

1.1

Friday, September 30, 2016 12:32 PM

1.1 Objective 1 - What is Statistics

September 29, 2016 07:16 AM

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

We must report a measure of our confidence in our results because we do not have 100% certainty our answers are correct.

The information referred to in the definition above is *data*. **Data** are a "fact or proposition used to draw a conclusion or make a decision." Data describe characteristics of an individual.

Key point: Data vary.

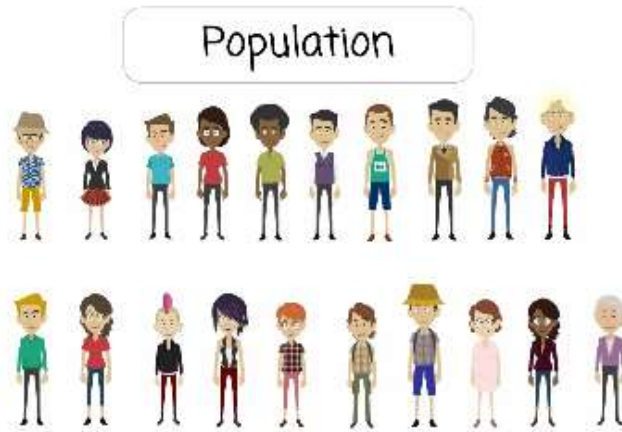
In fact, data vary when measured on ourselves as well. Do you sleep the same number of hours every night? No! Do you consume the same number of calories every day? No!

Statistics is not a math class.

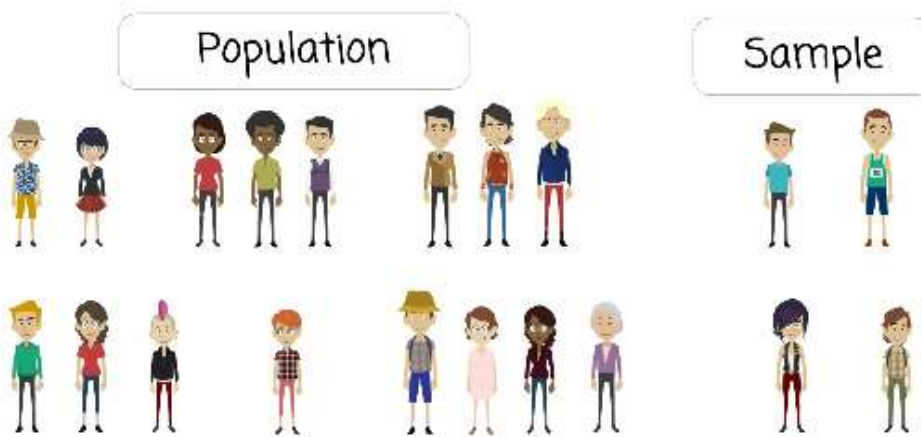
In statistics, the same approach to solving a problem can still lead to different results. This does not happen in a math class.

1.1 Objective 2 - Explain the Process of Statistics

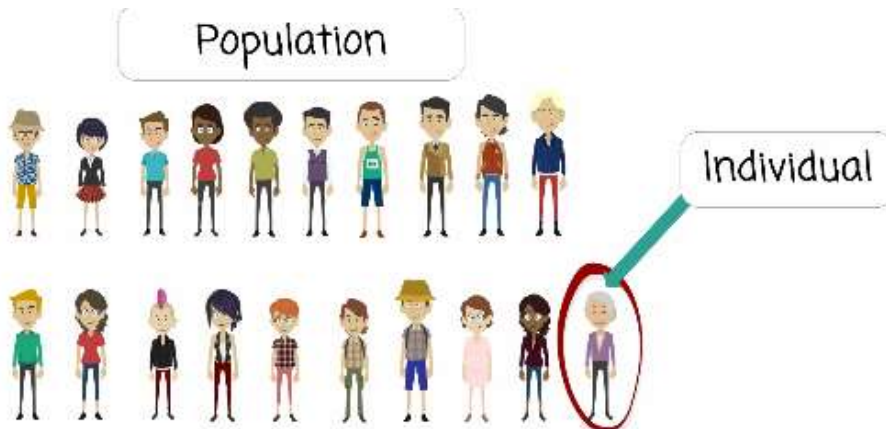
September 29, 2016 07:38 AM



The entire group to be studied is called the population.



