.

The curve is symmetrical about a vertical line drawn through the mean, μ .

In theory, the mean is the same as the median, because the graph is symmetric about μ .

As the notation indicates, the normal distribution depends only on the mean and the standard deviation.

Since the area under the curve must equal one, a change in the standard deviation, σ , causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ .

*****add a pic to describeeee

A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions.

*****add a pic to describeeee

One of special interest is called the **standard normal distribution**. The following video gives an example of data that would fall into a normal distribution.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=50#oembed-1>

COLLABORATIVE CLASSROOM ACTIVITY

We will record the heights of 200 men and 200 women, separately. Draw histograms of your data. Then draw a smooth curve through

each bar.

1. Is each curve somewhat bell-shaped?
2. Calculate the mean for each data set. Write the means on the x-axis of the appropriate graph below the peak.
3. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches in blue.
4. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches in red.
5. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

$$X \sim N(\mu, \sigma)$$

μ = the mean σ = the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}, \text{ where } \mu \text{ is the mean of}$$

the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

30. 6.1 The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z-scores**. A **z-score is measured in units of the standard deviation**. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$$

The z-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x - \mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ .

The following two videos give a description of what it means to have a data set that is “normally” distributed.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=51#oembed-1>



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=51#oembed-2>

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

$$z = \frac{x - \mu}{\sigma}$$

The z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ .

- Values of x that are larger than the mean have positive z-scores.
- Values of x that are smaller than the mean have negative z-scores.
- If x equals the mean, then x has a z-score of zero.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=51#oembed-3>

Note:

- The z-scores for $\mu+1\sigma$ and $\mu-1\sigma$ are +1 and -1, respectively.
 - The z-scores for $\mu+2\sigma$ and $\mu-2\sigma$ are +2 and -2, respectively.
 - The z-scores for $\mu+3\sigma$ and $\mu-3\sigma$ are +3 and -3 respectively.
-
-

Example 1

Suppose $X \sim N(5, 6)$.

This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$.

1. Suppose there is a raw data, x of 17. What is the z-score?

Show Answer

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is two standard deviations (2σ) above or to the right of the mean $\mu = 5$.

(given that the standard deviation is $\sigma = 6$.)

*** Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$) ***

2. What is the z-score for raw data $x = 1$?

Show Answer

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$.

Notice that: $5 + (-0.67)(6)$ is approximately equal to 1. (This has the pattern $\mu + z\sigma = \text{raw data.}$)

Summarizing,

- When z is positive, x is greater than or to the right of μ .
(when x is greater than μ , the corresponding z -score is positive,.)
 - When z is negative, x is less than or to the left of μ .
(When x is less than μ , the corresponding z -score is negative.)
-

Try It

What is the z -score of x , when $x = 1$ and $X \sim N(12,3)$?

Show Answer

$$z = \frac{1 - 12}{3} = -3.67$$

Example 2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal

distribution. Let X = the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

1. Suppose a person lost ten pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).

Show Answer

This z-score tells you that data 10 is 2.5 standard deviations to the **right** of the mean **five**.

2. Suppose a person gained three pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that data $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Show Answer

$z = -4$. This z-score tells you that data $x = -3$ is **4** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$.

If $x = 17$, then $z = 2$. (This was previously shown.)

If $y = 4$, what is z ?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1}.$$

The z-score for $y = 4$ is 2.

This means that raw data 4 is 2 standard deviations to the right of the mean.

Therefore, $x = 17$ and $y = 4$ are both two (**of their own**) standard deviations to the right of their respective means.

The z-score allows us to compare data that are scaled differently.

To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. (A negative weight gain would be a weight loss.)

Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain relative to their means.

Try It

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z-score when $x = 10$ is -1.5 . This score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

[practice-area rows="1"][/practice-area]

Show Answer

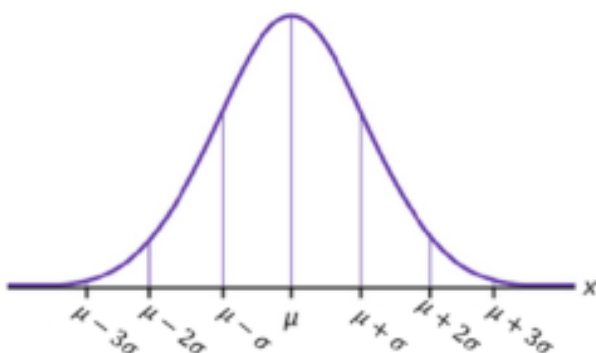
1.5, left, 16

The Empirical Rule

If X is a random variable and has a **normal distribution** with mean μ and standard deviation σ , then the **Empirical Rule** says the following:

- About 68% of the x values lie between the range between $\mu - \sigma$ and $\mu + \sigma$ (within one standard deviation of the mean).
 - About 95% of the x values lie between the range between $\mu - 2\sigma$ and $\mu + 2\sigma$ (within two standard deviations of the mean).
 - About 99.7% of the x values lie between the range between $\mu - 3\sigma$ and $\mu + 3\sigma$ (within three standard deviations of the mean).
- Notice that almost all the x -values/data lie within three standard deviations of the mean.

The empirical rule is also known as the 68-95-99.7 rule.



Example 3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm.

Male heights are known to follow a **normal distribution**.

Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010.

The z-score when $x = 168$ cm is $z =$ _____.

This z-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____.

Show Answer

a. -0.32, 0.32, left, 170

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$.

What is the male's height?

The z-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Show Answer

b. 177.98, 1.27, right

Try It

Use the information in Example 3 to answer the following questions.

- Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z = \underline{\hspace{2cm}}$. This z-score tells you that $x = 176$ cm is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean $\underline{\hspace{2cm}}$ (What is the mean?).
- Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = -2$. What is the male's height? The z-score ($z = -2$) tells you that the male's height is $\underline{\hspace{2cm}}$ standard deviations to the $\underline{\hspace{2cm}}$ (right or left) of the mean.

[practice-area rows="3"][/practice-area]

Show Answer

Solve the equation $z = \frac{x - \mu}{\sigma}$ for x . $x = \mu + (z)(\sigma)$

for x . $x = \mu + (z)(\sigma) = \frac{176 - 170}{0.96}$, This z-score tells

you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm. $X = 157.44$ cm, The z-score ($z = -2$) tells you that the male's height is two standard deviations to the left of the mean.

Example 4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

1. Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm.
2. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Show Answer

The z-score for $x = 160.58$ is $z = -1.5$.

The z-score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Try It

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

[practice-area rows="3"][/practice-area]

Show Answer

The z-score for $x_1 = 325$ is $z_1 = -1.14$. The z-score for $x_2 = 366.21$ is $z_2 = -1.14$. Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example 5

Suppose x has a normal distribution with mean 50 and standard

deviation 6.

Use empirical rule to interpret the data distribution.

Show Answer

- About 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50.

The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50.

The z-scores are -1 and $+1$ for 44 and 56, respectively.

- About 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$.

The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations of the mean 50.

The z-scores are -2 and $+2$ for 38 and 62, respectively.

- About 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50.

The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50.

The z-scores are -3 and $+3$ for 32 and 68, respectively.

Try It

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Show Answer

Between 20 and 30.

Example 6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm.

Let Y = the height of 15 to 18-year-old males in 1984 to 1985, $Y \sim N(172.36, 6.34)$.

(The variable Y is normally distributed.)

1. About 68% of the y values lie between what two values?
These values are _____. The z -scores are _____, respectively.
2. About 95% of the y values lie between what two values?
These values are _____. The z -scores are _____ respectively.
3. About 99.7% of the y values lie between what two values?
These values are _____. The z -scores are _____ respectively.

Show Answer

Since the variable Y is normally distributed, we can apply the Empirical Rule.

1. $\mu - 1\sigma = 172.36 - 6.34 = 166.02$; $\mu + 1\sigma = 172.36 + 6.34 = 178.7$
About 68% of the y values lie between 166.02 cm and 178.7 cm.
The z -scores are -1 and 1, respectively.
2. $\mu - 2\sigma = 172.36 - 2(6.34) = 159.68$; $\mu + 2\sigma = 172.36 + 2(6.34) = 185.04$
About 95% of the y values lie between 159.68 cm and 185.04 cm.
The z -scores are -2 and 2, respectively.
3. $\mu - 3\sigma = 172.36 - 3(6.34) = 153.34$; $\mu + 3\sigma = 172.36 + 3(6.34) = 191.38$

About 99.7% of the y values lie between 153.34 cm and 191.38 cm.

The z-scores are -3 and 3, respectively.

Try It

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

[practice-area rows="4"][/practice-area]

Show Answer

About 68% of the values lie between the values 41 and 63.

About 95% of the values lie between the values 30 and 74.

About 99.7% of the values lie between the values 19 and 85.

.

References

“Blood Pressure of Males and Females.” StatCrunch, 2013. Available online at <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960> (accessed May 14, 2013).

“The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores.” London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Concept Review

A z-score is a standardized value. Its distribution is the standard

normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ .

Formula Review

$Z \sim N(0, 1)$

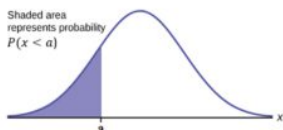
z = a standardized value (z-score)

mean = 0; standard deviation = 1

To convert z-score into raw data: $raw\ data = \mu + (z)\sigma$

To convert data into z-score: $z = \frac{x - \mu}{\sigma}$

3I. 6.2 Using the Normal Distribution



The **blue shaded area** in the following graph indicates the area to the left of x .

$P(X < a)$ = Area to the left of the vertical line through a . (blue area)

$P(X > a)$ = Area to the right of the vertical line through a . (white area)

$P(X > a)$ is also equal to $1 - P(X < a)$.

Remember,

$P(X < a) = P(X \leq a)$ and $P(X > a) = P(X \geq a)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.

Additionally, [this link](#) houses a tool that allows you to explore the normal distribution with varying means and standard deviations as well as associated probabilities. The following video explains how to use the tool.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=52#oembed-1>

Example 1

If the area to the left is 0.0228, what is the area to the right?

Show Answer

the area to the right is $1 - 0.0228 = 0.9772$.

Try It

If the area to the left of x is 0.012, then what is the area to the right?

[practice-area rows="2"][/practice-area]

Show Answer

$1 - 0.012 = 0.988$

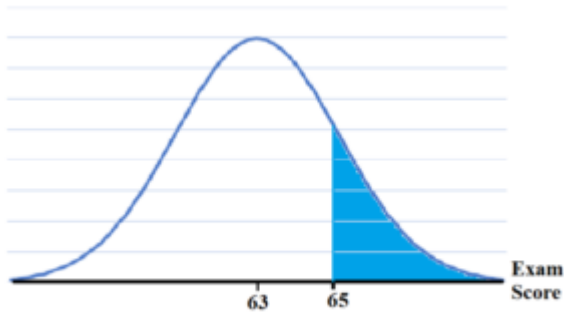
Example 2 (by using TI-83/84 Calculators)

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

1. Find the probability that a randomly selected student scored more than 65 on the exam.
2. Find the probability that a randomly selected student scored less than 85.
3. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).
4. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution

1. Find the probability that a randomly selected student scored more than 65 on the exam. Let X = a score on the final exam. $X \sim N(63, 5)$, where mean, $\mu = 63$ and standard deviation, $\sigma = 5$.
 1. Draw a graph. Then, find $P(x > 65)$.



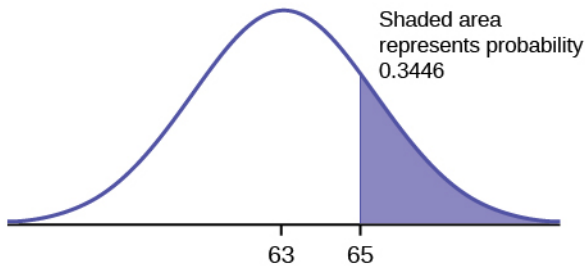
2. Using TI-Calculator:

Go into 2nd DISTR. Then press 2:normalcdf.

The syntax for the instructions are as follows: normalcdf (lower value, upper value, mean, standard deviation).

For this problem: normalcdf (65,10⁹⁹,63,5) and press “=”.
The answer is 0.3446.

3. Therefore, $P(x > 65) = 0.3446$.



The probability that any student selected at random scores more than 65 is 0.3446.

Extra Note:

The number 10⁹⁹ is way out in the right tail of the normal

curve. We are calculating the area when final exam score is between 65 and 10⁹⁹.

In some instances, the lower number of the area might be -10⁹⁹.

The number -10⁹⁹ is way out in the left tail of the normal curve.

$$z = \frac{65 - 63}{5} = 0.4$$

We can solve that area to the left of z-score 0.4 is 0.6554 by inputting normalcdf(-10⁹⁹, 65, 63, 5) to TI-Calculator.

Hence, $P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$

```
*****
*****
*****
```

To calculate the z-score:

Press 2nd ---> Distr, and choose 3:invNorm(*Enter the area to the left of z followed by), and press ENTER.

For this Example, the steps are

2nd ---> Distr ---> 3:invNorm(.6554) -> ENTER.

The answer is 0.3999 which rounds to 0.4.

```
*****
*****
*****
```

- Find the probability that a randomly selected student scored less than 85. Draw a graph to find $P(x < 85)$, and shade the graph. Using a computer or calculator, find $P(x < 85)$. Use calculator to try first before looking for the answer!

TI-Calculator: normalcdf(-10⁹⁹, 85, 63, 5). The answer is 1. (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

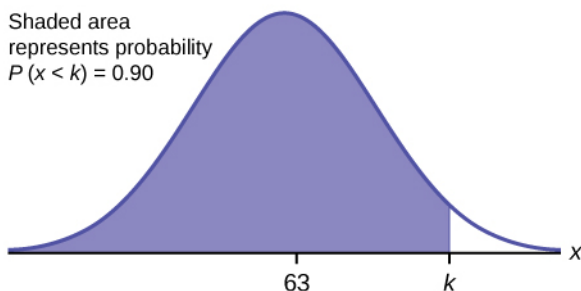
3. Find the 90th percentile. For each problem or part of a problem, draw a new graph.

Draw the x -axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. The variable k is located on the x -axis. $P(x < k)$ is the area to the left of k .

The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k , and ten percent are the same or higher.

The variable k is often called a **critical value**.



To get the answer on the calculator, follow this step: 2nd \rightarrow DISTR \rightarrow invNorm.

The format for this function is invNorm(area to the left, mean, standard deviation)

For this problem, we need to input "invNorm(0.90,63,5)" and press "=".

The answer is 69.4.

So the 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above.

4. Find the 70th percentile. Draw a new graph and label it appropriately. $k = 65.6$.

Use calculator to try first before looking for the answer!

$\text{invNorm}(0.70, 63, 5) = 65.6$

The 70th percentile is 65.6.

This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

Try It

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of 3.

Find the probability that a randomly selected golfer scored less than 65.

[practice-area rows="3"][/practice-area]

Show Answer

TI-Calculator: $\text{normalcdf}(1099, 65, 68, 3)$

The probability that a randomly selected golfer scored less than 65 is 0.1587.

Example 3

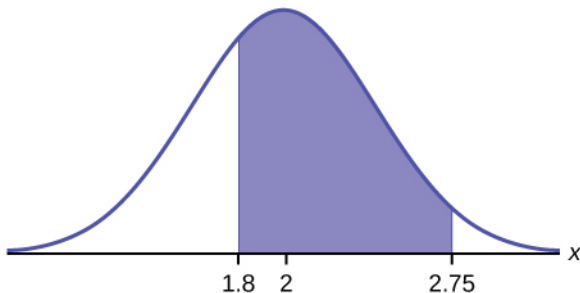
A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

1. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
2. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution

1. Let X = the amount of time (in hours) a household personal computer is used for entertainment.

$X \sim N(2, 0.5)$ where mean, $\mu = 2$ and standard deviation, $\sigma = 0.5$. Find $P(1.8 < x < 2.75)$. The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$.



TI-Calculator: $\text{normalcdf}(1.8, 2.75, 2, 0.5)$. The answer is 0.5886.

$$P(1.8 < x < 2.75) = 0.5886.$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

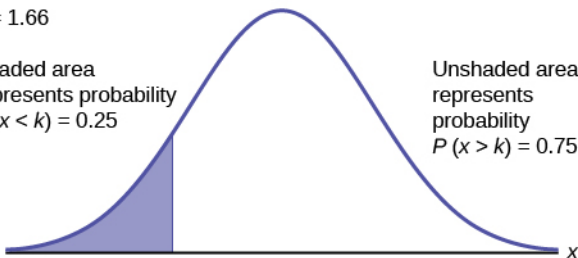
2. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, we need to **find the 25th percentile, k** .

$$P(x < k) = 0.25.$$

$$k = 1.66$$

Shaded area
represents probability
 $P(x < k) = 0.25$

Unshaded area
represents
probability
 $P(x > k) = 0.75$



$$\text{TI-Calculator: invNorm}(0.25, 2, 0.5) = 1.66.$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Try It

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

[practice-area rows="3"][/practice-area]

Show Answer

TI-Calculator: normalcdf(66,70,68,3)

The probability that a golfer scored between 66 and 70 is 0.4950

Example 4

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

1. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
2. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
3. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Solution:

Show Answer

1. $\text{normalcdf}(23,64.7,36.9,13.9) = 0.8186$

2. $\text{normalcdf}(-1099, 50.8, 36.9, 13.9) = 0.8413$

3. $\text{invNorm}(0.80, 36.9, 13.9) = 48.6$

The 80th percentile is 48.6 years.

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Try It

Use the information in Example 3 to answer the following questions.

1. Find the 30th percentile, and interpret it in a complete sentence.
2. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old.

Solution:

Let X = a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

1. To find the 30th percentile, find k such that $P(x < k) = 0.30$.

Show Answer

TI-Calculators: $\text{invNorm}(0.30, 36.9, 13.9)$

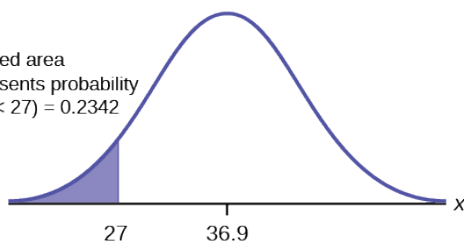
The solution is 29.6 years.

30% of smartphone users aged between 13 and

55+ are at most 29.6 years and 70% are at least 29.6 years.

2. Find $P(x < 27)$.

Shaded area
represents probability
 $P(x < 27) = 0.2342$



Show Answer

TI-Calculators: $\text{invNorm}(0.30, 36.9, 13.9)$

The solution is 29.6 years.

30% of smartphone users aged between 13 and 55+ are at most 29.6 years and 70% are at least 29.6 years.

(Note that $\text{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$. The two answers differ only by 0.0040.)

Example 5

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively.

Using this information, answer the following questions (round answers to one decimal place).

1. Calculate the interquartile range (IQR).
2. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution

1. $IQR = Q_3 - Q_1$

To find Q_3 (also known as 75th percentile), we input $\text{invNorm}(0.75, 36.9, 13.9)$ to TI-calculator. Hence $Q_3 = 46.2754$.

To find Q_1 (also known as 25th percentile), we input $\text{invNorm}(0.25, 36.9, 13.9)$ to TI-calculator. Hence $Q_1 = 27.5246$.

$IQR = Q_3 - Q_1 = 18.7508$.

2. Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.") $0.40 =$ the area to the right.

Area to the left $= 1 - 0.40 = 0.60$.

The area to the left of $k = 0.60$.

Input " $\text{invNorm}(0.60, 36.9, 13.9)$ " to TI-Calculator, The answer is 40.4215.

Hence, $k = 40.42$.

Forty percent of the ages that range from 13 to 55+ are at least 40.42 years.

Try It

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean

$\mu = 81$ points and standard deviation $\sigma = 15$ points.

1. Calculate the first- and third-quartile scores for this exam.

[practice-area rows="2"][/practice-area]

Show Answer

$Q1 = 25\text{th percentile} = \text{invNorm}(0.25, 81, 15) = 70.9$

$Q3 = 75\text{th percentile} = \text{invNorm}(0.75, 81, 15) = 91.9$

2. The middle 50% of the exam scores are between what two values?

[practice-area rows="2"][/practice-area]

Show Answer

The middle 50% of the scores are between 70.9 and 91.1.

Example 6

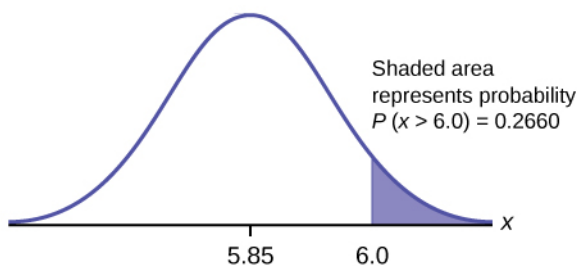
A citrus farmer who grows mandarin oranges finds that the

diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

1. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
2. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
3. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Solution

1. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$



2. $1 - 0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

The middle 20% means range from 40% to 60%.

Find k_1 , the 40th percentile, and k_2 , the 60th percentile.

$k_1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79 \text{ cm}$

$k_2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91 \text{ cm}$

3. 6.15: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Try It

Using the information from Example 5, answer the following:

1. The middle 40% of mandarin oranges from this farm are between _____ and _____.
2. Find the 16th percentile and interpret it in a complete sentence.