A sample is a subset of the population being studied.



An individual is a person or object that is a member of the population being studied.

A statistic is a numerical summary of a sample.

Descriptive statistics consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.

Inferential statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

Level of Confidence



We are 95% confident that between 76% and 84% of all students would return the money.

The Process of Statistics

Step 1: Identify the Research Objective. A researcher must determine the question or questions he or she wants answered. The question or questions must be detailed so that they identify the population that is to be studied.

Step 2: Collect the Data Needed to Answer the Question(s) Posed in (Step 1). Conducting research on an entire population is often difficult and expensive, so we typically look at a sample. This step is vital to the statistical process because if the data are not collected correctly, the conclusions drawn are meaningless. Do not overlook the importance of appropriate data collection.



Step 3: Describe the Data. Descriptive statistics allow the researcher to obtain an overview of the data and can help determine the type of statistical methods the researcher should use.

Step 4: Perform Inference. Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results.

1.1 Objective 3 - Qualitative and Quantitative Variables

September 29, 2016 07:59 AM

Vocabulary and Concepts

The characteristics of the **individuals** in a study are **variables**.

Variables vary, meaning that a variable can take on different values.

If variables did not vary, then they would be constants and **inferential statistics** would not be necessary.

One goal of research is to learn the causes of variability.

Variables can be classified into two groups: qualitative and quantitative.

Qualitative, or **categorical**, **variables** allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Caution!

A numeric value does not automatically suggest a variable is quantitative.

Caution!

A numeric value does not automatically a variable is quantitative.

1.1 Objective 4 - Discrete and Continuous Variables

September 29, 2016 08:00 AM

A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. A discrete variable cannot take on every possible value between any two possible values.

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

Figure 1 illustrates the relationship among qualitative, quantitative, discrete, and continuous variables.

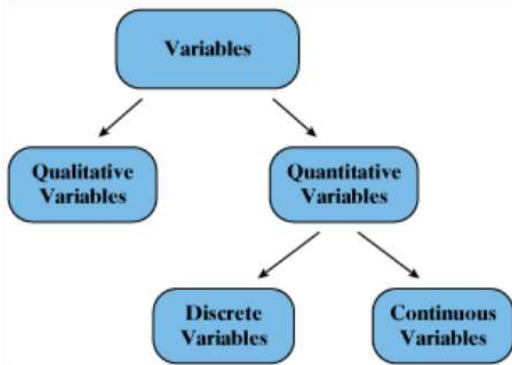


Figure 1

Vocabulary

The list of observed values for a variable is **data**.

Qualitative data are observations corresponding to a qualitative variable.

Quantitative data are observations corresponding to a quantitative variable.

Discrete data are observations corresponding to a discrete variable.

Continuous data are observations corresponding to a continuous variable.

1.1 Objective 5 - Level of Measurement of a Variable

September 29, 2016 08:09 AM

DEFINITIONS

A variable is at the **nominal level of measurement** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement. However, the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **interval level of measurement** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on the values of the variable.

A variable is at the **ratio level of measurement** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

1.2

Friday, September 30, 2016 12:32 PM

1.2 Objective 1- Distinguish between and Observational Study and a Designed Experiment

Friday, September 30, 2016 11:58 AM

Distinguish between an Observational Study and an Experiment

There are two methods for collecting data:

Observational Studies

Designed Experiments

Study 1

Cellular Phones and Brain Tumors

Researcher Elisabeth Cardis and her colleagues wanted “to determine whether mobile phone use increases the risk of [brain] tumors.”

(Source: Elisabeth Cardis et al. “Brain Tumour Risk in Relation to Mobile Telephone Use,” International Journal of Epidemiology 2010: 1–20)

Cellular Phones and Brain Tumors

To do so, the researchers identified 5117 individuals from 13 countries who were 30–59 years of age who had brain tumors diagnosed between 2000 and 2004 and matched them with 5634 individuals who did not have brain tumors. The matching was based on age, gender, and region of residence.

Cellular Phones and Brain Tumors

Both the individuals with tumors and the matched individuals were interviewed to learn about past mobile phone use, as well as sociodemographic background, medical history, and smoking status.

Cellular Phones and Brain Tumors

The researchers found no significant difference in cell phone use between the two groups. The researchers concluded there is “no increased risk of brain tumors observed in association with use of mobile phones.”

Study 2

Cellular Phones and Brain Tumors

Researchers Joseph L. Roti Roti and associates examined “whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analog cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats.”

Cellular Phones and Brain Tumors

To do so, the researchers randomly divided 480 rats into three groups.

The rats in group 1 were exposed to the analog cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 served as controls and received no radiation.

The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure.

Differences between the Two Studies

In study 1, no attempt was made to influence the individuals in the study.

The researchers simply interviewed people to determine their historical use of cell phones.

No attempt was made to influence the value of the explanatory variable, radio-frequency exposure.



An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables.

Differences between the Two Studies

In study 2, the researchers obtained 480 rats and divided the rats into three groups. Each group was *intentionally* exposed to various levels of radiation. The researchers then compared the number of rats that had brain tumors.

Clearly, there was an attempt to influence the individuals in this study because the value of the explanatory variable (exposure to radio frequency) was influenced.

If a researcher assigns the individuals in a study to a certain group, intentionally changes the value of an explanatory variable, and then records the value of the response variable for each group, the study is a **designed experiment**.

Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable.

You might wonder, why we should ever conduct an observational study if we cannot claim causation? Often, it is unethical to conduct a designed experiment.

Consider the link between smoking and lung cancer. In a designed experiment (on humans) to determine if smoking causes lung cancer, a researcher would divide a group of volunteers into two groups—Group 1 would smoke a pack of cigarettes every day for the next 10 years, and Group 2 would not smoke. Eating habits, sleeping habits, and exercise would be controlled so that the only

difference between the two groups would be smoking. After 10 years, the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group with the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes lung cancer. This designed experiment controls many potential cancer-causing factors that would not be controlled in an observational study. However, it is an unethical experiment. Do you see why?

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "Observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." *From Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized Controlled Trials." New England Journal of Medicine 342:1878-1886, 2000*

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online section to the textbook section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

DEFINITION

A **confounding variable** is an explanatory variable that was considered in the study whose effect cannot be distinguished from a second explanatory variable in the study.

The Difference between Lurking Variables and Confounding Variables

The big difference between lurking variables and confounding variables is that ***lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).***

Lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory variable and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable in a study does not necessarily have any association with the other explanatory variable, but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.2 Objective 2 - Explain the Various Types of Observational Studies

Friday, September 30, 2016 12:31 PM

Cross-sectional Studies Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.

Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals that have certain characteristics are matched with those that do not.

Cohort Studies A cohort study first identifies a group of individuals to participate in the study (cohort). The cohort is then observed over a period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are **prospective**.

Some Concluding Remarks about Observational Studies Versus Designed Experiments

Is a designed experiment superior to an observational study? Not necessarily.

- Because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designed experiments.
- It is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high-tension wires and leukemia (on humans). Do you see why?

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Have you heard this saying? "*There is no point in reinventing the wheel.*"

Here is how it applies to statistics: There is no sense expending energy to obtain data that already exist. If a researcher wants to conduct a study and appropriate data set exists, it would be silly to collect the data from scratch.

For example, various federal agencies regularly collect data that are available to the public. Some of these agencies include the Centers for Disease Control and Prevention (www.cdc.gov), the Internal Revenue Service (www.irs.gov), and the Department of Justice (<http://fjsrc.urban.org/index.cfm>). Another useful source of data is the General Social Survey (GSS), [www.gss.norc.org](http://www.gss.norc.uchicago.edu), administered by the University of Chicago. This survey regularly asks "demographic and attitudinal questions" of individuals around the country.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Another Source of Data: The Census

The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose primary residence is within the U.S. borders must fill out a questionnaire packet.

The cost of obtaining the census in 2010 was approximately \$5.4 billion; about 635,000 temporary workers were hired to assist in collecting the data.