Solution:

1. Show Answer

The middle area = 0.40, so each tail has an area of 0.30.

Find k_1 , the 30th percentile and k_2 , the 70th percentile.

$$k_1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72 \text{ cm}$$

$$k_2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98 \text{ cm}$$

2. Show Answer

$$\text{normalcdf}(5, 1099, 5.85, 0.24) = 0.9998$$

References

“Naegle’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegle's_rule (accessed May 14, 2013).

“403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at <http://www.thisamericanlife.org/radio-archives/episode/403/nummi> (accessed May 14, 2013).

“Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at <http://www.winatthelottery.com/public/departments40.cfm> (accessed May 14, 2013).

“Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at <http://visual.ly/smart-phone-users-numbers> (accessed May 14, 2013).

“Facebook Statistics.” Statistics Brain. Available online at <http://www.statisticbrain.com/facebook-statistics/> (accessed May 14, 2013).

Concept Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean

μ and the standard deviation σ . A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

Normal Distribution:

$X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.

Standard Normal Distribution:

$Z \sim N(0, 1)$.

Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)

Calculator function for the k th percentile: $k = \text{invNorm}$ (area to the left of k , mean, standard deviation)

PART VI

THE CENTRAL LIMIT THEOREM

32. Introduction: The Central Limit Theorem



If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.
- Apply and interpret the central limit theorem for sums.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

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

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distribution of sample means and the sums tend to follow the normal distribution.

The size of the sample, n , that is required in order to be “large enough” depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should

come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement.**



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=54#oembed-1>

33. 7.1 The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution).

Using a subscript that matches the random variable, suppose:

1. μ_X = the mean of X
2. σ_X = the standard deviation of X

If you draw random samples of size n , then as n increases, the random samples \bar{X} which consists of sample means, tend to be normally distributed.

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

The **central limit theorem** for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and **calculating their means**, the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by, the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

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

The random variable \bar{X} has a different z-score associated with it

from that of the random variable X. The mean \bar{x} is the value of \bar{X} in one sample.

$$z = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

$\mu_x = \mu_{\bar{x}}$ (mean of X = mean of \bar{X} .)

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ = standard deviation of \bar{X} and is called the

standard error of the mean.

Guide for TI-Calculator:

To find **probabilities** for means on the TI-calculator, follow these steps:

- “2nd”
- “DISTR”
- “2: normalcdf”
- normalcdf (lower value, upper value, mean, $\frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$)

where: mean is the mean of the original distribution, standard deviation is the standard deviation of the original distribution
sample size

Example 1

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

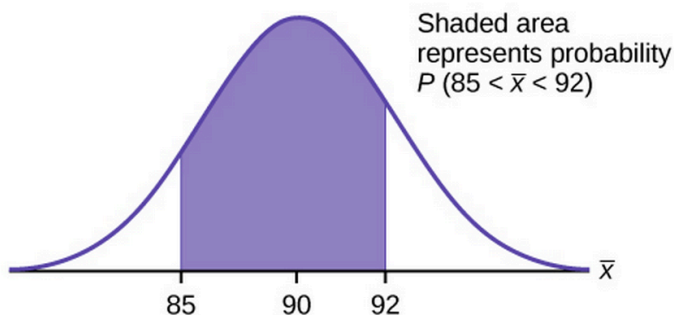
1. Find the probability that the **sample mean** is between 85 and 92.
2. Find the value that is two standard deviations above the

expected value, 90, of the sample mean.

Solution

lower value = 85, upper value = 92, mean $\mu = 90$, std dev $\sigma = 15$, sample size = 25.

1. In this example, the probability that the sample mean is btw 85 and 92 = area btw 85 and 92.



Since 25 samples are drawn, $\mu_x = \text{mean} = 90$, $\sigma_x = \frac{\text{std dev}}{\text{sample size}} = \frac{15}{\sqrt{25}}$. By using TI-83/84, $\text{normalcdf}(85, 92, 90, \frac{15}{\sqrt{25}}) = 0.0697$

2. To find the value that is two standard deviations above the expected value 90, use the formula:

$$\text{value} = \mu_x + (\# \text{ of STD DEV}) \left(\frac{\sigma_x}{\sqrt{n}} \right)$$

Value that is 2 std dev above 90

$$= 90 + 2 \left(\frac{15}{\sqrt{25}} \right)$$

$$= 96$$

The value that is two standard deviations above the expected

value is 96.

(Note: The standard error of the mean is $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$.)

Recall that the standard error of the mean ($\frac{\sigma}{\sqrt{n}}$) is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n .



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=55#oembed-1>

Try It

An unknown distribution has a mean of 45 and a standard deviation of 8. Samples of size $n = 30$ are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

[practice-area rows="2"][/practice-area]

Show Answer

TI-Calculator: $\text{normalcdf}(42, 50, 45, \frac{8}{\sqrt{30}})$

$P(42 < \bar{x} < 50) = 0.9797$

Example 2

The length of time, in hours, it takes an “over 40” group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size $n = 50$** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution:

In this example, mean $\mu = 2$, std dev $\sigma = 0.5$, sample size $n = 50$

Let \bar{X} = the **mean** time, in hours, it takes to play one soccer match.

We are looking for $P(1.8 < \bar{x} < 2.3)$.

TI-Calculator: $\text{normalcdf}(1.8, 2.3, 2, \frac{0.5}{\sqrt{50}})$

The probability that the mean time is between 1.8 hours and 2.3 hours

$$= P(1.8 < \bar{x} < 2.3)$$

$$= 0.9977.$$

Try It

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

[practice-area rows="1"][/practice-area]

Show Answer

$$\text{normalcdf}(2, 3, 2.5, \frac{0.25}{\sqrt{60}})$$

$$P(2 < \bar{x} < 3) = 1$$

Guide for TI-Calculator:

To find **percentiles** for means on the calculator, follow these steps.

- 2nd DISTR
- 3:invNorm

- $k = \text{invNorm}(\text{area to the } \underline{\text{LEFT}} \text{ of } k, \text{ mean}, \frac{\text{standard deviation}}{\sqrt{\text{sample size}}})$

where: k = the k^{th} percentile, *mean* is the mean of the original distribution, *standard deviation* is the standard deviation of the original distribution

Example 3

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

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

Solution

In this example, mean $\mu = 34$ years, std dev $\sigma = 15$ years, sample size $n = 100$

1. Since the sample mean tends to target the population mean, the mean for the sample mean ages of tablet users $\mu_X = \mu = 34$.

The sample standard deviation for the sample mean ages is

$$\text{given by } \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

2. The central limit theorem states that for large sample sizes(n), the sampling distribution will be approximately normal.

3. TI-Calculator: `normalcdf(30,1E99,34,1.5)`

The probability that the sample mean age is more than 30 = $P(X > 30) = 0.9962$

4. Let k = the 95th percentile.

$$\text{TI-Calculator: } \text{invNorm}(0.95, 34, \frac{15}{\sqrt{100}})$$

$k = 95^{\text{th}}$ percentile = 36.5.

Try It

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s

through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Show Answer

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game (also known as $P(29 < \bar{x} < 35)$.)

TI-Calculator: normalcdf = 0.0186 (29, 35, 28, $\frac{4.8}{\sqrt{100}}$)

$$P(29 < \bar{x} < 35) = 0.0186$$

You can conclude there is approximately a 2% chance that your game will be played by men whose mean age is between 29 and 35.

Example 4

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

1. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
2. What is the standard error of the mean?
3. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
4. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution

In this example, mean $\mu = 8.2$, std dev $\sigma = 1$, sample size $n = 60$

1. The mean for the sample mean number of app engagement by a tablet user = $\mu_{\bar{x}} = \mu = 8.2$.
2. The std dev for the sample mean number of app engagement by a tablet user = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$

This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.

3. Let k = the 90th percentile.

TI-Calculator: invNorm(0.9, 8.2, $\frac{1}{\sqrt{60}}$)

k = the 90th percentile = 8.37.

90 percent of the average app engagement time for table users is less than 8.37 minutes.

4. TI-Calculator: $\text{normalcdf}(8, 8.5, 8.2 \frac{1}{\sqrt{60}})$

$$P(8 < \bar{x} < 8.5) = 0.9293$$



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=55#oembed-2>

Try It

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are $n = 34, \bar{x} = 16.01$ ounces. If the cans are filled so that $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Show Answer

Ti-Calculator: $\text{normalcdf}(16.01, 1E99, 16, \frac{0.143}{\sqrt{34}})$

$$P(\bar{x} > 16.01) = 0.3417$$

Since there is a 34.17% probability that the average

sample weight is greater than 16.01 ounces, we should be skeptical of the company's claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

References

Baran, Daya. "20 Percent of Americans Have Never Used Email." WebGuild, 2010. Available online at <http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email> (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

Data from the United States Department of Agriculture.

Concept Review

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the

standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

Formula Review

The Central Limit Theorem for Sample Means: $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

34. 7.2 The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution) and suppose:

1. μ = the mean of X
2. σ = the standard deviation of X

If you draw random samples of size n , then as n increases, the random variable Σx consisting of sums tends to be **normally distributed** such that $\Sigma x \sim N[(n)(\mu), (\sqrt{n})(\sigma)]$.

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable Σx has the following z-score associated with it:

1. Σx is one sum.

$$2. \quad z = \frac{\Sigma x - (n)(\mu)}{(\sqrt{n})(\sigma)} \quad \text{or}$$

$$\Sigma x = (n)(\mu) + (z)(\sigma)(\sqrt{n})$$

(Do not memorize both formula. They are same!)

- $(n)(\mu) = \mu_{\Sigma x}$, the mean of Σx
- $(\sqrt{n})(\sigma) = \sigma_{\Sigma x}$, the standard deviation of Σx

Guide for TI-Calculator

To find probabilities for sums on the calculator, follow these steps.

- 2nd
 - DISTR
 - 2: normalcdf
 - normalcdf (lower value of the area, upper value of the area, n * mean, \sqrt{n} * standard deviation)
- where: *mean* is the mean of the original distribution *standard deviation* is the standard deviation of the original distribution
sample size = n
-
-
-

Example 1

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

1. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500.
2. Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution

Let X = one value from the original unknown population.

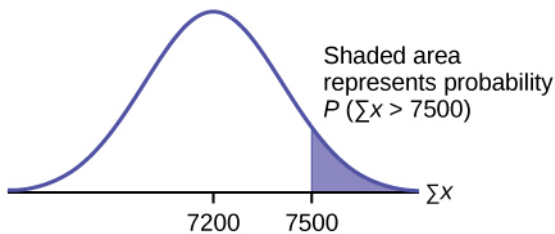
The probability question asks you to find a probability for **the sum (or total of) 80 values**.

$\sum x$ = the sum or total of 80 values, $\mu = 90$, $\sigma = 15$, and $n = 80$,

- mean of the sums, $\mu_{\sum x} = (n)(\mu) = (80)(90) = 7,200$
- standard deviation of the sums, $\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{80})(15)$

Therefore, $\sum x \sim N((80)(90), (\sqrt{80})(15))$.

1. Probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500
 $= P(\sum x > 7,500)$
 $= \text{Shaded area}$



TI-Calculator: $\text{normalcdf}(7500, 1E99, (80)(90), (\sqrt{80})(15)) = 0.0127$

Therefore, $P(\sum x > 7,500) = 0.0127$

2. Find $\sum x$ where $z = 1.5$.

$$\sum x = (n)(\mu) + (z)(\sqrt{n})(\sigma) = (80)(90) + (1.5)(\sqrt{80})(15) = 7401.2$$

Try It

An unknown distribution has a mean of 45 and a standard deviation of 8. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

[practice-area rows="1"][/practice-area]

Show Answer

$\sum x$ = the sum / total of 50 values,
 $\mu = 45$, $\sigma = 8$, and $n = 50$,

- mean of the sums = $(n)(\mu) = (50)(45)$
- standard deviation of the sums =

$$(\sqrt{n})(\sigma) = (\sqrt{50})(8)$$

TI-Calculator: normalcdf (2400, 1E99, (50)(45), $(\sqrt{50})(8)$)

Probability (the sum of the 50 values is more than 2,400)
= 0.0040

Guide for TI-Calculator

To find percentiles for sums on the calculator, follow these steps.

- 2nd
 - DIST
 - 3: invNormk = invNorm (area to the left of k , $(n)(\mu)$, $(\sqrt{n})(\sigma)$) where:
 k is the k th **percentile**,
 μ is the mean of the original distribution,
 σ is the standard deviation of the original distribution,
sample size = n
-

-

Example 2

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. The sample of size is 50.

1. What are the mean and standard deviation for the sum of the ages of tablet users?
2. What is the distribution?
3. Find the probability that the sum of the ages is between 1,500 and 1,800 years.
4. Find the 80th percentile for the sum of the 50 ages.

Solution

In this example, $\mu = 34$, $\sigma = 15$, and $n = 50$,

- mean of the sums, $\mu_{\sum x} = (n)(\mu) = (50)(34)$

- standard deviation of the sums, $\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{50})(15)$

1. $\mu_{\sum x} = n\mu = 50(34) = 1,700$ and

$$\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{50})(15) = 106.01$$

2. The distribution for the sum of ages of tablet users is normal by the central limit theorem.

3. TI-Calculator: normalcdf(1500, 1800, (50)(30), $(\sqrt{50})(15)$)
 $P(1500 < \sum x < 1800) = 0.7974$

4. Let k = the 80th percentile.

TI-Calculator: invNorm(0.80, (50)(34), $(\sqrt{50})(15)$)

k = 1789.3

Try It

In a recent study reported Oct.29, 2012 on the Flurry Blog, the mean age of tablet users is 35 years. Suppose the standard deviation is 10 years. The sample size is 39.

1. What are the mean and standard deviation for the sum of the ages of tablet users?

[practice-area rows="1"][/practice-area]

Show Answer

$$\mu_{\sum x} = n\mu = 1365 \text{ and}$$

$$\sigma_{\sum X} = \sigma = (\sqrt{n\sigma_x})(15) = 62.4.$$

2. What is the distribution?

[practice-area rows="1"][/practice-area]

Show Answer

The distribution is normal for sums by the central limit theorem.

3. Find the probability that the sum of the ages is between 1,400 and 1,500 years.

[practice-area rows="1"][/practice-area]

Show Answer

TI-Calculator: normalcdf(1400, 1500, (39)(35), 10)

$$P(1400 < \sum x < 1500) = 0.2723$$

4. Find the 90th percentile for the sum of the 39 ages.

[practice-area rows="1"][/practice-area]

Show Answer

Let k = the 90th percentile.

TI-Calculator: invNorm(0.90, (39)(35), $(\sqrt{39})(10)$)

$$k = 1445.0$$

Example 3

The mean number of minutes for app engagement by a tablet user

is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of size 70.

1. What are the mean and standard deviation for the sums?
2. Find the 95th percentile for the sum of the sample. Interpret this value in a complete sentence.
3. Find the probability that the sum of the sample is at least ten hours.

Solution

In this example, $\mu = 8.2$ minutes, $\sigma = 1$ minute, and $n = 70$,

- mean of the sums, $\mu_{\sum x} = (n)(\mu) = (70)(8.2)$
- standard deviation of the sums, $\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{70})(1)$

1. $\mu_{\sum x} = (n)(\mu) = 70(8.2) = 574$ minutes and $\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{70})(1) = 8.37$ minutes.
2. Let k = the 95th percentile.,
TI-Calculator: invNorm (0.95,(70)(8.2), $(\sqrt{70})(1)$)
 $k = 587.76$ minutes.
95% of the app engagement times are at most 587.76 minutes.
3. Covert 10 hours into 600 minutes.
TI-Calculator: normalcdf(600, 1E99,(70)(8.2), $(\sqrt{70})(1)$)
 $P(\text{the sum of the sample is at least ten hours}) = P(\sum x > 600 \text{ minutes}) = 0.0009$

Example 4

The mean number of minutes for app engagement by a table use is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

1. What is the probability that the sum of the sample is between seven hours and ten hours? What does this mean in context of the problem?
2. Find the 16th and 84th percentiles for the sum of the sample. Interpret these values in context.

Solution

In this example, $\mu = 8.2$ minutes, $\sigma = 1$ minute, and $n = 70$,

- mean of the sums, $\mu_{\sum x} = (n)(\mu) = (70)(8.2)$
- standard deviation of the sums, $\sigma_{\sum x} = (\sqrt{n})(\sigma) = (\sqrt{70})(1)$

1. 7 hours = 420 minutes and 10 hours = 600 minutes
TI-Calculator: $\text{normalcdf}(420, 600, (70)(8.2), \sqrt{70}(1))$
P (the sum of the sample is between seven hours and ten hours)
 $= P(420 \leq \sum x \leq 600)$
 $= 0.9991$.

This means that for this sample sums there is a 99.91% chance that the sums of usage minutes will be between 420 minutes and 600 minutes.

2. To find the 16th percentile, TI-Calculator: $\text{invNorm}(0.16, (70)(8.2), (\sqrt{70})(1)) = 565.68$ minutes.
To find the 84th percentile, TI-Calculator: invNorm

$(0.84, (70)(8.2), (\sqrt{70})(1)) = 582.32$ minutes. Since 84% of the app engagement times are at most 582.32 minutes and 16% of the app engagement times are at most 565.68 minutes, we may state that 68% of the app engagement times are between 565.68 minutes and 582.32 minutes.

References

Farago, Peter. "The Truth About Cats and Dogs: Smartphone vs Tablet Usage Differences." The Flurry Blog, 2013. Posted October 29, 2012. Available online at <http://blog.flurry.com> (accessed May 17, 2013).

Concept Review

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_X and a standard deviation of σ_X , the mean of the sums is $n\mu_X$ and the standard deviation is $(\sqrt{n})(\sigma_X)$ where n is the sample size.

Formula Review

The Central Limit Theorem for Sums:

$$\sum X \sim N[(n)(\mu), (\sqrt{n})(\sigma)]$$

The Central Limit Theorem for Sums z-score and standard deviation for sums:

- z for the sample mean of the sums: $z = \frac{\sum x - (n)(\mu)}{(\sqrt{n})(\sigma)}$
- Mean for Sums, $\mu_{\sum x} = (n)(\mu_x)$
- Standard deviation for Sums, $\sigma_{\sum x} = (\sqrt{n})(\sigma_x)$

35. 7.3 Using the Central Limit Theorem

It is important for you to understand when to use the **central limit theorem**. If you are being asked to find the probability of the mean, use the clt for the mean. If you are being asked to find the probability of a sum or total, use the clt for sums. This also applies to percentiles for means and sums.

Note

If you are being asked to find the probability of an **individual** value, do **not** use the clt. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger size from any population, then the mean \bar{x} of the sample tends to get closer and closer to the population mean μ .

The formula for the standard deviation of variable \bar{x} is $\frac{\sigma}{\sqrt{n}}$. If n is getting larger, then $\frac{\sigma}{\sqrt{n}}$ is getting smaller. Indirectly, the sample mean \bar{x} will be closer to the population mean μ .

We can say that μ is the value that the sample means approach as n gets larger.

The central limit theorem illustrates the law of large numbers.

Central Limit Theorem for the Mean and Sum

Example 1

A study involving stress is conducted among the students on a college campus.

The stress scores follow a [uniform distribution](#) with the lowest stress score = 1 and the highest score = 5.

Using a sample of 75 students, find

- The probability that the **mean stress score** for the 75 students is less than two.
- The 90th percentile for the **mean stress score** for the 75 students.
- The probability that the **total of the 75 stress scores** is less than 200.
- The 90th percentile for the **total stress score** for the 75 students.

Solution

Let X = one stress score. The sample size $n = 75$.

We are looking for a probability or a percentile for a **mean score** in problem a and b.

We are looking for a probability or a percentile for a **total or sum of score** in problem c and d.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where lowest score = 1 and highest score = 5.

$$\mu_X = \frac{a + b}{2} = \frac{1 + 5}{2} = 3$$
$$\sigma_X = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{\frac{(5 - 1)^2}{12}} = 1.15$$

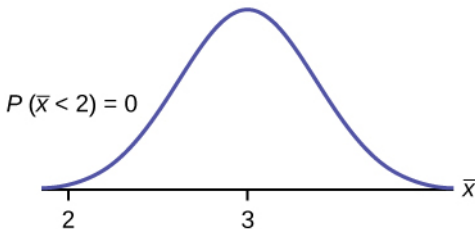
For problems a and b, let \bar{X} = the mean stress score for the 75 students.

Then, $\bar{X} \sim N(3, \frac{1.15}{\sqrt{75}})$ where $n = 75$.

Solution

a. Find the probability that the **mean stress score** for the 75 students is less than two.

We are asked to find $P(\bar{x} < 2)$. By plotting the graph,



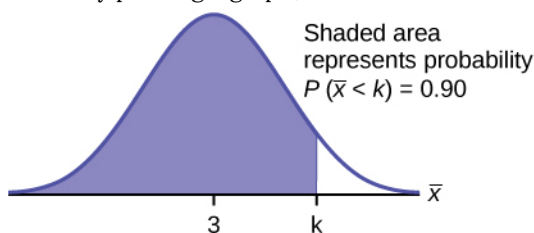
We will use TI-83/84 to solve for part (a).

$$\text{TI-Calculator: normalcdf} \left(1, 2, 3, \frac{1.15}{\sqrt{75}} \right) = 0$$

Remember that the smallest stress score is one.

b. Find the 90th percentile for the **mean stress score** for the 75 students.