Why is the U.S. Census so important? The results of the census are used to determine the number of representatives in each state in the House of Representatives, boundaries of congressional districts, distribution of funds for government programs (such as Medicaid), and planning for the construction of roads and schools. The first U.S. Census was conducted in 1790 under the direction of Thomas Jefferson. It is a constitutional mandate that a census be conducted every 10 years.

Is the United States successful in obtaining a census? Not entirely. Some individuals go uncounted due to illiteracy, language issues, and homelessness. Given the political stakes that are based on the census, politicians often consider how to count these individuals. Statisticians have offered solutions to the counting problem. If you wish, go to www.census.gov; in the search box, type *count homeless*. You will find many articles on the U.S. Census Bureau's attempt to count the homeless. The bottom line is that even census data has flaws.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

DEFINITION

A **census** is a list of individuals in a population along with certain characteristics of each individual.

1.3 Introduction

October 1, 2016 08:30 AM

Sampling

Observational studies can be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted. For example, the Gallup Organization regularly surveys Americans about various pop-culture and political issues. Often, the population of interest is Americans aged 18 years or older. Of course, the Gallup Organization cannot survey *all* adult Americans (there are over 200 million); instead, the group typically surveys a *random sample* of about 1000 adult Americans.

DEFINITION

Random sampling is the process of using chance to select individuals from a population to be included in the sample.

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let *chance* or *randomness*, rather than convenience, play a role in dictating which individuals are in the sample. **If convenience is used to obtain a sample, the results of the survey are meaningless.**

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Recognizing a Convenience Sample and Its Limitations

Suppose that Gallup wants to know the proportion of adult Americans who consider themselves to be baseball fans. If Gallup obtained a sample by standing outside Fenway Park (home of the Boston Red Sox professional baseball team), the survey results are not likely to be reliable. Why? Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population.

Suppose you wanted to learn the proportion of students on your campus who work. It might be convenient to survey the students in your statistics class, but do these students represent the overall student body? Does the proportion of freshmen, sophomores, juniors, and seniors in your class mirror the proportion of freshmen, sophomores, juniors, and seniors on campus? Does the proportion of males and females in your class resemble the proportion of males and females across campus? Probably not. What about evening (or day) students? For these (and many other) reasons, the convenient sample is not representative of the population, which means that any results reported from your survey are misleading.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Effective Sampling Techniques

We will discuss four basic sampling techniques:

1. Simple random sampling
2. Stratified sampling
3. Systematic sampling
4. Cluster sampling

These sampling methods are designed so that any selection biases the surveyor introduced (knowingly or unknowingly) during the selection process are eliminated. In other words, the surveyor does not have a choice as to which individuals are in the study. We will discuss simple random sampling in this section and the remaining three types of sampling in the next section.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.3 Objective 1 - Obtain a Simple Random Sample

October 1, 2016 08:34 AM

DEFINITION

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equal chance of occurring. The sample is then called a **simple random sample**.

The number of individuals in the sample is always less than the number of individuals in the population. That is, $n < N$.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Illustrating Simple Random Sampling

Sophie has four tickets to a concert. Six of her friends, Yolanda, Michael, Kevin, Marissa, Annie, and Katie, have all expressed an interest in going to the concert. Sophie decides to randomly select three of her six friends to attend the concert.

Example – Illustrating Simple Random Sampling

- (a) List all possible samples of size $n = 3$ from the population of size $N = 6$. Once an individual is chosen, he/she cannot be chosen again.
- (b) Comment on the likelihood of the sample containing Michael, Kevin, and Marissa.

Lower case n = Sample Size
Uppercase N = Population Size

Solution

(a) List all possible samples of size $n = 3$ from the population of size $N = 6$. Once an individual is chosen, he/she cannot be chosen again.

Let 1 represent Yolanda, 2 represent Michael,
3 represent Kevin, 4 represent Marissa,
5 represent Annie, and 6 represent Katie.

(a) List all possible samples of size $n = 3$ from the population of size $N = 6$.

- Yolanda, Michael, Kevin
- Yolanda, Michael, Marissa
- Yolanda, Michael, Annie
- Yolanda, Michael, Katie
- Yolanda, Kevin, Marissa
- Yolanda, Kevin, Annie

- Michael, Marissa, Annie
- Michael, Marissa, Katie
- Michael, Annie, Katie
- Kevin, Marissa, Annie
- Kevin, Marissa, Katie
- Kevin, Katie, Annie
- Marissa, Annie, Katie

(b) Comment on the likelihood of the sample containing Michael, Kevin, and Marissa.

There is a 1 in 20 chance that the simple random sample will contain these 3 friends.

How do we select the individuals in a simple random sample?

We could write the names of the individuals in the population on different pieces of paper and then select names from a hat. Often, however, the size of the population is so large that performing simple random sampling in this fashion is not practical.

Typically, each individual in the population is assigned a unique number between 1 and N , where N is the size of the population. Then n distinct random numbers are selected, where n is the size of the sample.

To number the individuals in the population, we need a **frame**—a list of all the individuals within the population.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

StatCrunch Steps for Simple Random Sample

StatCrunch Steps

- Data > Simulate > Discrete Uniform
- Enter the number of values in Rows and the number of samples in Columns.
- Enter Minimum and Maximum values.
- Select a seed, or select the single dynamic seed option.
- Click Compute.

1.4 Other Effective Sampling Methods

October 1, 2016 09:26 AM

The goal of sampling is to collect as much information as possible about the population at the least cost. Cost includes monetary outlays, time, and other resources. With this goal in mind, we may find it advantageous to use sampling techniques other than simple random sampling.

Here we cover three other sampling techniques:

1. Stratified sampling
2. Systematic sampling
3. Cluster sampling

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Under certain circumstances, a *stratified sample* provides more information about the population for less cost compared with simple random sampling.

DEFINITION

A **stratified sample** is obtained by dividing the population into non-overlapping groups called strata and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogenous (similar) in some way.

For example, suppose Congress was considering a bill that abolishes estate taxes. In an effort to determine the opinion of her constituency, a senator asks a pollster to conduct a survey within her state.

The pollster may divide the population of registered voters within the state into three strata: Republican, Democrat, and Independent. This grouping makes sense because the members within each of the three parties may have similar opinions regarding estate taxes, but opinions among parties may differ. The main criterion in performing a stratified sample is that each group (stratum) must have a common attribute that results in the individuals being similar within the stratum.

An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to be surveyed while it obtains the same or more information. This result occurs because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next. In addition, a stratified sample guarantees that each stratum is represented in the sample.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.4 Objective 1 - Obtaining a Stratified Sample

October 2, 2016 08:16 AM

Example – Obtaining a Stratified Sample

The president of DePaul University wants to conduct a survey to determine the community's opinion regarding campus safety.

The president divides the DePaul community into three groups: resident students, nonresident (commuting) students, and staff (including faculty) so that he can obtain a stratified sample.

Suppose there are 6204 resident students, 13,304 nonresident students, and 2401 staff, for a total of 21,909 individuals in the population.

- Resident students: $6204/21,909 = 28\%$
- Nonresident students: $13,304/21,909 = 61\%$
- Staff: $2401/21,909 = 11\%$

The president wants to obtain a sample of size 100, with the number of individuals selected from each stratum weighted by the population size.

A sample of size 100 requires a stratified sample of

- $0.28(100)=28$ resident students
- $0.61(100)=61$ nonresident students
- $0.11(100)=11$ staff

Solution

To obtain the stratified sample, construct a simple random sample within each group.

- Resident students: 28 out of 6204
- Nonresident students: 61 out of 13,304
- Staff: 11 out of 2401

Caution

Do not use the same seed for all the groups in a stratified sample, because we want the simple random samples within each stratum to be independent of each other.

Samples

Use the first 28 nonrepeating values in the Resident column, the first 61 nonrepeating values in the Nonresident column, and the first 11 nonrepeating values in the Staff column.

Advantage of Stratified Sampling

The researcher can determine characteristics within each stratum.

1.4 Objective 2 - Obtain a Systematic Sample

October 2, 2016 08:28 AM

In both simple random sampling and stratified sampling, a [frame](#) must exist. Therefore, these sampling techniques require some preliminary work before the sample can be found. A sampling technique that does not require a frame is *systematic sampling*.