We are asked to find the 90th percentile for the mean of 75 stress scores. By plotting a graph,



Let k = the 90th percentile. Find the value of k where $P(\bar{x} < k) = 0.90$.

$$\text{TI-Calculator: invNorm} \left(0.90, 3, \frac{1.15}{\sqrt{75}} \right)$$

$$k = 3.2$$

The 90th percentile for the mean of 75 scores is about 3.2.

This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

For problems c and d, let ΣX = the sum of the 75 stress scores. Then,

$$\sum X \sim N[(75)(3), (\sqrt{75})(1.15)]$$

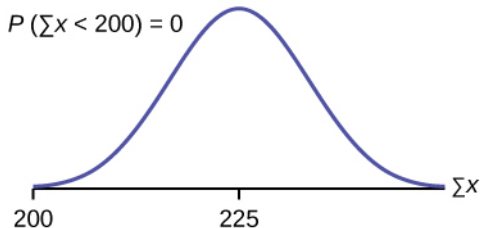
The mean of the sum of 75 stress scores is $(75)(3) = 225$.

The standard deviation of the sum of 75 stress scores is $(\sqrt{75})(1.15) = 9.96$

c. Find the probability that the **total of the 75 stress scores** is less

than 200.

We are asked to find $P(\Sigma x < 200)$. By plotting the graph,



TI-Calculator: `normalcdf(75, 200, (75)(3), ($\sqrt{75}$)(1.15))`

Therefore, $P(\Sigma x < 200) = 0$

The probability that the total of 75 scores is less than 200 is about zero..

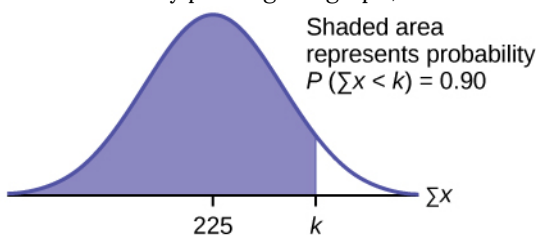
Remember,

since the smallest single score is 1, it is possible that we draw the smallest score of 1 for 75 times theoretically.

The smallest total of 75 stress scores is 75.

d. Find the 90th percentile for the **total stress score** for the 75 students.

We are asked to solve for the 90th percentile for the total of 75 stress scores. By plotting the graph,



Let k = the 90th percentile. Find the value of k where $P(\Sigma x < k) = 0.90$.

TI-Calculator: `invNorm(0.90, (75)(3), ($\sqrt{75}$)(1.15))`

$k = 237.8$

The 90th percentile for the sum of 75 scores is about 237.8.

This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8. = 237.8

Try It

A study involving stress is conducted among the students on a college campus.

The stress scores follow a uniform distribution with the lowest stress score = 1 and the highest score = 5.

Use a **sample size of 55** to answer the following questions.

1. Find $P(\bar{x} < 7)$.
2. Find $P(\sum x < 170)$.
3. Find $P(\sum x > 170)$.
4. Find the 80th percentile for the mean of 55 scores.
5. Find the 85th percentile for the sum of 55 scores.

[practice-area rows="2"][/practice-area]

Show Answer

$$1. \text{ TI-Calculator: normalcdf } (1, 7, 3, \frac{1.15}{\sqrt{55}}) = 0 =$$

1

$$2. \text{ TI-Calculator: normalcdf } (55, 170, (55)(3), (\sqrt{55})(1.15))$$

$$= 0.7211$$

$$3.1 - 0.7211 = 0.2789$$

$$4. \text{ TI-Calculator: } \text{invNorm}(0.80, 3, \frac{1.15}{\sqrt{55}}) = 3.13$$

$$5. \text{ TI-Calculator: } \text{invNorm}(0.85, (55)(3), \sqrt{55})(1.15) = 173.84$$

Example 2

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim \text{Exp}\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.

Let \overline{X} = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right) \text{ by the central limit theorem for sample}$$

means.

Using the clt to find probability,

1. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$. Draw the graph.
2. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
3. Explain why the probabilities in parts 1 and 2 are different.
4. Find the 95th percentile for the sample mean excess time for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

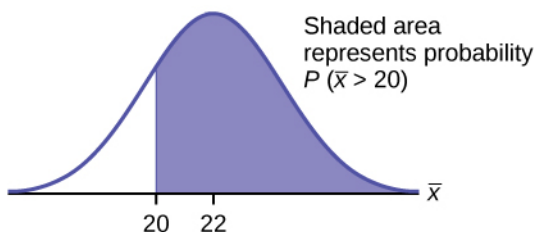
Solution

1. Find: $P(\bar{x} > 20)$

$$P(\bar{x} > 20) = 0.79199 \text{ using } \text{normalcdf}$$

$$(20, 1E99, 22, \frac{22}{\sqrt{80}})$$

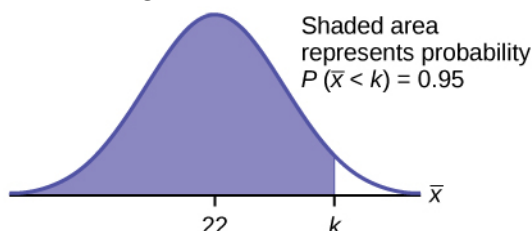
The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



Remember, $1\text{E}99 = 10^{99}$ and $-1\text{E}99 = -10^{99}$. Press the EE key for E.
Or just use 10^{99} instead of $1\text{E}99$.

2. Find $P(x > 20)$. Remember to use the exponential distribution for an individual: $X \sim \frac{1}{22}$.

$$P(x > 20) = e^{-(1/22)(20)} \quad \text{or} \quad e^{(-0.04545)(20)} = 0.4029$$
3. $P(x > 20) = 0.4029$ but $P(\bar{x} > 20) = 0.7919$. (1)
 The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means. (2) When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. (3) Use the clt with the normal distribution when you are being asked to find the probability for a mean.
4. Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$
 $k = 26.0$ using $\text{invNorm} = 26.0$



The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time. Ninety five percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

Try It

Use the information in Example 2, but change the sample size to 144.

1. Find $P(20 < \bar{x} < 30)$.
2. Find $P(\Sigma x \text{ is at least } 3,000)$.
3. Find the 75th percentile for the sample mean excess time of 144 customers.
4. Find the 85th percentile for the sum of 144 excess times used by customers.

[practice-area rows="2"][/practice-area]

Show Answer

1. 0.8623
2. 0.7377
3. 23.2
4. 3,441.6

Example 3

In the United States, someone is sexually assaulted every two

minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.

1. Find the median, the first quartile, and the third quartile for the sample mean time of sexual assaults in the United States.
2. Find the median, the first quartile, and the third quartile for the sum of sample times of sexual assaults in the United States.
3. Find the probability that a sexual assault occurs on the average between 1.75 and 1.85 minutes.
4. Find the value that is two standard deviations above the sample mean.
5. Find the IQR for the sum of the sample times.

Solution

1. We have $\mu_x = \mu = 2$ and $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$.

Therefore:

$$50\text{th percentile} = \mu_x = \mu = 2 \quad \text{TI-83/84:}$$

$$\text{invNorm}(0.5, 2, 0.05),$$

$$25\text{th percentile} = 1.97 \quad \text{TI-83/84:}$$

$$\text{invNorm}(0.25, 2, 0.05),$$

$$75\text{th percentile} = 2.03 \quad \text{TI-83/84:}$$

$$\text{invNorm}(0.75, 2, 0.05).$$

2. We have $\mu_{\sum x} = n\mu_x = 100(2) = 200$ and $\sigma_{\sum x} = \sqrt{100}(0.5) = 5$.

Therefore:

$$50\text{th percentile} = \mu_{\sum x} = 200 \quad \text{TI-83/84: invNorm}(0.50, 200, 0.05),$$

$$25\text{th percentile} = 196.63 \quad \text{TI-83/84: invNorm}(0.25, 200, 0.05),$$

$$75\text{th percentile} = 203.37 \quad \text{TI-83/84: invNorm}(0.75, 200,$$

- 0.05)
3. $P(1.75 < \bar{x} < 1.85) = 0.0013$
TI-83/84: `normalcdf(1.75,1.85,2,0.05)`
 4. Using the z-score equation, $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$.
To solve for \bar{x} , $z = 2$, $\mu_{\bar{x}} = 2$, $\sigma_{\bar{x}} = 2$
Then we have $\bar{x} = 2(0.05) + 2 = 2.1$
 5. The IQR of $\sum x$ is 75th percentile - 25th percentile = 203.37
- 196.63 = 6.74

Try It

Try It

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follow a normal distribution.

1. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
2. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.

3. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Show Answer

1. $P(x > 120) = 0.0272$.

TI-Calculator: $\text{normalcdf}(120, 99, 114.8, 13.1)$

There is about a 3%, that the randomly selected woman will have systolic blood pressure greater than 120.

2. Since we are selecting 40 women randomly, $P(x > 120) = 0.006$.

TI-83/84: $\text{normalcdf}(120, 99, 114.8, \frac{13.1}{\sqrt{40}})$

There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.

3. The central limit theorem could not be used if the sample size were four and we did not know the original distribution was normal. The sample size would be too small.

Example 4

A study was done about violence against prostitutes and the

symptoms of the post-traumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

1. In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?
2. Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.
3. In a sample of 49 prostitutes, what is the probability that the sum of the ages is no less than 1,600?
4. Is it likely that the sum of the ages of the 49 prostitutes is at most 1,595? Interpret the results.
5. Find the 95th percentile for the sample mean age of 65 prostitutes. Interpret the results.
6. Find the 90th percentile for the sum of the ages of 65 prostitutes. Interpret the results.

Solution

$$1. P(\bar{x} < 35) = 0.9886$$

TI-83/84: `normalcdf(-E99,35,30.9,1.8)`

$$2. P(\bar{x} > 50) \approx 0.$$

TI-83/84: `normalcdf(50, E99,30.9,1.8)`

For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.

$$3. P(\Sigma x \geq 1,600) = 0.0864$$

TI-8/84: `normalcdf(1600,E99,1514.10,63)`

$$4. P(\Sigma x \leq 1,595) = 0.9005.$$

TI-83/84: `normalcdf(-E99,1595,1514.10,63)`

This means that there is a 90% chance that the sum of the ages for the sample group $n = 49$ is at most 1595.

$$5. \text{The 95th percentile} = 32.7.$$

TI-83/84: $\text{invNorm}(0.95, 30.9, 1.1)$

This indicates that 95% of the prostitutes in the sample of 65 are younger than 32.7 years, on average.

6. The 90th percentile = 2101.5.

TI-83/84: $\text{invNorm}(0.90, 2008.5, 72.56)$

This indicates that 90% of the prostitutes in the sample of 65 have a sum of ages less than 2,101.5 years.

Try It

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69.0 inches and a standard deviation of 2.8 inches.

1. What mean doorway height would allow 95% of men to enter the aircraft without bending?
2. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
3. For engineers designing the 757, which result is more relevant: the height from part 1 or part 2? Why?

Show Answer

1. We know that $\mu_x = \mu = 69$ and we have $\sigma_x = 2.8$.
The 95th percentile = the height of the doorway that would allow 95th of men to enter the aircraft without bending.

TI-83/84: invNorm(0.95,69,2.8)

The 95th percentile = 73.61

2. We know that $\mu_x = \mu = 69$ and $\sigma_x = \frac{2.8}{\sqrt{100}}$.

We are looking the 95th percentile for a random sample of 100 men.

TI-83/84: invNorm(0.95,69, $\frac{2.8}{\sqrt{100}}$)

The mean doorway height that satisfies the condition is 69.49 inches.

3. When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible.

Therefore, we need to use the result based on part 1.

Historical Note: Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one

of the most important applications of the central limit theorem. Binomial probabilities with a small value for n (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n , you had to use the binomial formula, which could be very complicated. Using the **normal approximation to the binomial** distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure
- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the **continuity correction factor** and is used in the following example.

Example 5

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.

2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **fewer than 147** favor a charter school.
5. Find the probability **that exactly 175** favor a charter school.

Solution

Let X = the number that favor a charter school for grades K through

5. $X \sim B(n, p)$ where $n = 300$, $p = 0.53$, and $q = 1 - p = 1 - 0.53 = 0.47$.

Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial.

The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$.

The mean = $300 * 0.53 = 159$ and the standard deviation = $\sqrt{300 * 0.53 * 0.47} = 8.6447$.

The random variable for the normal distribution is Y . $Y \sim N(159, 8.6447)$.

1. you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

TI-83/84: `normalcdf(149.5,10^99,159,8.6447)`.

2. you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

TI-83/84: `normalcdf(0,160.5,159,8.6447) = 0.5689`

3. you **exclude 155** so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$.

TI-83/84: `normalcdf(155.5,10^99,159,8.6447) = 0.6572`.

4. you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

TI-83/84: `normalcdf(0,146.5,159,8.6447) = 0.0741`

5. $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

TI-83/84: `normalcdf(174.5,175.5,159,8.6447) = 0.0083`

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators, and they easily calculate probabilities for the binomial distribution. If you type in "binomial probability distribution calculation" in an Internet browser, you can find at least one online calculator for the binomial.

For Example 3, the probabilities are calculated using the following binomial distribution: ($n = 300$ and $p = 0.53$).

Compare the binomial and normal distribution answers.

1. $P(X \geq 150) : 1 - \text{binomialcdf}(300, 0.53, 149) = 0.8641$
2. $P(X \leq 160) : \text{binomialcdf}(300, 0.53, 160) = 0.5684$
3. $P(X > 155) : 1 - \text{binomialcdf}(300, 0.53, 155) = 0.6576$
4. $P(X < 147) : \text{binomialcdf}(300, 0.53, 146) = 0.0742$
5. $P(X = 175) : \text{binomialpdf}(300, 0.53, 175) = 0.0083$ (You need to use the binomial pdf.)

Try It

In a city, 46 percent of the population favor the incumbent, Dawn Morgan, for mayor. A simple random

sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

[practice-area rows="1"][/practice-area]

Show Answer

$n = 500, p = 0.46, q = 1 - p = 1 - 0.46 = 0.54$

0.0401

TI-Calculator: normalcdf (250, 1E99, 500*0.46,

$\sqrt{500 * 0.46 * 0.54}$)

References

Data from the Wall Street Journal.

"National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed May 17, 2013).

Concept Review

The central limit theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample

size you take from a population, the closer the sample mean gets to μ .

PART VII

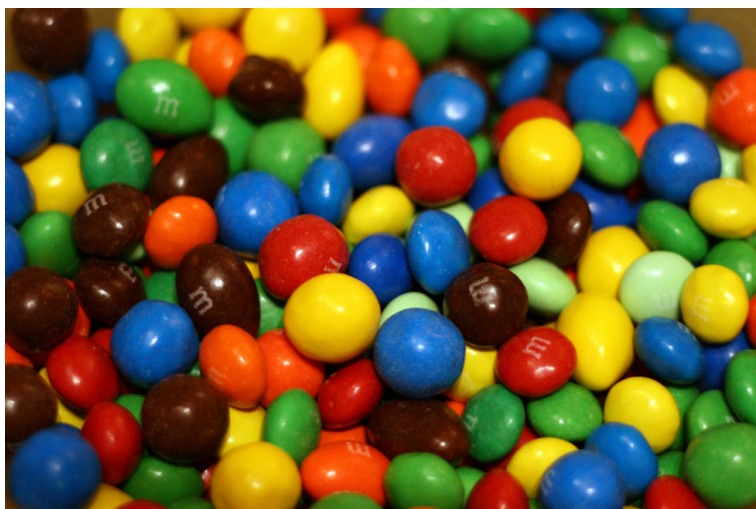
CONFIDENCE INTERVALS

Create confidence intervals and perform hypothesis tests and use them in statistical inference

Learning Outcomes

- Create confidence intervals for means and proportions for one and two populations
- Perform hypothesis tests for means and proportions for one and two populations
- Perform Chi-Square distribution tests, including Goodness of Fit test
- Perform Chi-Square distribution tests, including Test for Independence
- Perform Chi-Square distribution tests, including Test for Homogeneity
- Perform a One-Way Anova test to test for the equivalence of three or more population means

36. Introduction: Confidence Intervals



Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Learning Objectives

By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the Student's t distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean.

Another example is that you are trying to determine the percentage of times you make a basket when shooting a basketball. You might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion of successful shot.

We use sample data to make generalizations about an unknown population. This part of statistics is called inferential statistics. **The sample data help us to make an estimate for a population.** We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point

estimates, we construct interval estimates, called **confidence intervals**.

Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable.

It is the population parameter that is fixed.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=59#oembed-1>

For example,

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes.

You could conduct a survey and calculate the sample mean, \bar{x} , and the sample standard deviation, s .

You would use \bar{x} to estimate the population mean and s to estimate the population standard deviation.

The sample mean, \bar{x} , is the point estimate for the population mean, μ .

The sample standard deviation, s , is the point estimate for the population standard deviation, σ .

Each of \bar{x} and s is called a statistic.

A confidence interval is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the iTunes example, we do not know the population mean μ , but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$

The [empirical rule](#), which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean (\bar{x}) will be within two standard deviations of the population mean μ .

For our iTunes example, mean $\mu = 0.1$ and standard deviation $\sigma = 0.1$. Therefore, two standard deviations is $(2)(0.1) = 0.2$.

The sample mean $\bar{x} = 0.1$ is likely to be within 0.2 units of μ .

Because \bar{x} is within 0.2 units of μ , which is unknown, then μ is likely to be within 0.2 units of \bar{x} in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $(2)(0.1)$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between $\bar{x} - 0.2 = 2 - 0.2 = 1.8$ and $\bar{x} + 0.2 = 2 + 0.2 = 2.2$.

The 95% confidence interval is (1.8, 2.2). We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2.

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an \bar{x} that is not within 0.2 units of the true mean μ . The second possibility happens for only 5% of all the samples (95–100%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ . Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase “margin of error.” Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Note

Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

37. 8.1 A Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**, we need \bar{x} is the **point estimate** of the unknown population mean μ .

The confidence interval estimate will have the form:

(point estimate – error bound, point estimate + error bound) or,
in symbols, $(\bar{x} - EBM, \bar{x} + EBM)$

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

90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level, CL. α is the probability that the interval does not contain the unknown population parameter.

Given that CL is the probability that the calculated confidence interval estimate will contain the true population parameter,
 α is the probability that the interval does not contain the unknown population parameter,

mathematically, we can conclude that $\alpha + CL = 1$.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=60#oembed-1>

Example 1

Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.

The sample mean is seven, and the error bound for the mean is 2.5.

$\bar{x} = 7$. At 95% confidence level, $EBM = 2.5$.

General form to find confidence interval:

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

The confidence interval is $(7 - 2.5, 7 + 2.5)$, and calculating the values gives $(4.5, 9.5)$.

Since we calculate the interval at 95% confidence level, we estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5.”

Try It

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

What is the confidence interval estimate for the population mean?

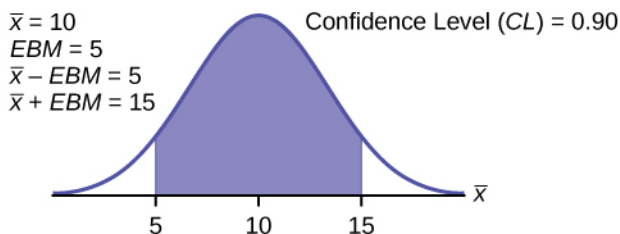
[practice-area rows="1"][/practice-area]

Show Answer

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where $EBM = 5$.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10\%$ in both tails, or 5% in each tail, of the normal distribution.



To capture the central 90%, we must go out 1.645 “standard deviations” on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the “standard deviation” used must be appropriate for the parameter we are estimating, so in this section

we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the “standard error of the mean” in order to distinguish clearly the standard deviation for a mean from the population standard deviation σ .

In summary, as a result of the central limit theorem:

- \bar{X} , that is, $\bar{X} \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$
 - **When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**
-

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean \bar{X} from the sample data. Remember, in this section we already know the population standard deviation σ .
 - Find the z-score that corresponds to the confidence level.
 - Calculate the error bound EBM.
 - Construct the confidence interval.
 - Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)
-

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ , we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL , is the area in the middle of the standard normal distribution.

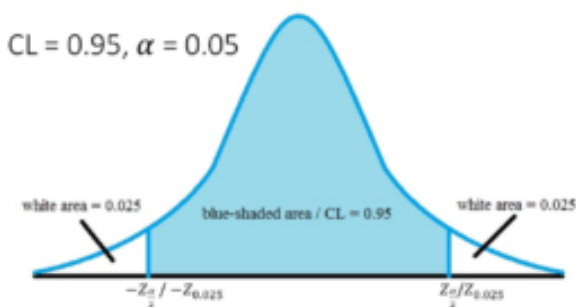
$CL = 1 - \alpha$, so α is the area that is split equally between the two tails.

Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $Z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$,
then $\frac{\alpha}{2} = 0.025$. We can write $Z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.



$Z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a standard normal probability table.

To find out $z_{0.025}$, we can use TI-Calculator.

TI-Calculator: `invNorm(0.975, 0, 1)`

We will find $z_{0.025} = 1.96$

Note

Remember to use the area to the LEFT of ; in this chapter the last two inputs in the `invNorm` command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound (EBM)

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

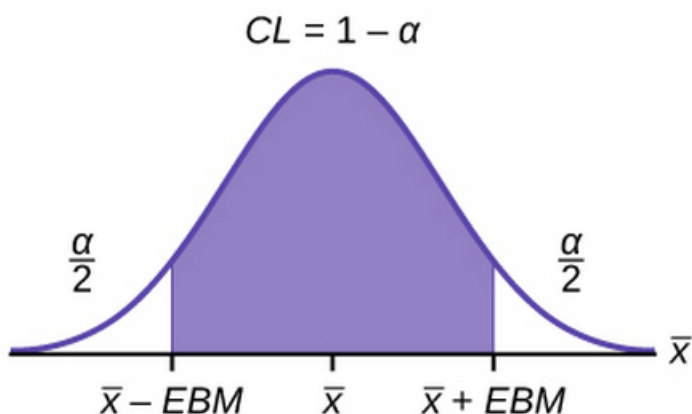
$$\bullet \text{ EBM} = (Z_{\frac{\alpha}{2}}) \left(\frac{\sigma}{\sqrt{n}} \right)$$

The confidence interval estimate has the format $(\bar{x} - \text{EBM}, \bar{x} + \text{EBM})$.

The graph gives a picture of the entire situation.

Blue-shaded area + 2 white areas on each side

$$\begin{aligned} &= \text{CL} + \frac{\alpha}{2} + \frac{\alpha}{2} \\ &= \text{CL} + \alpha \\ &= 1 \end{aligned}$$



Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). “We estimate with ____% confidence that the true population mean (include the context of the problem) is between ____ and ____ (include appropriate units).”

Example 2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

You can use technology to calculate the confidence interval directly. The first solution is shown step-by-step (Solution A). The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A:

To find the confidence interval, you need the sample mean, and the EBM.

$$\bar{x} = 68, \sigma = 3, n = 6$$
$$EBM = (Z_{\frac{\alpha}{2}}) \left(\frac{\sigma}{\sqrt{n}} \right)$$

The confidence level is 90% ($CL = 0.90$)

$CL = 0.90$ so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = 0.05, z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is $1 - 0.05 = 0.95$.

$$\frac{z_{\alpha}}{2} = z_{0.05} = 1.645$$

TI-83/84: invNorm(0.95, 0, 1)

This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$$EBM = (1.645) \left(\frac{3}{\sqrt{36}} \right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is (67.1775, 68.8225).

Solution B:

Press **STAT** and arrow over to **TESTS**.

Arrow down to **7: ZInterval**.

Press **ENTER**.

Arrow to **Stats** and press **ENTER**.

Arrow down and enter three for σ , 68 for \overline{X} , 36 for n , and .90 for C-level.

Arrow down to **Calculate** and press **ENTER**.

The confidence interval is (67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Try It

Suppose average pizza delivery times are normally

distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time.

Show Answer

(34.1347, 37.8653)

Example 3

The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. This table shows the highest SAR level for a random selection of cell phone models as measured by the FCC.

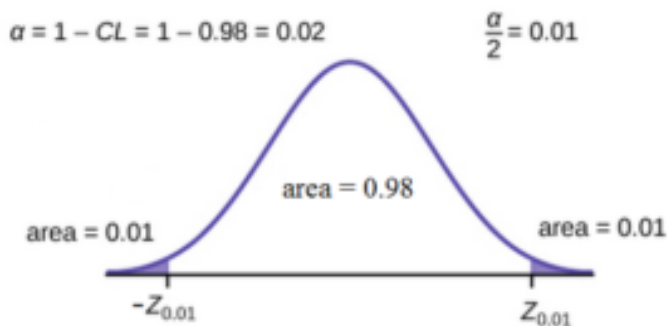
Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones.
Assume that the population standard deviation is $\sigma = 0.337$.

Solution A:

To find the confidence interval, start by finding the point estimate: the sample mean, $\bar{x} = 1.024$

Next, find the EBM. Because you are creating a 98% confidence interval, $CL = 0.98$.



You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99.

Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (Z_{0.01})\left(\frac{\sigma}{\sqrt{n}}\right) = (2.236)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $\bar{x} \pm EBM$

$$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$$

$$\bar{x} + EBM = 1.024 + 0.1431 = 1.1671$$

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution B:

- Press STAT and arrow over to TESTS.
- Arrow down to 7: ZInterval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:

- σ : 0.337
- \bar{x} : 1.024
- n : 30
- C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The 98% confidence interval is (0.881, 1.167).(to three decimal places)

Try It

This table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SA R	Phone Model	SA R
Blackberry Pearl 8120	1.4 8	Nokia E71x	1.5 3
HTC Evo Design 4G	0.8	Nokia N75	0.6 8
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.3 6	Sagem Puma	1.2 4
LG Fathom	0.7 7	Samsung Fascinate	0.5 7
LG Optimus Vu	0.4 62	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.3 6	Samsung Nexus S	0.5 1
Motorola Droid Pro	1.3 9	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.7 3
Nokia 7705 Twist	0.7	ZTE C79	0.8 69

[practice-area rows="2"][/practice-area]

Show Answer

$$\begin{aligned}\bar{x} &= 0.940 \\ \frac{\alpha}{2} &= \frac{1 - CL}{2} = \frac{1 - 0.93}{2} = 0.035 \\ z_{0.05} &= 1.812 \\ EBM &= (z_{0.05}) \left(\frac{\sigma}{\sqrt{n}} \right) = (1.812) \left(\frac{0.337}{\sqrt{20}} \right) = \\ 0.1365 \quad \bar{x} - EBM &= 0.940 - 0.1365 = 0.8035 \\ \bar{x} + EBM &= 0.940 + 0.1365 = 1.0765\end{aligned}$$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in Example 3 and the Try It just completed. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 4

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval

estimate for the population mean exam score (the mean score on all exams).

Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution:

To find the confidence interval, you need the sample mean, \bar{x} , and the EBM.

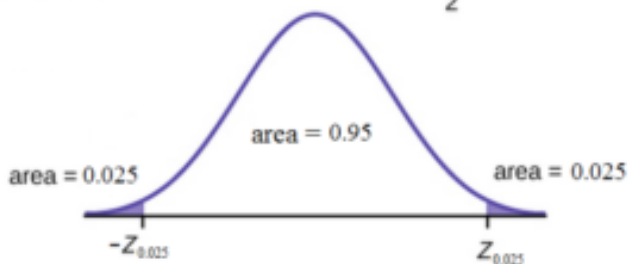
$$\bar{x} = 68, \sigma = 3, n = 36$$

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

$$CL = 0.95 \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025$$



The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

TI-83/84: invnorm(0.975,0,1)

This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

Interpretation

We estimate with 95% confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

95% of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

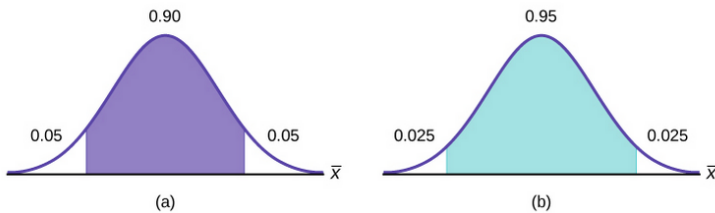
Comparing the Results

In Example 2, the 90% confidence interval is (67.18, 68.82).

In Example 4, the 95% confidence interval is (67.02, 68.98).

The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider.

To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Try It

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. The sample mean deliver time of 20 deliverers is 36 minutes. Find a 95% confidence interval estimate for the true mean pizza delivery time.

[practice-area rows="1"][/practice-area]

Show Answer

(33.37, 38.63)

Example 5

What happens to the error bound(EBM) if the sample size is changed?

In example 2, we suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68.

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Leave everything the same except the sample size. Use the original 90% confidence level.

What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$?

What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\mu = 68$
- $\sigma = 3$
- $n = 100$
- The confidence level is 90% ($CL=0.90$); .

Solution:

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

If we **increase** the sample size n to 100, we **decrease** the error bound.

If we **decrease** the sample size n to 25, we **increase** the error bound.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Try It

The mean pizza delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is 50 restaurants. Find a 90% confidence interval estimate for the population mean delivery time.

[practice-area rows="1"][/practice-area]

Show Answer

(34.6041, 37.3958)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
 - OR, average the upper and lower endpoints of the confidence interval.
-

Example 6

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Suppose we know that a confidence interval is (67.18, 68.82) and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $EBM = 68.82 - 68 = 0.82$.
- If we don't know the sample mean: .

Calculate the Sample Mean:

- If we know the error bound: $= 68.82 - 0.82 = 68$
- If we don't know the error bound: .

Try It

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

[practice-area rows="1"][/practice-area]

Show Answer

Sample mean is 45, error bound is 2.88

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is

The formula for sample size is , found by solving the error bound formula for n .

In this formula, z is , corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.
- using the sample size equation.

- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Try It

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

[practice-area rows="1"][/practice-area]

Show Answer

35 students

References

“American Fact Finder.” U.S. Census Bureau. Available online at

<http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t> (accessed July 2, 2013).

“Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).

“Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall.” Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30, 2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. “2000 CDC Growth Charts for the United States: Methods and Development.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/growthcharts/2000growthchart-us.pdf> (accessed July 2, 2013).

La, Lynn, Kent German. “Cell Phone Radiation Levels.” c|net part of CBX Interactive Inc. Available online at <http://reviews.cnet.com/cell-phone-radiation-levels/> (accessed July 2, 2013).

“Mean Income in the Past 12 Months (in 2011 Inflation-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates.” American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

“Metadata Description of Candidate Summary File.” U.S. Federal Election Commission. Available online at <http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml> (accessed July 2, 2013).

“National Health and Nutrition Examination Survey.” Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/nchs/nhanes.htm> (accessed July 2, 2013).

Concept Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (EBM). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate - EBM, point estimate + EBM)

The calculation of EBM depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding EBM increases as well. As the sample size increases, the EBM decreases. By the central limit theorem,

Given a confidence interval, you can work backwards to find the error bound (EBM) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the EBM formula for n to discover the size of the sample that is needed to achieve this goal:

Formula Review

$\bar{X} \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$. The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

(lower bound, upper bound) = (point estimate - EBM, point estimate + EBM)

$$= (\bar{x} - \text{EBM}, \bar{x} + \text{EBM})$$

$$= \left(\bar{x} - z \frac{\alpha}{\sqrt{n}}, \bar{x} + z \frac{\alpha}{\sqrt{n}}\right)$$

$\text{EBM} = z \frac{\alpha}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - \text{CL}$ = the proportion of confidence intervals that will not contain the population

$z \frac{\alpha}{\sqrt{n}}$ = the z-score with the property that the area to the right of the z-score is $\alpha/2$ this is the z-score used in the calculation of "EBM where $\alpha = 1 - \text{CL}$.

$$n = \frac{z^2 \sigma^2}{\text{EBM}^2}$$

(n) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate - error bound, point estimate + error bound)

To find the error bound when you know the confidence interval

$$\text{error bound} = \text{upper value} - \text{point estimate} \text{ OR } \text{error bound} = \frac{\text{upper value} - \text{lower value}}{2}$$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format $EBM = (\bar{x} - EBM, \bar{x} + EBM)$

38. 8.2 A Single Population Mean using the Student t Distribution

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

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

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

If you draw a simple random sample of size n from a population that has an approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t -score:

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is from its mean μ . For each sample size n , there is a

different Student’s t-distribution.

The **degrees of freedom, $n - 1$** , come from the calculation of the sample standard deviation s . Because the sum of the deviations is zero, we can find the last deviation once we know the other $n - 1$ deviations. The other $n - 1$ deviations can change or vary freely. **We call the number $n - 1$ the degrees of freedom (df).**

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the

probability for given values of t . The grammar for the `tcdf` command is **tcdf(lower bound, upper bound, degrees of freedom)**.

However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the `invT` command on the DISTRibution menu.

The `invT` command requires two inputs: **invT(area to the left, degrees of freedom)**

The output is the t -score that corresponds to the area we specified.

The TI-83 and 83+ do not have the `invT` command. (The TI-89 has an inverse T command.)

A probability table for the Student's t -distribution can also be used. The table gives t -scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an `invT` program or command, so if you are using that calculator, you need to use a probability table for the Student's t -Distribution.) When using a t -table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A [Student's \$t\$ table](#) gives t -scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t -probabilities.**

The notation for the Student's t -distribution (using T as the random variable) is:

$$T \sim t_{df} \text{ where } df = n - 1.$$

For example, if we have a sample of size $n = 20$ items, then we calculate the degrees of freedom as $df = n - 1 = 20 - 1 = 19$ and we write the distribution as $T \sim t_{19}$.

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

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

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

The format for the confidence interval is: $(\bar{x} - \text{EBM}, \bar{x} + \text{EBM})$

In other words, the formula for the confidence interval is

$$(\bar{x} - (t_{\frac{\alpha}{2}})(\frac{\sigma}{\sqrt{n}}), \bar{x} + (t_{\frac{\alpha}{2}})(\frac{\sigma}{\sqrt{n}}))$$

Calculate the Confidence Interval bu using
TI-Calculator:

Press STAT.

Arrow over to TESTS. Arrow down to 8:Interval and press ENTER (or just press 8).



One or more interactive elements has been excluded
from this version of the text. You can view them online

here: [https://library.achievingthedream.org/
odessastatistics/?p=61#oembed-1](https://library.achievingthedream.org/odessastatistics/?p=61#oembed-1)

Example 1

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

The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

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

-
- The first solution is step-by-step (Solution A).
 - The second solution uses the TI-83+ and TI-84 calculators (Solution B).
-

Solution A:

To find the 95% confidence interval, you need the sample mean, \bar{x} , and the EBM.

$$\bar{X} = 8.2267, s = 1.6722, n = 15, df = 15 - 1 = 14$$
$$CL = 0.95.$$

$$\text{Then } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$t_{\frac{\alpha}{2}} = t_{0.025}$$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is $1 - 0.025 = 0.975$

TI-84+ Calculator : invT(.975,14) or use t-table to find $t_{\frac{\alpha}{2}}$ when df = 14.

$$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$$

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

$$\text{Hence, EBM} = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = (0.924)$$

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8:TInterval** and press **ENTER** (or you can just press **8**).

Arrow to **Data** and press **ENTER**.

Arrow down to **List** and enter the list name where you put the data.

There should be a 1 after **Freq**.

Arrow down to **C-level** and enter 0.95

Arrow down to **Calculate** and press **ENTER**.

The 95% confidence interval is (7.3006, 9.1527)

Note: When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

Try It

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

[practice-area rows="1"][/practice-area]

Show Answer

(8.1634, 9.8032)

Example 2

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States.

The cord blood of the “In utero/newborn” group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. This table shows how many of the targeted chemicals were found in each infant’s cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant’s blood.

Solution A:

From the sample, you can calculate \overline{x} =127.45
 and $s = 25.965$. There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90% confidence interval: $CL = 0.90$,
 so $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = 0.05$$

$$(t_{\frac{\alpha}{2}} = t_{0.05}$$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = (1.729) \left(\frac{25.965}{\sqrt{20}} \right) = 10.038$$

$$\overline{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\overline{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90% confidence that the mean number of all

targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution B:

Enter the data as a list.

Press `STAT` and arrow over to `TESTS`.

Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).

Arrow to `Data` and press `ENTER`.

Arrow down to `List` and enter the list name where you put the data.

Arrow down to `Freq` and enter 1.

Arrow down to `C-level` and enter 0.90

Arrow down to `Calculate` and press `ENTER`.

The 90% confidence interval is (117.41, 137.49).

Example 3

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in This table. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution A:

$\bar{x} = 6.133$, $s = 5.514$, $n = 15$, and $df = 15 - 1 = 14$

$CL = 0.98$, so $\alpha = 1 - CL = 1.0 - 0.98 = 0.02$

$(t_{\frac{\alpha}{2}} = t_{0.01})$

$(t_{\frac{\alpha}{2}} = t_{0.01} = 2.624)$

EBM

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

$= 2.624 \left(\frac{5.514}{\sqrt{15}} \right) = 3.736$

$\bar{x} - EBM = 6.133 - 3.736 = 2.397$

$\bar{x} + EBM = 6.133 + 3.736 = 9.869$

We estimate with 98% confidence that the mean number of all hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution B:

Enter the data as a list.

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8: TInterval**.

Press **ENTER**.

Arrow to **Data** and press **ENTER**.

Arrow down and enter the name of the list where the data is stored.

Enter **Freq: 1** Enter **C-Level: 0.98**

Arrow down to **Calculate** and press **Enter**.

The 98% confidence interval is (2.3965, 9,8702).

References

“America’s Best Small Companies.” Forbes, 2013. Available online at <http://www.forbes.com/best-small-companies/list/> (accessed July 2, 2013).

Data from Microsoft Bookshelf.

Data from <http://www.businessweek.com/>.

Data from <http://www.forbes.com/>.

“Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012.” Federal Election Commission. Available online at <http://www.fec.gov/data/index.jsp> (accessed July 2, 2013).

“Human Toxome Project: Mapping the Pollution in People.” Environmental Working Group. Available online at <http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn> (accessed July 2, 2013).

“Metadata Description of Leadership PAC List.” Federal Election Commission. Available online at <http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml> (accessed July 2, 2013).

Concept Review

In many cases, the researcher does not know the population standard deviation, σ , of the measure being studied. In these cases,

it is common to use the sample standard deviation, s , as an estimate of σ . The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t -distribution is much better. Define a t -score using the following formula:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The t -score follows the Student's t -distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$ where $t_{\frac{\alpha}{2}}$ is the t -score with area to the right equal to $\frac{\alpha}{2}$. s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $\frac{\alpha}{2}$ for a given α .

Formula Review

s = the standard deviation of sample values.

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

t is the formula for the t -score which measures how far away a measure is from the population mean in the Student's t -distribution
 $df = n - 1$; the degrees of freedom for a Student's t -distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T , has a Student's t -distribution with df degrees of freedom

t is the formula for the t -score which measures how far away a measure is from the population mean in the Student's t -distribution
 $df = n - 1$; the degrees of freedom for a Student's t -distribution where n represents the size of the sample

$T \sim t_{df}$ the random variable, T , has a Student's t -distribution with df degrees of freedom

$t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$ = the error bound for the population mean when the population standard deviation is unknown

is the t -score in the Student's t -distribution with area to the right equal to

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

$$= (\text{point estimate} - \text{EBM}, \text{point estimate} + \text{EBM}) = \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}} \right)$$

39. 8.3 Confidence Interval for Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: $(0.40 - 0.03, 0.40 + 0.03)$.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=62#oembed-1>

How do you know you are dealing with a proportion problem?

First, the underlying **distribution is a binomial distribution**. (There is no mention of a mean or average.) If X is a binomial random

variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a proportion, take X , the random variable for the number of successes and divide it by n , the number of trials (or the sample size).

The random variable P' (read “P prime”) is that proportion, $P' = \frac{X}{n}$. (Sometimes the random variable is denoted as \hat{p} , read “P hat”.)

When n is large and p is not close to zero or one, we can use the **normal distribution** to approximate the binomial.

$$X \sim N(np, \sqrt{npq})$$

If we divide the random variable, the mean, and the standard deviation by

n , we get a normal distribution of proportions with P' , called the estimated proportion, as the random variable. (Recall that a proportion as the number of successes divided by n .)

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$

Using algebra to simplify:

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

P' follows a normal distribution for proportions:

$$\frac{X}{n} = P' \sim N\left(\frac{np}{n}, \sqrt{\frac{pq}{n}}\right)$$

The confidence interval has the form $(p' - \text{EBP}, p' + \text{EBP})$ given that EBP is error bound for the proportion.

$$\bullet \quad p' = \frac{x}{n}$$

p' = the **estimated proportion** of successes (p' is a **point estimate** for p , the true proportion.)
 x = the **number** of successes
 n = the size of the sample

$$\bullet \quad \text{The error bound for a proportion is } \text{EBP} = \left(z_{\frac{\alpha}{2}}\right) \left(\sqrt{\frac{p'q'}{n}}\right)$$

where $q' = 1 - p'$.

p' = the **estimated proportion** of successes
 q' = the **estimated proportion** of failures

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five. This formula is similar to the error bound formula for a mean, except that the “appropriate standard deviation” is different.

For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$.

For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p'q'}{n}}$ as the

standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the sample proportions p' and q' are estimates of the unknown population proportions p and q . The estimated proportions p' and q' are used because p and q are not

known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

Note

For the normal distribution of proportions, the z-score formula is as

follows. If $P' \sim N(p, \sqrt{\frac{pq}{n}})$ then the z-score formula is $z =$

$$\frac{p' - p}{\sqrt{pqn}}$$

Example 1

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones.

Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution A (Step-by-Step Solution)

Let X = the number of people in the sample who have cell phones. To calculate the confidence interval, you must find p' , q' , and EBP . $n = 500$, x

= the number of successes = 421 $p' = \frac{x}{n} = \frac{421}{500} = 0.842$ $p' = 0.842$ is the

sample proportion; this is the point estimate of the population proportion. $q' = 1 - p' = 1 - 0.842 = 0.158$

Since $CL = 0.95$, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$.

$$\frac{\alpha}{2} = 0.025$$

$$\text{Then } Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

Use the TI-83, 83+, or 84+ calculator command $\text{invNorm}(0.975, 0, 1)$ to find $Z_{0.025}$.

Remember that the area to the right of $Z_{0.025}$ is 0.025 and the area to the left of $Z_{0.025}$ is 0.975.