DEFINITION

A **systematic sample** is obtained by selecting every k 'th individual from the population. The first individual selected corresponds to a number between 1 and k .

For example, to learn about the outcome of an election, a pollster might survey every tenth individual that leaves a polling place.

Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population. Also, systematic samples typically provide more information for a given cost than does simple random sampling. In addition, systematic sampling is easier to employ; so there is less likelihood of interviewer error occurring, such as selecting the wrong individual to be surveyed.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Obtaining a Systematic Sample without a Frame

The manager of Kroger Food Stores wants to measure the satisfaction of the store's customers. Design a sampling technique that can be used to obtain a sample of 40 customers.

Solution

The manager decides to obtain a systematic sample by surveying every 7th customer.

To determine which customer will be the first customer surveyed, the manager needs to randomly choose a customer between the 1st and 7th customer.

The survey will include customers ...

$$3, 10, 17, 24, \dots, 276$$

↑
 $3 + 39(7)$

Choosing a Value for k

When using systematic sampling, how would we select the value of k ?

If the size of the population is unknown, there is no mathematical way to determine k . The value of k must be small enough to achieve our desired sample size and large enough to obtain a sample that is representative of the population. The following video illustrates the importance of k .

From <https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>

For Example 2, suppose $k = 30$.

$$3, 33, 63, \dots, 1173$$

$39 \cdot 30 + 3$

If Kroger does not have 1173 shoppers, the desired sample size will not be achieved.

For Example 2, suppose $k = 4$.

$$3, 7, 11, \dots, 159$$

$39 \cdot 4 + 3$

The 159th shopper might leave the store at 3pm, so our survey would not include any of the evening shoppers.

An estimate of the size of the population would help to determine an appropriate value of k .

Systematic Sampling

Determining k when N is Known

If possible, approximate the population size N .

Determine the sample size desired, n .

Divide N by n and round down to the nearest integer. This value is k .

Suppose $N = 20,325$ and we desire a sample size of $n = 100$.

$$\frac{20,325}{100} = 203.25 \rightarrow k = 203$$

If you start at 90

$$90, 293, 496, \dots, 20,187$$

$$99 \cdot 203 + 90$$

Steps in Systematic Sampling

Step 1 If possible, approximate the population size, N .

Step 2 Determine the sample size, n .

Step 3 Compute $\frac{N}{n}$ and round down to the nearest integer. This value is k .

Step 4 Randomly select a number between 1 and k . Call this number p .

Step 5 The sample will consist of the following individuals:

$$p, p + k, p + 2k, \dots, p + (n - 1)k$$

Step 1:

$$p, p + k,$$

Step 2:

$$p + 2k, \dots$$

Step 3:

$$p + (n - 1)k$$

1.4 Objective 3 - Obtain a Cluster Sample

October 2, 2016 08:55 AM

A fourth sampling method is called *cluster sampling*. Like the previous three sampling methods, this method has benefits under certain circumstances.

DEFINITION

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

Suppose a school administrator wants to learn the characteristics of students enrolled in online classes. Rather than obtaining a simple random sample based on the frame of all students enrolled in online classes, the administrator treats each online class as a cluster and then finds a simple random sample of these clusters. The administrator then surveys *all* students in the selected clusters. This reduces the number of classes that get surveyed.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Obtaining a Cluster Sample

A sociologist wants to gather data regarding household income within the city of Boston. Obtain a sample using cluster sampling.

Solution

The city of Boston can be set up so that each city block is a cluster. Once the city blocks have been identified, we obtain a simple random sample of the city blocks and survey all households on the blocks selected.

Suppose there are 10,493 city blocks in Boston. First, the sociologist must number the blocks from 1 to 10,493.

Suppose the sociologist only has enough time and money to survey 20 clusters (city blocks).

Selected Blocks

3951	6676	8408	3462	10321
2532	5585	8198	8500	4025
1682	6633	4528	9887	5709
6917	7919	8200	2685	8142

Advantages of Cluster Sampling

Reduces travel time that would likely be required with stratified sampling or simple random sampling.