This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(Z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.96) \sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The 95% confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.81, 0.874)$.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to and enter 421.
- Arrow down to and enter 500.
- Arrow down to C-Level and enter .95.
- Arrow down to Calculate and press ENTER.

The 95% confidence interval is (0.81003, 0.87397).

Try It

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

[practice-area rows="1"][/practice-area]

Show Answer

(0.3315, 0.4525)

Example 2

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution A (Step-by-Step Solution)

$$x = 300 \text{ and } n = 500 \Rightarrow p' = \frac{x}{n} = \frac{300}{500} = 0.600 \text{ Since } CL = 0.90, \text{ then } \alpha = 1 - CL = 1 - 0.90 = 0.10$$
$$\frac{\alpha}{2} = 0.05 \Rightarrow Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command $\text{invNorm}(0.95, 0, 1)$ to find $Z_{0.05}$.

Remember that the area to the right of $Z_{0.05}$ is 0.05 and the area to the left of $Z_{0.05}$ is 0.95.

This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(Z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.645) \sqrt{\frac{(0.6)(0.4)}{500}} = 0.036$$

$$p' - EBP = 0.600 - 0.036 = 0.564$$

$$p' + EBP = 0.600 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.564, 0.636)$.

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to and enter 300.
- Arrow down to and enter 500.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.564, 0.636).

Try It

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

[practice-area rows="2"][/practice-area]

Show Answer

(0.7731, 0.8269)

We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.31% and 82.69%.

Example 3

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms.

She surveys 600 students and finds that 480 are against the new legislation.

In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Solution A (Step-by-Step Solution)

Sixty-eight percent (68%) of students own an iPod and a smart phone. $p' = 0.68$, $q' = 1 - p' = 1 - 0.68 = 0.32$. Since $CL = 0.97$, we know $\alpha = 1 - 0.97 = 0.03$. The area to the left of $Z_{0.015}$ is 0.015, and the area to the right of $Z_{0.015}$ is $1 - 0.015 = 0.985$.

Using the TI 83, 83+, or 84+ calculator function $\text{InvNorm}(0.985, 0, 1)$ / Standard Normal Table, the z-score $Z_{0.015} = 2.17$

$$EBP = \left(z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.645) \sqrt{\frac{(0.68)(0.32)}{300}} = 0.0269$$

$$p' - EBP = 0.68 - 0.0269 = 0.6531$$

$$q' - EBP = 0.68 + 0.0269 = 0.7069$$

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.6531 and 0.7069.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to x and enter $300 \cdot 0.68$.
- Arrow down to n and enter 300.
- Arrow down to C-Level and enter 0.97.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.6531, 0.7069).

“Plus Four” Confidence Interval for p

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to

use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example 4

A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Six students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution A (Step-by-Step Solution)

Six students out of 25 reported smoking within the past week, so $x = 6$ and $n = 25$.

Because we are using the “plus-four” method, we will use $x = 6 + 2 = 8$ and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} = 0.276$$

$$q' = 1 - p' = 1 - 0.276 = 0.724$$

Since $CL = 0.95$, we know $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

Use the TI-83, 83+, or 84+ calculator command $\text{invNorm}(0.975, 0, 1)$ / Standard Normal probability table to find $Z_{0.025}$.

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

Remember that the area to the right of $Z_{0.025}$ is 0.025 and the area to the left of $Z_{0.025}$ is 0.975.

$$EBP = \left(Z_{\frac{\alpha}{2}} \right) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.96) \sqrt{\frac{(0.276)(0.724)}{29}} = 0.1627$$

$$p' - EBP = 0.276 - 0.1627 = 0.1133$$

$$p' + EBP = 0.276 + 0.1627 = 0.4387$$

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.1133 and 0.4387.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.

Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

- Arrow down to x and enter eight. Arrow down to n and enter 29.
- Arrow down to C-Level and enter 0.95.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.113, 0.439).

Example 5

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus-four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution A (Step-by-Step Solution)

Using “plus four,” we have $x = 31 + 2 = 33$ and $n = 65 + 4 = 69$. $p' = \frac{x}{n} = \frac{33}{69}$

$$q' = 1 - \frac{x}{n} = 1 - \frac{33}{69} = \frac{36}{69}$$

Since $CL = 0.96$, we know $\alpha = 1 - CL = 1 - 0.96 = 0.04$

$$\frac{\alpha}{2} = \frac{0.04}{2} = 0.02$$

Remember that the area to the right of $Z_{0.02}$ is 0.02 and the area to the left of $Z_{0.02}$ is 0.98.

Use the TI-83, 83+, or 84+ calculator command `invNorm(0.98,0,1)` / Standard Normal probability table to find $Z_{0.02}$.

$$Z_{\frac{\alpha}{2}} = Z_{0.02} = 2.0537$$
$$EBP = (Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) = (2.0537) \sqrt{\frac{(\frac{33}{69})(\frac{36}{69})}{69}} = 0.1235$$
$$p' - EBP = \frac{33}{69} - 0.1235 = 0.3548$$
$$p' + EBP = \frac{33}{69} + 0.1235 = 0.6018$$

We are 96% confident that between 35.48% and 60.18% of all freshmen at State U have declared a major.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to x and enter 33.
- Arrow down to n and enter 69.
- Arrow down to C-Level and enter 0.96.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.3548, 0.6018).

Example 6

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the “plus four” method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution A (Step-by-Step Solution)

Using “plus-four,” we have $x = 13 + 2 = 15$ and $n = 50 + 4 = 54$. $p' = \frac{x}{n} = \frac{15}{54}$

$$q' = 1 - \frac{x}{n} = 1 - \frac{15}{54} = \frac{39}{54}$$

Since $CL = 0.90$, we know $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

Remember that the area to the right of $Z_{0.05}$ is 0.05 and the area to the left of $Z_{0.05}$ is 0.95.

Use the TI-83, 83+, or 84+ calculator command `invNorm(0.95,0,1)` / Standard Normal probability table to find $Z_{0.05}$. $Z_{\frac{\alpha}{2}} =$

$$Z_{0.05} = 1.6449 \text{ EBP} = (Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.6449)^*$$

$$\sqrt{\frac{(\frac{15}{54})(\frac{39}{54})}{54}} = 0.1003 \quad p' - \text{EBP} = \frac{15}{54} - 0.1003 = 0.1775 \quad p' + \text{EBP} = \frac{15}{54} + 0.1003 = 0.3781$$

We are 90% confident that between 17.75% and 37.81% of all teens would report having more than 500 friends on Facebook.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to x and enter 15.
- Arrow down to n and enter 54.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.1775, 0.3781).

Example 7

The Berkman Center Study referenced in Example 6 talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the “plus-four” method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in Example 6.

Solution A (Step-by-Step Solution)

Using “plus-four,” we have $x = 159 + 2 = 161$ and $n = 588 + 4 = 592$.

$$p' = \frac{x}{n} = \frac{161}{592}$$

$$q' = 1 - \frac{x}{n} = 1 - \frac{161}{592} = \frac{431}{592}$$

Since $CL = 0.90$, we know $\alpha = 1 - CL = 1 - 0.90 = 0.10$

$$\frac{\alpha}{2} = \frac{0.1}{2} = 0.05$$

Remember that the area to the right of $Z_{0.05}$ is 0.05 and the area to the left of $Z_{0.05}$ is 0.95.

Use the TI-83, 83+, or 84+ calculator command `invNorm(0.95,0,1)` / [Standard Normal probability table](#) to find $Z_{0.05}$.

$$Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.6449 \text{ EBP} = (Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.6449)^*$$

$$\sqrt{\frac{\left(\frac{161}{592}\right)\left(\frac{431}{592}\right)}{592}} = 0.0301$$

$$p' - \text{EBP} = \frac{161}{592} - 0.0301 = 0.2418$$

$$p' + \text{EBP} = \frac{161}{592} + 0.0301 = 0.3021$$

We are 90% confident that between 24.18% and 30.21% of all teens would report having more than 500 friends on Facebook.

Solution B (Using TI-Calculator)

- Press STAT and arrow over to TESTS.
- Arrow down to A:1-PropZint.
- Press ENTER.
- Arrow down to x and enter 161.
- Arrow down to n and enter 592.
- Arrow down to C-Level and enter 0.90.
- Arrow down to Calculate and press ENTER.

The confidence interval is (0.242, 0.302).

Conclusion

The confidence interval for the larger sample is narrower than the interval from Example 6. Larger samples will always yield more precise confidence intervals than smaller samples. The “plus four” method has a greater impact on the smaller sample.

It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the EPB, changing it from 0.102 to 0.100.

In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592).

It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is $EBP =$

$$(Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right)$$

Solving for n gives you an equation for the sample size.

$$n = \frac{(Z_{\frac{\alpha}{2}})^2 (p'q')}{EBP^2}$$

Example 8

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones.

How many customers aged 50+ should the company survey in order

to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Solution:

From the problem, we know that $EBP = 3\% = 0.03$ and the confidence level is 90%.

However, in order to find n , we need to know the estimated (sample) proportion p' .

Remember that $q' = 1 - p'$. But, we do not know p' yet.

Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product.

(Try other products: $(0.6)(0.4) = 0.24$; $(0.3)(0.7) = 0.21$; $(0.2)(0.8) = 0.16$ and so on).

The largest possible product gives us the largest n .

This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion.

To calculate the sample size n , use the formula

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2} \text{ and make the substitutions.}$$

- $Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$
- $p'q' = (0.5)(0.5) = 0.25$
- $EBP = 3\% = 0.03$

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$$

$n = \left[\frac{1.645^2 * 0.25}{0.03^2}\right] = 751.6736 = 752$ (Round the answer to the next whole number.)

The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Try It

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Show Answer

271 customers should be surveyed.

References

Jensen, Tom. "Democrats, Republicans Divided on Opinion of Music Icons." Public Policy Polling. Available online at

<http://www.publicpolicypolling.com/Day2MusicPoll.pdf> (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. "Teens, Social Media, and Privacy." PewInternet, 2013. Available online at <http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx> (accessed July 2, 2013).

Prince Survey Research Associates International. "2013 Teen and Privacy Management Survey." Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~media/Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. "Three in Four U.S. Workers Plan to Work Past Retirement Age: Slightly more say they will do this by choice rather than necessity." Gallup® Economy, 2013. Available online at <http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx> (accessed July 2, 2013).

The Field Poll. Available online at <http://field.com/fieldpollonline/subscribers/> (accessed July 2, 2013).

Zogby. "New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure 'Investment' for National Security." Zogby Analytics, 2013. Available online at <http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll> (accessed July 2, 2013).

"52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

Concept Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n , where x represents the number of successes and n represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

(lower bound, upper bound)

The “plus four” method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate p' and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

$p' = x / n$ where x represents the number of successes and n represents the sample size.

The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p'$$

The variable p' has a binomial distribution that can be approximated

with the normal distribution shown here.

$$EBP = (z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right)$$

Confidence interval for a proportion:

$$(\text{lower bound, upper bound}) = (p' - EBP, p' + EBP) = (p' -$$

$$(Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right), p' + (Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right))$$

$$n = \frac{(Z_{\frac{\alpha}{2}})^2 p'q'}{EBP^2}$$
 provides the number of participants needed to

estimate the population proportion with confidence $1 - \alpha$ and margin of error EBP .

Use the normal distribution for a single population proportion p'

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

$$EBP = (Z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) (p' + q') = 1$$

The confidence interval has the format $(p' - EBP, p' + EBP)$.

\bar{x} is a point estimate for μ

p' is a point estimate for ρ

s is a point estimate for σ

PART VIII

HYPOTHESIS TESTING WITH ONE SAMPLE

40. Introduction: Hypothesis Testing with One Sample

Page by: OpenStax College

Summary



If you want to test a claim that the mean sales of restaurants in Odessa for a month is more than \$20,000 you can conduct a hypothesis test.

Learning Objectives

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population

parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year.

A statistician will make a decision about these claims. This process is called

“hypothesis testing.” A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists

of two contradictory hypotheses or statements, a decision based on the data, and a conclusion.

To perform a hypothesis test, a statistician will:

- Set up two contradictory hypotheses.
- Collect sample data (in homework problems, the data or summary statistics will be given to you).
- Determine the correct distribution to perform the hypothesis test.

- Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
 - Make a decision and write a meaningful conclusion.
-
-

Glossary

Confidence Interval:

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard

deviation).

- The sample and its size.

Hypothesis Testing:

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

4I. 9.1 Null and Alternative Hypotheses

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

H_0 (**The null hypothesis**): It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

H_a (**The alternative hypothesis**): It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0 .

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

After you have determined which hypothesis the sample supports, you make a decision. There are two options for a **decision**.

They are either:

1. “reject H_0 ” if the sample information favors the alternative hypothesis.
2. “do not reject H_0 ” / “decline to reject H_0 ” if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a :

H_0	H_a
equal (=)	not equal (\neq) or greater than ($>$) or less than ($<$)
greater than or equal to (\geq)	less than ($<$)
less than or equal to (\leq)	more than ($>$)

Note:

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it.

The choice of symbol depends on the wording of the hypothesis test.

However, be aware that many researchers (including one of the co-authors in research work) use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis.

This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.



One or more interactive elements has been excluded from this version of the text. You can view them online

here: <https://library.achievingthedream.org/odessastatistics/?p=65#oembed-1>

Example 1

We want to test if More than 30% of the registered voters in Santa Clara County voted in the primary election. State the null and alternative hypothesis.

Null Hypothesis

H_0 : No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$

Show Answer

H_a : More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Try It

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

[practice-area rows="2"][/practice-area]

Show Answer

H_0 : The drug reduces cholesterol by 25%. $p = 0.25$

H_a : The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example 2

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

Null Hypothesis

$$H_0: \mu = 2.0$$

Alternative Hypothesis

$$H_a: \mu \neq 2.0$$

Try It

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol ($=$, \neq , \geq , $<$, \leq , $>$) for the null and alternative hypotheses.

$$H_0: \mu _ _ 66$$

$$H_a: \mu _ _ 66$$

[practice-area rows="1"][/practice-area]

Show Answer

$$H_0 : \mu = 66$$

$$H_a : \mu \neq 66$$

Example 3

We want to test if college students take less than five years to

graduate from college, on the average.

The null and alternative hypotheses are:

Null Hypothesis

$$H_0: \mu \geq 5$$

Alternative Hypothesis

$$H_a: \mu < 5$$

Try It

We want to test if the time to teach a lesson plan is fewer than 45 minutes. State the null and alternative hypotheses.

[practice-area rows="1"][/practice-area]

Show Answer

$$H_0: \mu \geq 45$$

$$H_a: \mu < 45$$

Example 4

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

Null Hypothesis

$$H_0: p \leq 0.066$$

Alternative Hypothesis

$$H_a: p > 0.066$$

Try It

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try.

State the null and alternative hypotheses.

[practice-area rows="2"][/practice-area]

Show Answer

$$H_0: p = 0.40$$

$$H_a: p > 0.40$$

Concept Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we: Evaluate the **null hypothesis**, typically denoted with H_0 . The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality ($=$, \leq or \geq) Always write the **alternative hypothesis**, typically denoted with H_a or H_1 , using less than, greater than, or not equals symbols, i.e., (\neq , $>$, or $<$). If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Formula Review

H_0 and H_a are contradictory.

42. 9.2 Outcomes, Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not.

The outcomes are summarized in the following table:

Action	H ₀ is actually	
	True	False
Do not reject H ₀	Correct Outcome	Type II Error (β)
Reject H ₀	Type I Error (α)	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject H₀** when **H₀ is true (correct decision)**.
2. The decision is to **reject H₀** when **H₀ is true** (incorrect decision known as a **Type I error**).
3. The decision is **not to reject H₀** when **H₀ is false** (incorrect decision known as a **Type II error**).
4. The decision is to **reject H₀** when **H₀ is false (correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is $1 - \beta$.
 Since β is probability of making type II error, we want this probability to be small.
 In other words, we want the value $1 - \beta$ to be as closed to one as possible.
 Increasing the sample size can increase the Power of the Test.

Example 1

Suppose the null hypothesis, H_0 , is: Frank's rock climbing equipment is safe.

- **Type I error:** Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- **Type II error:** Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = **Probability** that Frank thinks his rock climbing equipment may not be safe when it really is safe.

β = **Probability** that Frank thinks his rock climbing equipment may be safe when it is not safe.

Frank's decision	Null Hypothesis: The rock climbing equipment is safe.	
	True (The equipment is safe)	False (The equipment is not safe.)
Not reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

Notice that, in this case, the error with the greater consequence is the Type II error.

(If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Try It

Suppose the null hypothesis, H_0 , is: the blood cultures contain no traces of pathogen X.
State the Type I and Type II errors.

[practice-area rows="2"][/practice-area]

Solution

Type I error: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.

Type II error: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Null hypothesis: The blood cultures contain no traces of pathogen X.		
Decision (What the researcher thinks...)	True (The blood cultures contain no traces of pathogen X.)	False (The blood cultures contain traces of pathogen X.)
Not reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

Example 2

Suppose the null hypothesis, H_0 : The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

State the 4 possible outcomes of performing a hypothesis test.

Solution:

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = $P(\text{Type I error})$.

β = **probability** that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.

Null hypothesis: The victim's situation is alive.		
Decision	True (The victim is alive.)	False (The victim is dead.)
Not reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)



One or more interactive elements has been excluded

from this version of the text. You can view them online here: <https://library.achievingthedream.org/odessastatistics/?p=66#oembed-1>

Try It

Suppose the null hypothesis, H_0 , is: A patient is not sick. Which type of error has the greater consequence, Type I or Type II?

[practice-area rows="1=2"][/practice-area]

Solution

Type I Error: The patient will not be thought well when, in fact, he is not sick.

Type II Error: The patient will be thought well when, in fact, he is sick.

Null hypothesis: A patient is not sick.		
Decision	True (The patient is not sick.)	False (The patient is sick.)
Not reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick.
He will not be able to get treatment.

Example 3

Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0 , is: Boy Genetic Labs has no effect on gender outcome. Which type of error has the greater consequence, Type I or Type II?

Solution:

H_0 : Boy Genetic Labs has no effect on gender outcome.

H_a : Boy Genetic Labs has effect on gender outcome.

Null Hypothesis : Boy Genetic Labs has no effect on gender outcome.		
Decision	True (No Effect)	False (Effect)
Not reject H_0	Correct decision	Type II Error
reject H_0	Type I Error	Correct decision

- **Type I error:** This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that Boy Genetic Labs influences the gender outcome, when in fact it has no effect.
The probability of this error occurring is denoted by the Greek letter alpha, α .
- **Type II error:** This results when we fail to reject a false null hypothesis. In context, we would state that Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does.
The probability of this error occurring is denoted by the Greek letter beta, β .

The error of greater consequence would be the Type I error since couples would use the Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Try It

“Red tide” is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside.

Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Solution

In this scenario, an appropriate null hypothesis would be

H_0 : the mean level of toxins is at most 800 μg . ($H_0 : \mu_0 \leq 800 \mu\text{g}$)

H_a : the mean level of toxins exceeds 800 μg . ($H_a : \mu_0 > 800 \mu\text{g}$)

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800 μg .

The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than 800 μg). The DMF lifts the harvesting ban.

Null Hypothesis: The mean level of toxins is at most 800 μg .		
Decision	True	False
Not reject H_0	Correct decision	Type II Error
Reject H_0	Type I Error	Correct decision

This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food. In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example 4

A certain experimental drug claims a cure rate of higher than 75% for males with prostate cancer.

Describe both the Type I and Type II errors in context. Which error is the more serious?

Solution:

H_0 : The cure rate is less than 75%.
 H_a : The cure rate is higher than 75%.

Null hypothesis (The cure rate is less than 75%)		
Decision	True (The cure rate is less than 75%)	False (The cure rate is higher than 75%)
Not reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

- **Type I:** A cancer patient believes the cure rate for the drug is more than 75% when the cure rate actually is less than 75%.
- **Type II:** A cancer patient believes the the cure rate is less than 75% cure rate when the cure rate is actually higher than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Try It

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0 , that states the percentage of adults with jobs is at least 88%.

Identify the Type I and Type II errors from these four statements.

a) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%

b) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

c) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

d) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Type I error:

c

Type II error:

b

Example 5

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0 , that states the percentage of adults with jobs is at least 88%.

Identify the Type I and Type II errors from these four statements.

a) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%

b) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

c) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

d) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Solution:

If H_0 : The percentage of adults with jobs is at least 88%, then H_a : The percentage of adults with jobs is less than 88%.

Type I error:

c

Type II error:

b

Concept Review

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II** error occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β , for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

α = probability of a Type I error = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

43. 9.3 Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.**

Perform tests of a population mean using a **normal distribution** or a **Student's t-distribution**.

(Remember, use a Student's t-distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.)

We perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$$\overline{X} \sim N\left(\mu_x \frac{\sigma_x}{\sqrt{n}}\right) \quad \text{or} \quad t_{df}$$

- The population parameter is μ .
- The estimated value (point estimate) for μ is \overline{x} , the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

- The population parameter is p .
- The estimated value (point estimate) for p is p' .

$p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean μ** using a **Student's t-distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly.

1. Your data should be a **simple random sample**.
2. Your data comes from a population that is approximately **normally distributed**.
3. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean μ** using a normal distribution (often called a z-test), the assumptions are:

1. You take a simple random sample from the population.
2. The population you are testing is normally distributed or your sample size is sufficiently large.
3. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion p** , you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are as follows:

1. There are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p . The quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$).
2. The shape of the binomial distribution needs to be similar to

the shape of the normal distribution. The binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Concept Review

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > 5$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

If there is no given preconceived α , then use $\alpha = 0.05$.

Types of Hypothesis Tests:

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation.

$$\text{Means: } \mu = \mu_{\bar{x}} \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation.

$$\text{Proportions: } \mu = p \quad \text{and} \quad \sigma = \sqrt{\frac{pq}{n}}$$

44. 9.4 Rare Events, the Sample, Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption**—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the **p-value**. The p-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large p -value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the p -value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p -value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example 1

A baker bakes 10 loaves of bread. The **mean** height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. The distribution of heights is normal. He claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him.

To persuade his customers that he is right, the baker decides to do a hypothesis test.

Solution:

Since the baker knows the standard deviation from baking hundreds of loaves of bread, we will run **Normal Z-Test**.

The null hypothesis could be $H_0: \mu \leq 15$.

The alternate hypothesis is $H_a: \mu > 15$.

The words “**is more than**” translates as a “ $>$ ” so “ $\mu > 15$ ” goes into

the alternate hypothesis.

The null hypothesis must contradict the alternate hypothesis.

Since σ is **known** ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with

- mean $\mu = 15$ and
- standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p -value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p -value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.

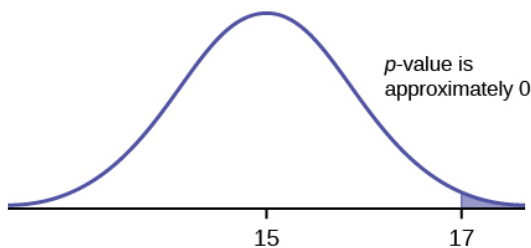


Figure 1

$$\begin{aligned} & p\text{-value} \\ &= P(\bar{X} > 17) \end{aligned}$$

$$= P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{17 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{17 - 15}{\frac{0.5}{\sqrt{10}}}\right)$$

= P(Z > 12.64911) which is approximately zero.

A *p*-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on average.

That is, almost 100% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm.

Because the outcome of 17 cm is so **unlikely to happen (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.).

There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Using TI-83/84:

1. Press STAT.
2. Arrow right to TESTS.
3. Choose 1: Z-Test.
4. For Inpt, arrow right to STATS and press ENTER.
5. Input the following information.

$$\mu_0 = 15$$

$$\beta = 0.5$$

$$\overline{X} = 17$$

$$n = 10$$

$$\mu : > \mu_0$$

6. Then arrow down to CALCULATE.

The result:

$$\mu > 15$$

$$z = 12.64911064$$

$$p = 5.854831 \times 10^{-37}$$

$$\overline{X} = 17$$

$$n = 10$$

Interpretation of the result:

The z-score of height 17cm is 12.64911. The blue shaded area of Figure 1, also known as p-value, is 5.854831×10^{-37} .

Try It

A normal distribution has a standard deviation of 1.
We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

$$H_0: \mu \leq 12$$

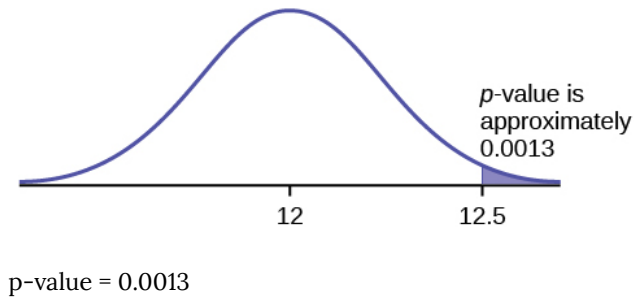
$$H_a: \mu > 12$$

The p -value is 0.0013

Draw a graph that shows the p -value.

[practice-area rows="1"][/practice-area]

Solution



Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the null hypothesis is to compare the p -value and a **preset or preconceived α (also called a “significance level”)**. A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_0 , do as follows:

- If $\alpha > p\text{-value}$, reject H_0 .
The results of the sample data are significant.
There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a , may be correct.
($p\text{-value} < \alpha$, then **reject H_0** . Hence we have **sufficient evidence** to conclude H_a .)
- If $\alpha \leq p\text{-value}$, do not reject H_0 . The results of the sample data are not significant.
There is not sufficient evidence to conclude that the alternative hypothesis, H_a , may be correct.
($p\text{-value} > \alpha$, then **not reject H_0** . Hence we have **no sufficient evidence** to conclude H_a .)

When you “do not reject H_0 ,” it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Example 2

When using the p -value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

If the p -value is low, the null must go.

If the p -value is high, the null must fly.

This memory aid relates a p -value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates

a p -value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Solution:

Reject the null hypothesis when

Show Answer

the p -value is less than the established value of α .

The results of the sample data

Show Answer

support the alternative hypothesis.

Do not reject the null hypothesis when

Show Answer

the p -value is greater than the established value of α .

The results of the sample data

--
Show Answer

do not support the alternative hypothesis.

Try It

CuteBaby Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

$$H_0: p = 0.50, H_a: p > 0.50$$

$$\alpha = 0.01$$

$$p\text{-value} = 0.025$$

Interpret the results and state a conclusion in simple, non-technical terms.

Show Answer

Since the p-value is greater than the established value of α (the p-value is higher), we do not reject the null hypothesis.

There is not enough evidence to support Cutebaby Genetics Labs' stated claim that their procedures improve the chances of a boy being born.

45. 9.5 Additional Information and Full Hypothesis Test Examples

- In a hypothesis test problem, you may see words such as “the level of significance is 1%.” The “1%” is the preconceived or preset α .
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, a common standard to use is $\alpha = 0.05$.**
- When you calculate the p -value and draw the picture, the p -value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternative hypothesis**, H_a , tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the p -value:** A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p -value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p -value such as 0.4, as opposed to a p -value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 1

$$H_0: \mu = 5$$

$$H_a: \mu < 5$$

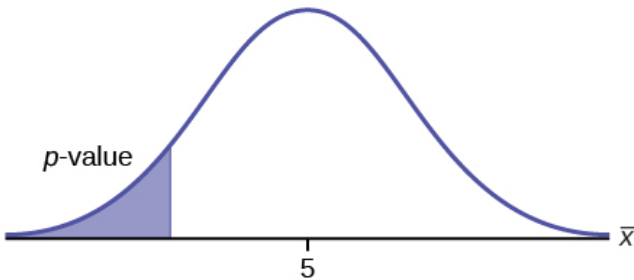
Significance level = 5%

Assume the p -value is 0.0243.

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.
3. Draw the picture of the p -value.
4. Do we reject null hypothesis, $H_0: \mu = 5$?
5. Do we have enough evidence to conclude that $\mu < 5$?

[Click here to show solution:](#)

1. Test of a single population mean.
2. H_a tells you the test is left-tailed.
3. The picture of the p -value is as follows:



4. Since $p\text{-value} < \text{significance level}$, we reject null hypothesis, $H_0: \mu = 5$.
5. We have enough evidence to conclude that $H_a: \mu < 5$.

Try It

$$H_0: \mu = 10$$

$$H_a: \mu < 10$$

Significance level = 5% = 0.05

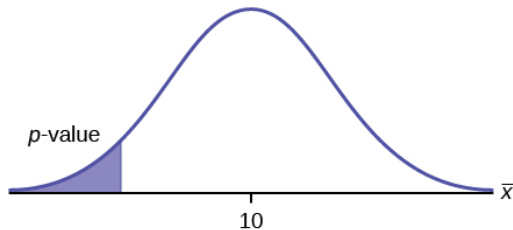
Assume the p -value is 0.0435.

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.
3. Draw the picture of the p -value.
4. Do we reject null hypothesis, $H_0: \mu = 10$?
5. Do we have enough evidence to conclude that $\mu < 10$?

[practice-area rows="1"][/practice-area]

Click here to show solution:

1. Test of a single population mean.
2. left-tailed test
3. The picture of the p -value is as follows:



4. Since $p\text{-value} < \text{significance level}$, we reject null hypothesis, $H_0: \mu = 10$.
5. We have enough evidence to conclude $H_a: \mu < 10$.

10.

Example 2

$H_0: p \leq 0.2$

$H_a: p > 0.2$

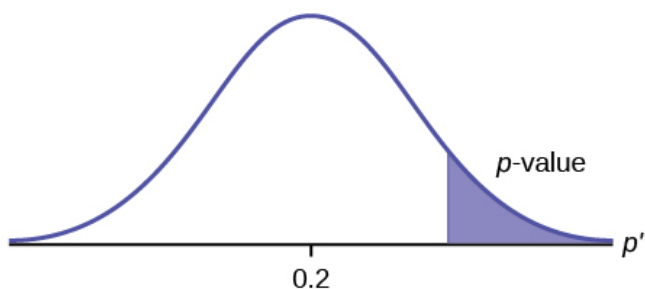
Significance level = 0.05

Assume the p -value is 0.0719.

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.
3. Draw the picture of the p -value.
4. Do we reject null hypothesis, $H_0: p \leq 0.2$?
5. Do we have enough evidence to conclude that $p > 0.2$?

Click here to show solution:

1. This is a test of a single population proportion.
2. H_a tells you the test is right-tailed.
3. The picture of the p -value is as follows:



4. Since $p\text{-value} > \text{significance level}$, we do not reject null hypothesis ($H_0: p \leq 0.2$).
 5. We do not have enough evidence to conclude $H_a: p > 0.2$.
-

Try It

$$H_0: \mu \leq 1$$

$$H_a: \mu > 1$$

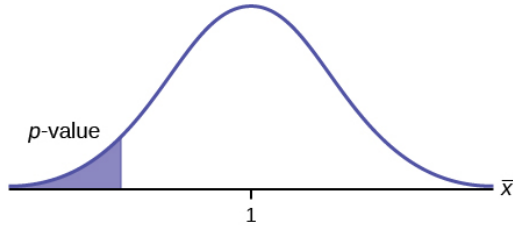
Significance level = 1%

Assume the $p\text{-value}$ is 0.1243.

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.
3. Draw the picture of the $p\text{-value}$.
4. Do we reject null hypothesis, $H_0: \mu \leq 1$?
5. Do we have enough evidence to conclude that $\mu > 1$?

Show Answer

1. Test of a population mean.
2. right-tailed test
3. The picture of the p-value is as follows:



4. Since $p\text{-value} > \text{significance level}$, we do not reject null hypothesis, $H_0: \mu \leq 1$.
5. We do not have enough evidence to conclude that $\mu > 1$.

Example 3

$$H_0: p = 50$$

$$H_a: p \neq 50$$

Significance level = 1%

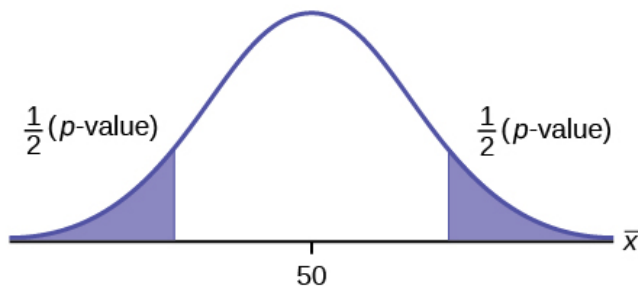
Assume the p -value is 0.0005

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.

3. Draw the picture of the p -value.
4. Do we reject null hypothesis, $H_0: p = 50$?
5. Do we have enough evidence to conclude that $p \neq 50$?

Show Answer

1. This is a test of a single population mean.
2. H_a tells you the test is two-tailed.
3. The picture of the p -value is as follows:



4. Since $p\text{-value} < \text{significance level}$, we reject null hypothesis, $H_0: p = 50$.
5. We have enough evidence to conclude $H_a: p \neq 50$.

Try It

$$H_0: p = 0.5$$

$$H_a: p \neq 0.5$$

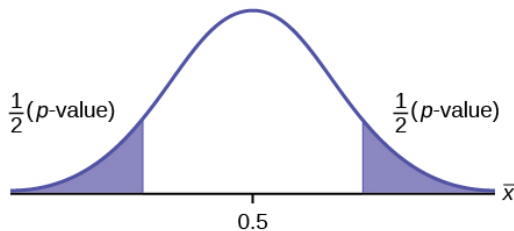
Significance level = 0.05

Assume the p -value is 0.2564.

1. What type of test is this?
2. Determine if the test is left, right, or two-tailed.
3. Draw the picture of the p -value.
4. Do we reject null hypothesis, $H_0: p = 0.5$?
5. Do we have enough evidence to conclude that $p \neq 0.5$?

[Click here to show solution:](#)

1. Hypothesis test of a single population proportion.
2. two-tailed test
3. The picture of the p -value is as follows:



4. Since $p\text{-value} > \text{significance level}$, we do not reject null hypothesis ($H_0: p = 0.5$).
5. We do not have enough evidence to conclude $H_a: p \neq 0.5$.

Steps to set up a hypothesis test:

- Set up H_0 and H_a .
- Determine the significance level.
- Find p-value.
- Compare p-value and significance level, (α).
- Decide if we reject / not reject H_0 .

If p-value < significance level, then reject H_0 . Therefore, enough evidence to conclude H_a .

If p-value > significance level, then not reject H_0 . Therefore, not enough evidence to conclude H_a .

- Conclusion.
-
-
-

Full Hypothesis Test Examples

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**.

For the 15 swims, **Jeffrey's mean time was 16 seconds**. Frank **thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds**.

Conduct a hypothesis test using 5% significance level. Assume that the swim times for the 25-yard freestyle are normal.

Solution:

mean = 16.43 seconds, standard deviation = 0.8 seconds.

Since the problem is about a mean, this is a **test of a single population mean**.

1. What are we testing?

$$H_0: \mu = 16.43$$

$$H_a: \mu < 16.43$$

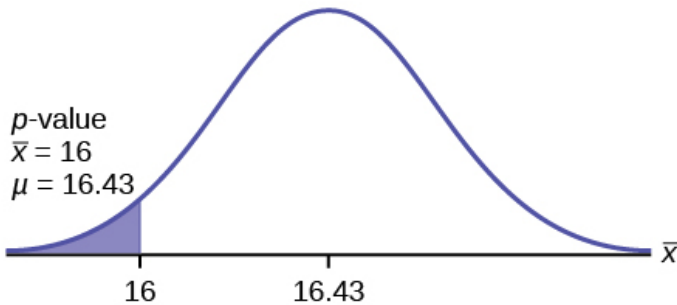
For Jeffrey to swim faster, his time will be less than 16.43 seconds. The “<” tells you this is left-tailed.

2. What is the significance level?

significance level, $\alpha = 5\% = 0.05$

3. What is the p-value?

Graph:



\bar{X} = the mean time to swim the 25-yard freestyle.

\bar{X} is normal. Population standard deviation is known: $\sigma = 0.8$

$$\bar{X} = 16,$$

$\mu = 16.43$ (comes from H_0 and not the data.)

$\sigma = 0.8$, and

$n = 15$.

Calculate the p-value using the normal distribution for a mean:

$$\begin{aligned}
& \overset{p\text{-value}}{=} P\left(\bar{X} < 16\right) \\
& = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{16 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
& = P\left(Z < \frac{16 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
& = P\left(Z < \frac{16 - 16.43}{\frac{0.8}{\sqrt{15}}}\right) \\
& = 0.0187
\end{aligned}$$

where the sample mean in the problem is given as 16.

4. Comparison between p-value and significance level.

p-value = 0.0187 (This is called the **actual level of significance**.)

The p-value is the area to the left of the sample mean is given as 16.

p-value = 0.0187, $\alpha = 0.05$

Therefore, $\alpha > p\text{-value}$.

Interpretation of the p-value:

If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

5. Decision?

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

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

6. Conclusion?

At the 5% significance level, we conclude that Jeffrey swims faster

using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

Using TI-Calculator to find p-value:

- Press STAT and arrow over to TESTS .
- Press 1:Z-Test . Arrow over to Stats and press ENTER .
- Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ , 16 for the sample mean, and 15 for n .
- Arrow down to μ : (alternate hypothesis) and arrow over to $< \mu_0$.
- Press ENTER .
- Arrow down to Calculate and press ENTER .

The calculator not only calculates the p -value ($p = 0.0187$), but it also calculates the test statistic (z-score) for the sample mean.
 $\mu < 16.43$ is the alternative hypothesis.

Do this set of instructions again except arrow to Draw (instead of Calculate) and press ENTER .

A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ (p -value).

Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the Z-Test function finds the p -value by doing a normal probability calculation using the central limit theorem:

To find $P(\bar{x} < 16)$, we will use TI-Calculator.

Ti-Calculator: 2nd DISTR normcdf ($-\infty, 16, 16.43, \frac{0.8}{\sqrt{15}}$).

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43

seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Try It

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of 2 yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the p -value, sketch the graph, and state your conclusion.

[practice-area rows="4"][/practice-area]

Step-by-Step Solution

Since the problem is about a mean, this is a test of a single population mean.

- $H_0 : \mu = 40$

- $H_a : \mu > 40$
- $\alpha = 0.05$
- $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{45 - 40}{\frac{2}{\sqrt{20}}} = 11.1803$
- When $\bar{X} = 45$, its corresponding z-score is 11.1803.

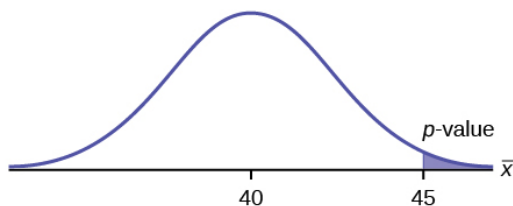
The area to the right when $\bar{X} = 45$

= The area to the right of z-score = 11.1803

= blue shaded area

= p-value

= $2.6115 * 10^{-29}$



- Because $p < \alpha$, we reject the null hypothesis.
- There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

Using TI-Calculator to solve:

- Press STAT and arrow over to TESTS.
- Press 1:Z-Test.
- Arrow over to Stats and press ENTER.
- Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ , 45 for the sample mean, and 20 for n.

- Arrow down to μ : (alternative hypothesis) and set it either as $<$, \neq , or $>$.
- Press ENTER.
- Arrow down to Calculate and press ENTER.

The p-value = 0.0062. $\alpha = 0.05$.

Because $p < \alpha$, we reject the null hypothesis.

There is sufficient evidence to suggest that the change in grip improved Marco's throwing distance.

The calculator not only calculates the p-value but it also calculates the test statistic (z-score) for the sample mean. Select $<$, \neq , or $>$; for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and p-value. Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 4

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3) 215(3) 225(1) 241(2) 252(2) 265(2) 275(2) 313(2) 316(5) 338(2) 341(1) 345(2) 368(2) 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution:

1. What are we testing?

Since the problem is about a mean weight, this is a **test of a single population mean**.

$$H_0: \mu = 275$$

$$H_a: \mu > 275$$

$$\text{Significance level, } \alpha = 2.5\% = 0.025$$

This is a right-tailed test.

2. What is the significance level, α ?

$$\alpha = 2.5\% = 0.025$$

3. Find the p-value.

\bar{X} = the mean weight (in pounds) lifted by the football players.

Distribution for the test: X is normally distributed because σ is known.

$$\bar{x} = 286.2, n = 30, \sigma = 55 \text{ pounds (Always use } \sigma \text{ if you know it.)}$$

We assume $\mu = 275$ pounds unless our data shows us otherwise.

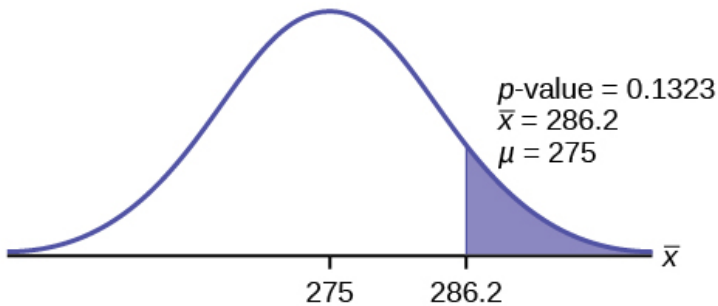
Calculate the p -value using the normal distribution for a mean and using the sample mean as input.

$$\begin{aligned} & p\text{-value} \\ &= P(\bar{x} > 286.2) \\ &= P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{286.2 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\ &= P\left(Z > \frac{286.2 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\ &= P\left(Z > \frac{286.2 - 275}{\frac{55}{\sqrt{30}}}\right) \\ &= P(Z > 1.1153623) \\ &= 0.1323. \end{aligned}$$

4. Comparison between p -value and significance level.

Interpretation of the p -value:

If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



$$\alpha = 0.025, p\text{-value} = 0.1323$$

5. Decision?

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

6. Conclusion?

Conclusion:

At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p -value can easily be calculated.

- Put the data and frequencies into lists.
- Press STAT and arrow over to TESTS .
- Press 1:Z-Test .
- Arrow over to Data and press ENTER .
- Arrow down and enter 275 for μ_0 , 55 for σ , the name of the list where you put the data, and the name of the list where you put the frequencies.
- Arrow down to μ : and arrow over to $> \mu_0$.
- Press ENTER .
- Arrow down to Calculate and press ENTER .

The calculator not only calculates the p -value ($p = 0.1331$, a little different from the previous calculation – in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z -score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER . A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p -value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Example 5

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

He performs a hypothesis test using a 5% level of significance.

Solution:

1. What are we testing?

Since we do not know population standard deviation, we are going to run **Student's t Test**.

This is a test of a **single population mean**.

$$H_0: \mu = 65$$

$$H_a: \mu > 65$$

A 5% level of significance means that $\alpha = 0.05$.

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

2. What is the significance level, α ?

The significance level, $\alpha = 5\% = 0.05$

3. Find the p-value.

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

Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution.

This means that the distribution for the test is a **student's t-test**.

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

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

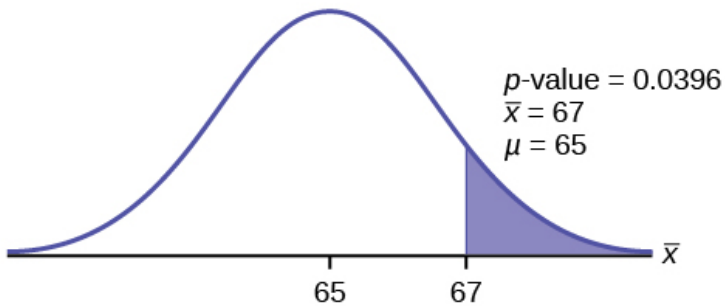
Given that sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data,

$$\begin{aligned} & p\text{-value} \\ &= P(\bar{x} > 67) \\ &= P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{67 - 65}{\frac{3.1972}{\sqrt{10}}}\right) \\ &= P(t > 1.978) \\ &= 0.0396 \end{aligned}$$

4. Comparison between p-value and significance level.

Interpretation of the p-value:

If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 65 or more.



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

5. Decision?

Since $\alpha > p\text{-value}$, reject H_0 .

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

6. Conclusion?

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

The p -value can easily be calculated.

- Put the data into a list.
- Press STAT and arrow over to TESTS .
- Press 2:T-Test . (as we do not have the population standard deviation.)
- Arrow over to Data and press ENTER .
- Arrow down and enter 65 for μ_0 , the name of the list where you put the data, and 1 for Freq: .
- Arrow down to μ : and arrow over to $> \mu_0$.
- Press ENTER . A
- Arrow down to Calculate and press ENTER .

The calculator not only calculates the p -value ($p = 0.0396$) but it also calculates the test statistic (t -score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER . A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p -value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Try It

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock

won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows:

\$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2.

Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p -value, state your conclusion, and identify the Type I and Type II errors.

Show Answer

We run **Student's t-test** as we do not know the population standard deviation.

$$H_0: \mu = 5$$

$$H_a: \mu < 5$$

$$p = 0.0082$$

Because $p < \alpha$, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

Type I Error: To conclude that the stock price is growing slower than \$5 a week when, in fact, the stock price is growing at \$5 a week (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that the stock price is growing at a rate of \$5 a week when, in fact, the stock price is growing slower than \$5 a week (do not reject the null hypothesis when the null hypothesis is false).

Example 6

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution:

1. What are we testing?

This is a **Normal test** of a single population proportion.

$$H_0: p = 0.50$$

$$H_a: p \neq 0.50$$

The 1% level of significance means that $\alpha = 0.01$.

The words “**is different from**” tell you this is a two-tailed test.

2. What is the significance level, α ?

the significance level, $\alpha = 1\% = 0.01$

3. Find the p-value.

This is a two-tailed test. We will include both **left tail** and **right tail** in the hypothesis test.

P' = the percent of of first-time brides who are younger than their grooms.

The proportion of first-time brides who reply that they are younger than their grooms in sample of 100 brides = $\frac{53}{100} = 0.53$.

0.53 is the **right tail** of this test as 0.53 is larger than 0.5 (the population mean).

How about the **left tail**?

Since 0.53 is on the right side of 0.50, the left tail will fall on the left side of 0.50.

Moreover, 0.03 is the difference between 0.53 and 0.50. Hence the distance between the left tail and 0.50 is equal to 0.03 as well.

$$0.50 - 0.03 = 0.47.$$

The **left tail** of this test is 0.47.

Given that $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$,

p-value

= area to the right of right tail + area to the left of left tail

= area to the right of 0.53 + area to the left of 0.47

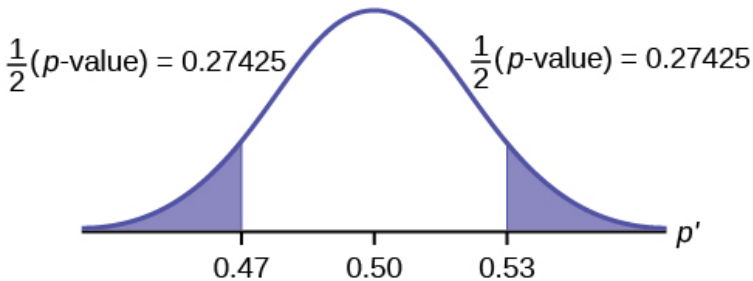
= $P(p' > 0.53) + P(p' < 0.47)$

$$= P \left(\frac{p' - p}{\sqrt{\frac{pq}{n}}} > \frac{0.53 - 0.50}{\sqrt{\frac{(0.5)(0.5)}{100}}} \right) + P \left(\frac{p' - p}{\sqrt{\frac{pq}{n}}} < \frac{0.47 - 0.50}{\sqrt{\frac{(0.5)(0.5)}{100}}} \right)$$

$$= P(Z > 0.6) + P(Z < -0.6)$$

$$= 0.27425 + 0.27425$$

$$= 0.5485$$



4. Comparison between p-value and significance level.

p-value = 0.5485, significance level = 0.01

5. Decision?

Since $p\text{-value} > \text{significance level}$, we do not reject H_0 .

6. Conclusion?

There is no sufficient evidence to suggest that the percentage is different than 50%.

Solution (Using TI-Calculator)

- Press `STAT` and arrow over to `TESTS`.
- Press `5:1-PropZTest`. Enter .5 for p_0 , 53 for x and 100 for n .
- Arrow down to `Prop` and arrow to `not equals p_0` . Press `ENTER`.
- Arrow down to `Calculate` and press `ENTER`.
- The calculator calculates the $p\text{-value}$ ($p = 0.5485$) and the test statistic ($z\text{-score}$). `Prop not equals .5` is the alternate hypothesis.

Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ ($p\text{-value}$). Make sure when you use `Draw` that no other equations are highlighted in `Y =` and the plots are turned off.

Try It

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The

teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p -value, sketch the graph, and state your conclusion.

Show Answer

Since the problem is about percentages, this is a test of single population proportions.

$$H_0 : p = 0.85$$

$$H_a : p \neq 0.85$$

$$p = 0.7554$$

Because $p > \alpha$, we fail to reject the null hypothesis.

There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example 7

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company survey 150 households with the result that 43 of the households have three cell phones. They believe that the proportion

is less than 30%. Conduct a hypothesis test to check their claim at 5% significance level.

Show Answer

This is a Normal test for a population proportion.

$$H_0 : p = 0.3$$

$$H_a : p < 0.3$$

5% significance level means $\alpha = 0.05$.

Given that $p = 0.3$, $q = 1 - p = 1 - 0.3 = 0.7$, $\bar{p} = \frac{43}{150}$, $n = 150$

$$\begin{aligned} & \text{p-value} \\ &= P (\bar{p} < 0.30) \\ &= P (\frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} < \frac{\frac{43}{150} - 0.30}{\sqrt{\frac{(0.3)(0.7)}{150}}}) \\ &= P (Z < -0.3563) \end{aligned}$$

$$= 0.3608$$

Since $p\text{-value} > \alpha$, we do not reject H_0 .

We do not have sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Try It

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the p -value, state your conclusion, and identify the Type I and Type II errors.

[Click here to show solution:](#)

$H_0: p = 0.92$

$H_a: p < 0.92$

$p\text{-value} = 0.0046$

Because $p\text{-value} < 0.05$, we reject the null hypothesis.

There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

Example 8

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass. (Assume the population is normal.)

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

Is there convincing evidence that the average conductivity of this type of glass is greater than 1?

Use a significance level of 0.05.

Step-by-step Solution:

This is a **Student's t-test** as we do not know the population standard deviation.

$$H_0 : \mu = 1$$

$$H_a : \mu > 1$$

Significance level is 5% ($\alpha = 0.05$)

p-value

$$\begin{aligned} &= P(\bar{X} > 1) \\ &= P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{1.04 - 1}{\frac{0.0659}{\sqrt{11}}}\right) \\ &= P(Z > 2.013126) \\ &= 0.0359 \end{aligned}$$

Since p-value < significance level, we reject null hypothesis.

We have sufficient evidence to conclude that the average conductivity of this type of glass is greater than 1.

Using TI-Calculator to solve:

This is a **Student's t-test** as we do not know the population standard deviation.

$$H_0 : \mu = 1$$

$$H_a : \mu > 1$$

Significance level is 5% ($\alpha = 0.05$)

We will input the sample data into the TI-83 as follows.

L1	L2	L3	1
1.08			
.98			
.98			
1.02			
.95			
.95			
1.00			

L1(12) =

T-Test

Inpt: Date Stats

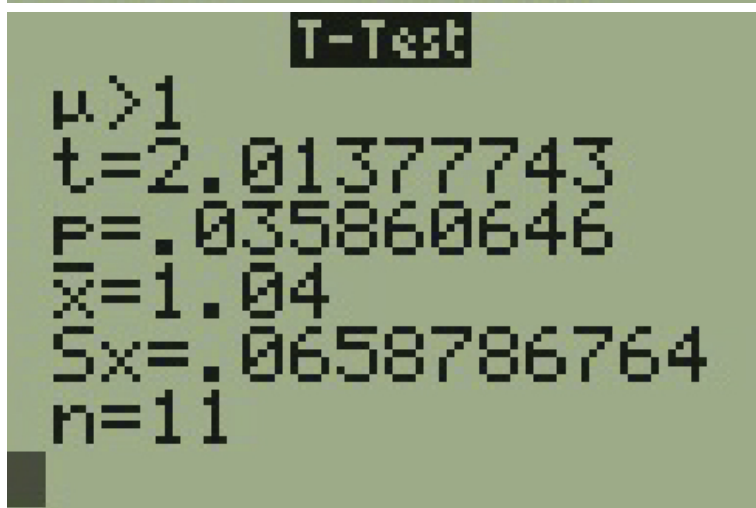
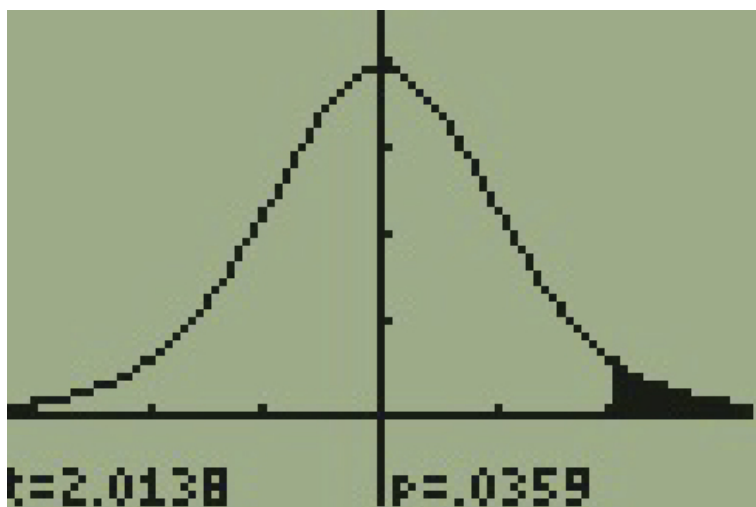
μ_0 : 1

List: L1

Freq: 1

μ : $\neq \mu_0$ $< \mu_0$ $> \mu_0$

Calculate Draw



p-value = 0.03586

Since p-value < significance level, we reject null hypothesis.

We have sufficient evidence to conclude that the average conductivity of this type of glass is greater than 1.

Try It

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). (Since this is a critical issue, use a 0.5% significance level to run the hypothesis test.

Step-by-step solution:

This is a **Normal test** for a single population proportion.

$$H_0 : p = 0.00034$$

$$H_a : p > 0.00034$$

Significance level is 0.5% ($\alpha = 0.005$)

$$\text{Given that } \bar{p} = \frac{172}{420,019}, p = 0.00034, q = 1 - p = 1 - 0.00034 = 0.99966, n = 420019$$

p-value

$$= P(\bar{p} > 0.00034)$$

$$= P\left(\frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} > \frac{\frac{172}{420,019} - 0.00034}{\sqrt{\frac{0.00034 \cdot 0.99966}{420,019}}}\right)$$

$$= P(Z > 2.44336)$$

$$= 0.00728$$

Since p-value < significance level, we reject null hypothesis.

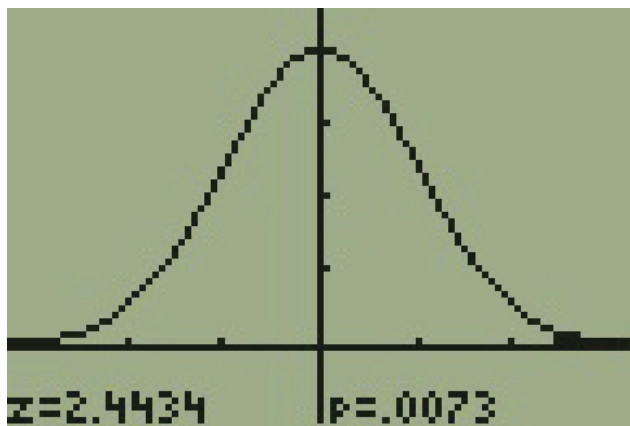
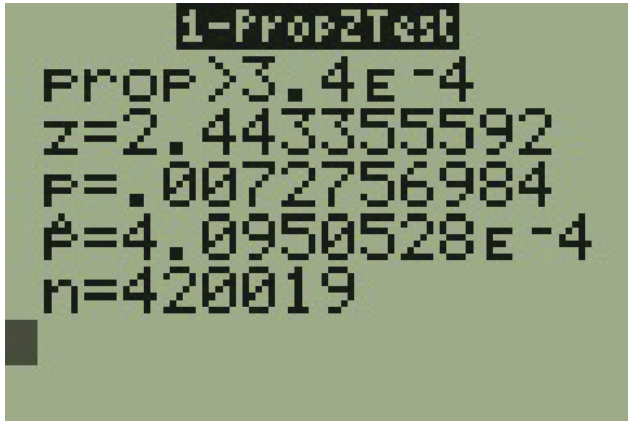
We have sufficient evidence to conclude that cell phone users developed brain cancer at a greater rate.

Using TI-Calculator to solve:

$$H_0 : p = 0.00034$$

$$H_a : p > 0.00034$$

Significance level is 0.5% ($\alpha = 0.005$)



p-value = 0.00728

Since $p\text{-value} < \text{significance level}$, we reject null hypothesis.

We have sufficient evidence to conclude that cell phone users developed brain cancer at a greater rate.

Example 9

According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.07734%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use 1% significance level.

Show Answer

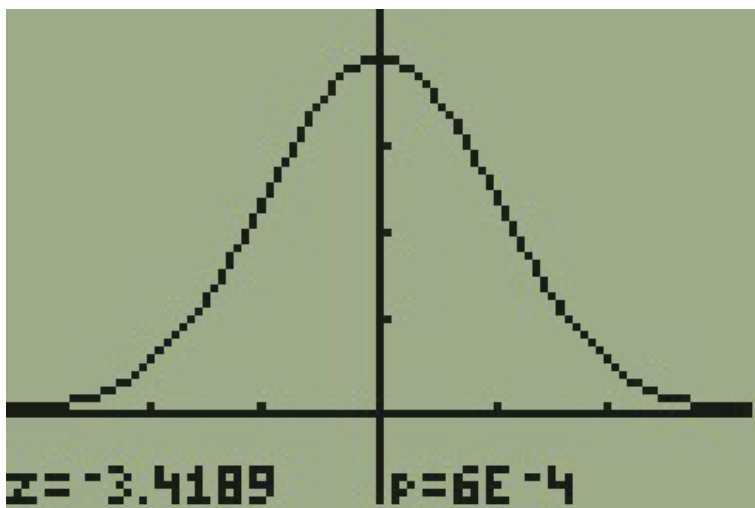
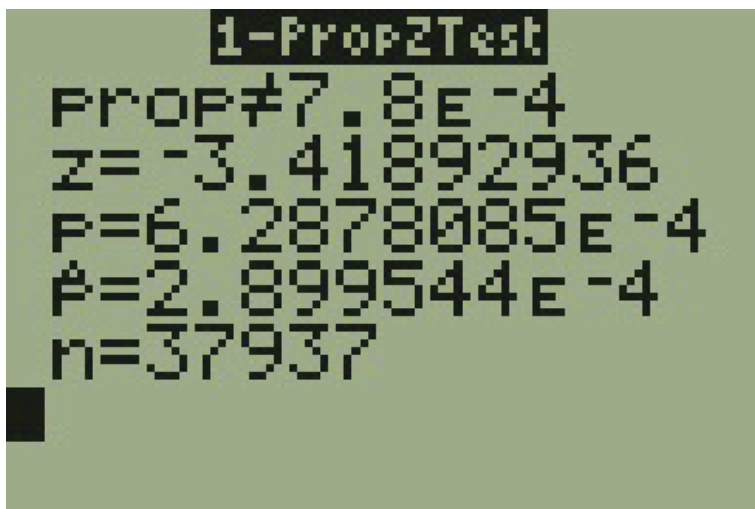
This is a **Normal test** for a single population proportion.

$H_0 : p = 0.00077734$

$H_a : p \neq 0.0007734$

Significance level is 1% ($\alpha = 0.001$)

The following screen shots display the summary statistics from the hypothesis test.



p-value = 0.0006288

Since p-value < significance level, we reject null hypothesis.
There is sufficient evidence to conclude that there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage.

Concept Review

The hypothesis test itself has an established process. This can be summarized as follows:

Determine H_0 and H_a . Remember, they are contradictory.

Determine the random variable.

Determine the distribution for the test.

Draw a graph, calculate the test statistic, and use the test statistic to calculate the p -value. (A z -score and t -score are examples of test statistics.)

Compare the preconceived α with the p -value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β . β is needed to help determine the sample size of the data that is used in calculating the p -value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

PART IX

LINEAR REGRESSION AND CORRELATION

46. Introduction: Linear Regression and Correlation



Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience.

(credit: Joshua Rothhaas)

Learning Objectives

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is bivariate data — “bi” for two variables. In reality, statisticians use multivariate data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, “linear regression” with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

47. 12.1 Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + b * x \text{ where } a \text{ and } b \text{ are constant numbers.}$$

The variable **x is the independent variable, and y is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Examples of linear equations:

$$y = 3 + 2x$$

$$y = -0.01 + 1.2x$$

Try It

Is the following an example of a linear equation?

$$y = -0.125 - 3.5x$$

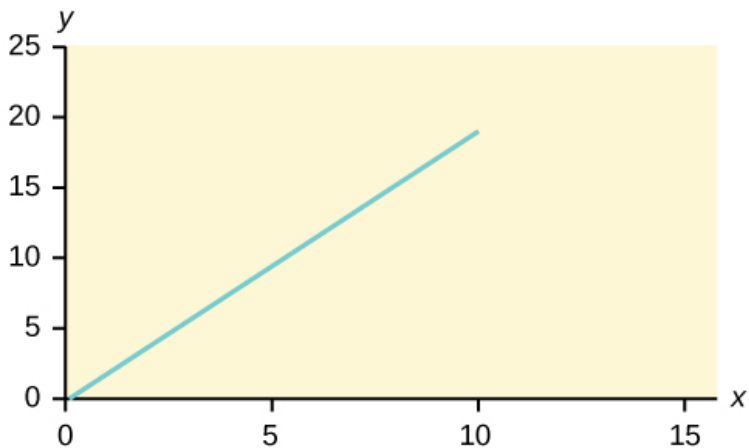
Show Answer

Yes

The graph of a linear equation of the form $y = a + bx$ is a **straight line**.

Any **straight line that is not vertical** can be described by this equation.

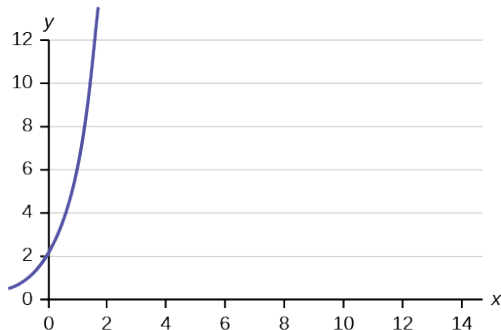
Graph the equation $y = -1 + 2x$.



Try It

Is the following an example of a linear equation? Why

or why not?



Show Answer

No, the graph is not a straight line; therefore, it is not a linear equation.

Example 2

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is \$32 per hour plus a \$31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Show Answer:

Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

The \$31.50 is a fixed cost.

If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only.

The total cost is: $y = 31.50 + 32x$

Try It

Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of \$50 per class as well as \$20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Show Answer

$$y = 50 + 20x$$

Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.



(a)



(b)



(c)

Three possible graphs of $y = a + bx$.

- (a) If $b > 0$, the line slopes upward to the right.
- (b) If $b = 0$, the line is horizontal.
- (c) If $b < 0$, the line slopes downward to the right.

Example 3

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of \$25 plus \$15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Show Answer

The independent variable (x) is the number of hours Svetlana tutors each session.

The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y -intercept is 25 ($a = 25$).

At the start of the tutoring session, Svetlana charges a one-time fee of \$25 (this is when $x = 0$).

The slope is 15 ($b = 15$).

For each session, Svetlana earns \$15 for each hour she tutors.

Try It

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges \$25 plus \$20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y -intercept and what is the slope? Interpret them using complete sentences.

Show Answer

The independent variable (x) is the number of hours Ethan works each visit.

The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y -intercept is 25 ($a = 25$).

At the start of a visit, Ethan charges a one-time fee of \$25 (this is when $x = 0$).

The slope is 20 ($b = 20$).

For each visit, Ethan earns \$20 for each hour he works.

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Concept Review

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y -intercept. Graphically, the y -intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The **y -intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

Formula Review

$y = a + bx$ where a is the y -intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

48. 12.2 Scatter Plots

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y . The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

Example 1

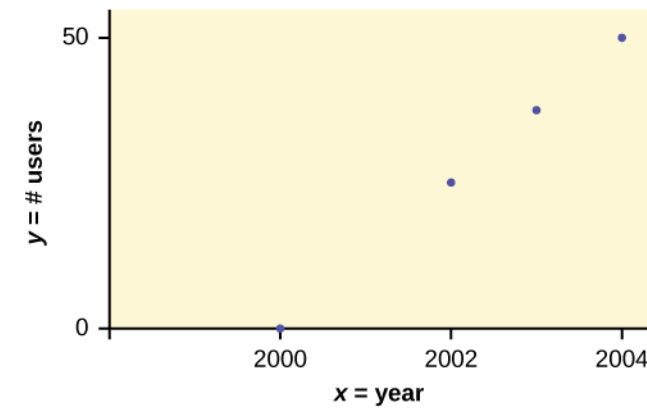
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot.

Solution:

Let x = the year and let y = the number of m-commerce users, in millions.

(year)	(# of users)
2000	0.5
2002	20.0
2003	33.0
2004	47.0

Table showing the number of m-commerce users (in millions) by year.



Scatter plot showing the number of m-commerce users (in millions) by year.

Creating a Scatter Plot (TI-Calculator):

1. Enter your X data into list L1 and your Y data into list L2.
 2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER.
(Make sure the other plots are OFF.)
 3. For TYPE: highlight the very first icon, which is the scatter plot, and press ENTER.
 4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
 5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
 6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
 7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data.
You can press WINDOW to see the scaling of the axes.
-
-

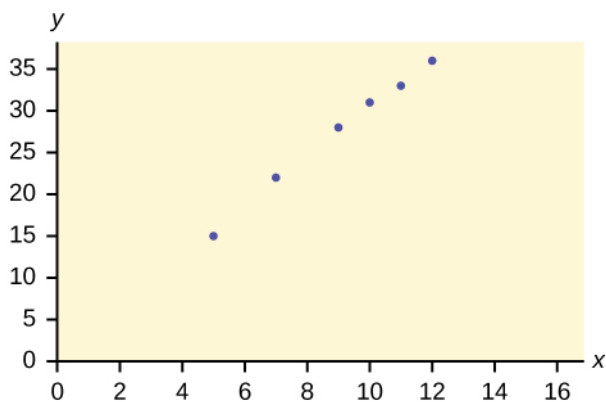
Try It

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	Y (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatter plot and state if what Amelia thinks appears to be true.

Show Answer



Yes, Amelia's assumption appears to be correct.
The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is either: High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable. High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a “perfect fit.” The horizontal line would in fact show no relationship.

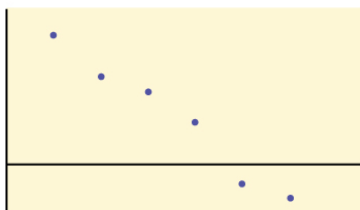
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



(a) Positive linear pattern (strong)



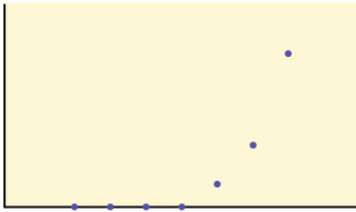
(b) Linear pattern w/ one deviation



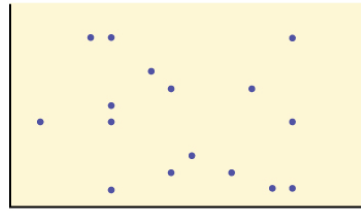
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

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

Concept Review

Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

49. 12.3 The Regression Equation

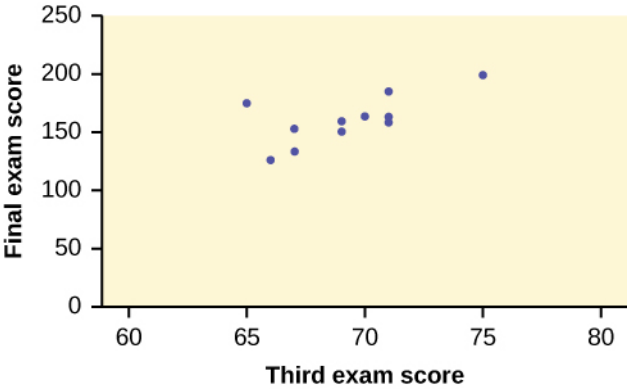
Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to “fit” a straight line. This is called a **Line of Best Fit** or **Least-Squares Line**.

Example 1

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

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

Table showing the scores on the final exam based on scores from the third exam.



Scatter plot showing the scores on the final exam based on scores from the third exam.

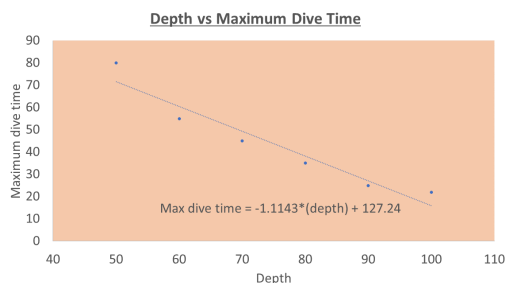
Try It

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table show different depths with the maximum dive times in minutes.

Depth (in feet)	Maximum dive time (in minutes)
50	80
60	55
70	45
80	35
90	25
100	22

1. Can you predict the maximum dive time of a random diver if you know the depth?
2. What is the maximum dive time if a diver dives at 110 feet?

Show Answer

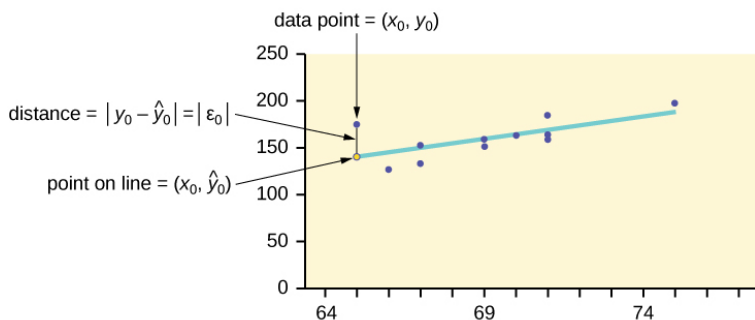


1. $\text{max dive time} = 127.24 - 1.1143 * \text{depth}$
2. At 110 feet, a diver could dive for only five minutes.

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

Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x\hat{y})$.

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



The term $y_0 - \hat{y}_0 = \epsilon_0$ is called the “**error**” or **residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y . In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

- If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y .
- If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y .

In the diagram above, $y_0 - \hat{y}_0 = \epsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ϵ = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors,

$$\epsilon_i = y_i - \hat{y}_i \text{ for } i = 1, 2, 3, \dots, 11.$$

Each $|\epsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ϵ values. If you square each ϵ and add, you get

$$(\epsilon_1)^2 + (\epsilon_2)^2 + \dots + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make

the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx$$

where $a = \bar{y} - b\bar{x}$ and $b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$.

The sample means of the x -values and the y -values are \bar{x} and \bar{y} .

The slope b can be written as $b = r \left(\frac{s_y}{s_x} \right)$ where

- s_y = the standard deviation of the y -values,
- s_x = the standard deviation of the x -values,
- r is the correlation coefficient, which is discussed in the next section.

Least Squares Criteria for Best Fit

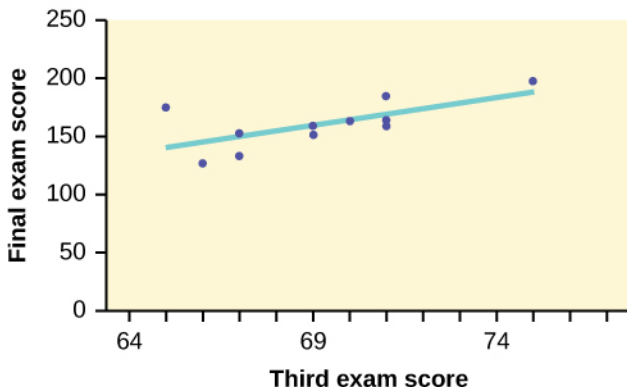
The process of fitting the best-fit line is called **linear regression**. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made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

Note:

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section.

Third Exam vs Final Exam Example

The graph of the line of best fit for the third-exam/final-exam example is as follows:



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

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

Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x -values in the sample data, **but not necessarily for x -values outside that domain.** You could use the line to predict the final

exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x -values in the sample data, which are between 65 and 75.

Understanding Slope

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

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

Third Exam vs Final Exam Example: $\hat{y} = -173.51 + 4.83x$

Slope: The slope of the line is $b = 4.83$.

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

Using the Linear Regression T Test: LinRegTTest

1. In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x, y) values are next to each other in the lists. (If a particular pair of values is

- repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.)
 3. On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
 4. On the next line, at the prompt β or ρ , highlight " $\neq 0$ " and press ENTER
 5. Leave the line for "RegEq:" blank
 6. Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

<p>LinRegTTest Xlist: L1 Ylist: L2 Freq: 1 β or ρ: $\neq 0$ <0 >0 RegEQ: Calculate</p>	<p>LinRegTTest $y = a + bx$ $\beta \neq 0$ and $\rho \neq 0$ $t = 2.657560155$ $p = .0261501512$ $df = 9$ $a = -173.513363$ $b = 4.827394209$ $s = 16.41237711$ $r^2 = .4396931104$ $r = .663093591$</p>
--	---

TI-83+ and TI-84+
calculators

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$, and $b = 4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$. The two items at the bottom are $r^2 = 0.43969$ and $r = 0.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

1. We are assuming your X data is already entered in list L1 and your Y data is in list L2
2. Press 2nd STATPLOT ENTER to use Plot 1

3. On the input screen for PLOT 1, highlight On, and press ENTER
 4. For TYPE: highlight the very first icon which is the scatterplot and press ENTER
 5. Indicate Xlist: L1 and Ylist: L2
 6. For Mark: it does not matter which symbol you highlight.
 7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat"); the calculator will fit the window to the data
 8. To graph the best-fit line, press the "Y=" key and type the equation $-173.5 + 4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
 9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax
-

Note

Another way to graph the line after you create a scatter plot is to use LinRegTTest. Make sure you have done the scatter plot. Check it on your screen. Go to LinRegTTest and enter the lists. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit. Press Y = (you will see the regression equation). Press GRAPH. The line will be drawn."

The Correlation Coefficient, r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the

correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y .

The **correlation coefficient**, r , developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y .

The correlation coefficient is calculated as

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

where n = the number of data points.

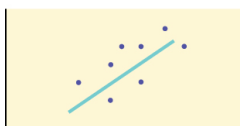
If you suspect a linear relationship between x and y , then r can measure how strong the linear relationship is.

What the VALUE of r tells us: The value of r is always between -1 and $+1$: $-1 \leq r \leq 1$. The size of the correlation r indicates the strength of the linear relationship between x and y . Values of r close to -1 or to $+1$ indicate a stronger linear relationship between x and y . If $r = 0$ there is absolutely no linear relationship between x and y (no linear correlation). If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

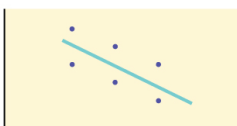
What the SIGN of r tells us: A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (positive correlation). A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (negative correlation).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

Note:

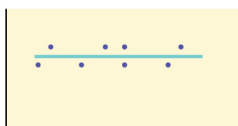
Strong correlation does not suggest that x causes y or y causes x . We say “correlation does not imply causation.”



(a) Positive correlation



(b) Negative correlation



(c) Zero correlation

(a) A scatter plot showing data with a positive correlation. $0 < r < 1$

(b) A scatter plot showing data with a negative correlation. $-1 < r < 0$

(c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r . The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination, r^2

The variable r^2 is called the **coefficient of determination** and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

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

Third Exam vs Final Exam Example:

The line of best fit is $\hat{y} = -173.51 + 4.83x$

The correlation coefficient is $r = 0.6631$

The coefficient of determination is $r^2 = 0.6631^2 = 0.4397$

Interpretation of r^2 in the context of this example: Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line. Therefore, approximately 56% of the variation ($1 - 0.44 = 0.56$) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Concept Review

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y . The Sum of Squared Errors, when set to its minimum, calculates the

points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y . The variable r has to be between -1 and $+1$. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2 , is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

50. 12.4 Prediction

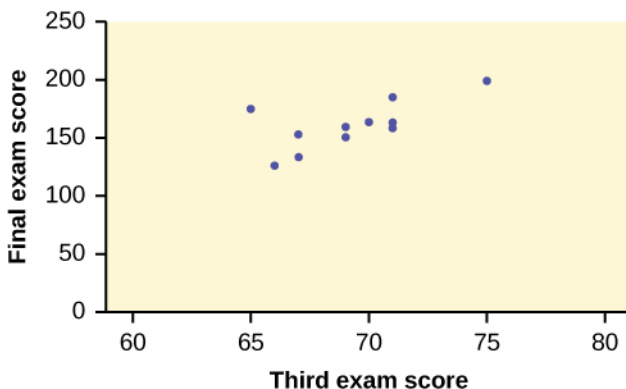
Recall this example from earlier content:

A random sample of 11 statistics students produced the following data, where

x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

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

Table showing the scores on the final exam based on scores from the third exam.



Scatter plot showing the scores on the final exam based on scores from the third exam.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores (**x-values**) range from 65 to 75. **Since 73 is between the x-values 65 and 75**, substitute $x = 73$ into the equation. Then:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 1

Use the data above for this example:

1. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Show Answer

When third exam is 66, final exam score = $-17351 + 4.83(66) = 145.27$

2. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Show Answer

The x values in the data are between 65 and 75.

90 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.)

To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution $x=90$ into the equation.

$$\hat{y} = -173.51 + 4.83(90) = 261.19$$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

Note:

The process of predicting inside of the observed x values observed in the data is called **interpolation**.

The process of predicting outside of the observed x values observed in the data is called **extrapolation**.

Try It

Data are collected on the relationship between the number of hours per week practicing a musical instrument (x) and scores on a math test (y). The line of best fit is as follows:

$$\hat{y} = 72.5 + 2.8x$$

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Show Answer

$$\text{Score} = 72.5 + (2.8)(5) = 86.5$$

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Data from the United States Census Bureau. Available online at http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

Data from the National Center for Health Statistics.

Concept Review

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

51. 12.5 Testing the Significance of the Correlation Coefficient

The correlation coefficient, r , tells us about the strength and direction of the linear relationship between x and y . However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n , together.

We perform a hypothesis test of the “**significance of the correlation coefficient**” to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r , the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r , is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ , the Greek letter “rho.”
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient

ρ is “close to zero” or “significantly different from zero”. We decide this based on the sample correlation coefficient r and the sample size n .

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is “significant.”

Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero. What the conclusion means: There is a significant linear relationship between x and y . We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is “not significant.”

Conclusion: “There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.” What the conclusion means: There is not a significant linear relationship between x and y . Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
 - If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
 - If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.
-

Performing the Hypothesis Test

- Null Hypothesis: $H_0: \rho = 0$
- Alternate Hypothesis: $H_a: \rho \neq 0$

What the Hypotheses Mean in Words

- Null Hypothesis H_0 : The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- Alternate Hypothesis H_a : The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

Drawing a Conclusion

There are two methods of making the decision. The two methods are equivalent and give the same result.

- Method 1: Using the p -value
- Method 2: Using a table of critical values

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

Note

Using the p -value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

Method 1: Using a p -value to make a decision

To calculate the p -value using LinRegTTEST:

- On the LinRegTTEST input screen, on the line prompt for β or ρ , highlight “ $\neq 0$ ”
- The output screen shows the p -value on the line that reads “ $p =$ ”.
- (Most computer statistical software can calculate the p -value.)

If the p -value is less than the significance level ($\alpha = 0.05$)

- Decision: Reject the null hypothesis.
- Conclusion: “There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.”

If the p -value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: “There is insufficient evidence to conclude that

there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero.”

Calculation Notes:

- You will use technology to calculate the p -value. The following describes the calculations to compute the test statistics and the p -value:
- The p -value is calculated using a t -distribution with $n - 2$ degrees of freedom.

- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t , is shown in the computer or calculator output along with the p -value. The test statistic t has the same sign as the correlation coefficient r .
- The p -value is the combined area in both tails.

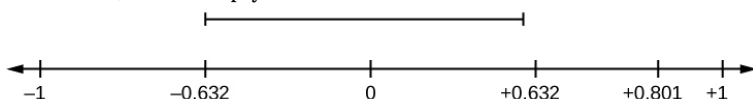
An alternative way to calculate the p -value (**p**) given by LinRegTTest is the command `2*tcdf(abs(t),10^99, n-2)` in 2nd DISTR.

Method 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r is significant or not. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example 1

Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and $+0.632$. If $r < \text{negative critical value}$ or $r > \text{positive critical value}$, then r is **significant**. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.



r is not significant between -0.632 and $+0.632$. $r = 0.801 > +0.632$. Therefore, r is significant.

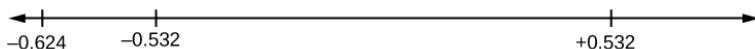
Try It

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points and the critical value is 0.576 . Can the line be used for prediction? Why or why not?

If the scatter plot looks linear then, yes, the line can be used for prediction, because $r > \text{the positive critical value}$.

Example 2

Suppose you computed $r = -0.624$ with 14 data points. $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532 . Since $-0.624 < -0.532$, r is significant and the line can be used for prediction



$r = -0.624$ -0.532. Therefore, r is significant.

Try It

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

No, the line cannot be used for prediction, because $r <$ the positive critical value.

Example 3

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811 . Since $-0.811 < 0.776 < 0.811$, r is not significant, and the line should not be used for prediction.



$-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

Try It

For a given line of best fit, you compute that $r = -0.7204$ using $n = 8$ data points, and the critical value is $= 0.707$. Can the line be used for prediction? Why or why not?

Yes, the line can be used for prediction, because $r <$ the negative critical value.

Example 4

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = -0.567$ and the sample size, n , is 19. The $df = n - 2 = 17$. The critical value is -0.456 . $-0.567 < -0.456$ so r is significant.
 2. $r = 0.708$ and the sample size, n , is nine. The $df = n - 2 = 7$. The critical value is 0.666 . $0.708 > 0.666$ so r is significant.
 3. $r = 0.134$ and the sample size, n , is 14. The $df = 14 - 2 = 12$. The critical value is 0.532 . 0.134 is between -0.532 and 0.532 so r is not significant.
 4. $r = 0$ and the sample size, n , is five. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.
-

Try It

For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

No, the line cannot be used for prediction no matter what the sample size is.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between

x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

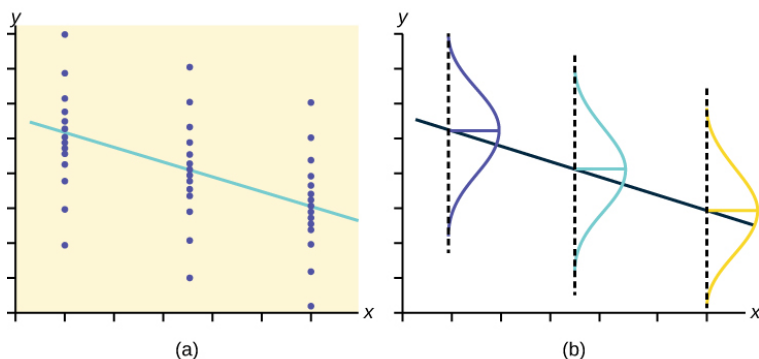
The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x . In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the

line are equal for each value of x . In other words, each of these normal distributions of y values has the same shape and spread about the line.

- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.



The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Concept Review

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear:** In the population, there is a linear relationship that models the average value of y for different values of x .
- **Independent:** The residuals are assumed to be independent.
- **Normal:** The y values are distributed normally for any value of x .
- **Equal variance:** The standard deviation of the y values is equal

for each x value.

- **Random:** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y , σ , use the standard deviation of the residuals, s .

$s = \sqrt{\frac{SSE}{n - 2}}$ The variable ρ (rho) is the population correlation coefficient.

To test the null hypothesis $H_0: \rho = \text{hypothesized value}$, use a linear regression t-test. The most common null hypothesis is $H_0: \rho = 0$ which indicates there is no linear relationship between x and y in the population.

The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit: $\hat{y} = a + bx$

where a = y -intercept, b = slope

Standard deviation of the residuals:

$$s = \sqrt{\frac{SSE}{n - 2}}$$

where

SSE = sum of squared errors

n = the number of data points