No need to obtain a frame of all the households.

Issues to Consider in Cluster Sampling

The following questions arise in cluster sampling:

- How do I cluster the population?
- How many clusters do I sample?
- How many individuals should be in each cluster?

First, we must determine whether the individuals within the proposed cluster are *homogeneous* (similar individuals) or *heterogeneous* (dissimilar individuals).

In Example 3, city blocks tend to have similar households. Survey responses from houses on the same city block are likely to be similar. This results in duplicate information. We conclude that **if the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.**

What if the cluster is heterogeneous? Under this circumstance, the heterogeneity of the cluster likely resembles the heterogeneity of the population. In other words, each cluster is a scaled-down representation of the overall population.

For example, a quality control manager might use shipping boxes that contain 100 lightbulbs as a cluster. The manager does this because the rate of defects within the cluster resembles the rate of defects in the population, assuming that the bulbs are randomly placed in the box. Thus, **when each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.**

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Convenience Sampling

In the four sampling techniques just presented (simple random sampling, stratified sampling, systematic sampling, and cluster sampling), the individuals are selected randomly. Often, however, inappropriate

sampling methods are used in which the individuals are *not* randomly selected.

Have you ever been stopped in the mall by someone holding a clipboard? These folks are responsible for gathering information, but their methods of data collection are inappropriate, and the results of their analysis are suspect because they collect data using a *convenience sample*.

DEFINITION

*In a **convenience sample** the individuals are easily obtained and not based on randomness.*

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Examples of Convenience Samples

The most popular convenience samples are those in which the individuals in the sample are **self-selected**, meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called **voluntary response** samples. One example of self-selected sampling is phone-in polling—a radio personality will ask his or her listeners to phone or text the station to submit their opinions. Another example is the use of the Internet to conduct surveys. For example, a TV news show will present a story regarding a certain topic and ask its viewers to "tell us what you think" by completing an online questionnaire or tweeting an opinion with a hashtag.

Both of these samples are poor designs because the individuals who decide to be in the sample generally have strong opinions about the topic. A more typical individual in the population will not bother phoning, texting, or tweeting to complete a survey. Any inference made regarding the population from this type of sample should be made with extreme caution.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Multistage Sampling

In practice, most large-scale surveys obtain samples using a combination of the techniques just presented.

As an example of multistage sampling, consider Nielsen Media Research. Nielsen randomly selects households and, through a People Meter, monitors the television programs these households are watching. The meter is an electronic box connected to each TV within the household. The People Meter measures what program is being watched and who is watching it. Nielsen selects the households with the use of a two-stage sampling process.

Stage 1: Using U.S. Census data, Nielsen divides the country into geographic areas (strata). The strata are typically city blocks in urban areas and geographic regions in rural areas. About 6000 strata are randomly selected.

Stage 2: Nielsen sends representatives to the selected strata and lists households within the strata. The households are then randomly selected through a simple random sample.

Nielsen sells the information obtained to television stations and companies. These results are used to help determine prices for commercials.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

An Example of Multistage Sampling

Consider the sample used by the Census Bureau for the Current Population Survey. This survey requires five stages of sampling:

Stage 1: Stratified sample

Stage 2: Cluster sample

Stage 3: Stratified sample

Stage 4: Cluster sample

Stage 5: Systematic sample

This survey is very important because it is used to obtain demographic estimates of the United States in noncensus years. Details about the Census Bureau's sampling methods can be found in *The Current Population Survey: Design and Methodology*, Technical Paper No. 40.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Sample Size Considerations

Throughout our discussion of sampling, we did not mention how to determine the sample size. Determining the sample size is key in the overall statistical process. Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error.

Researchers must find a balance between the reliability of the results and the cost of obtaining these results. The bottom line is that time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference will be.

Later in the course, we will discuss techniques for determining the sample size required to estimate characteristics regarding the population within some margin of error. (For a detailed discussion of sample size considerations, consult a text on sampling techniques such as *Elements of Sampling Theory and Methods* by Z. Govindarajulu, Pearson, 1999.)

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.5 Bias in Sampling

October 3, 2016 12:27 PM

So far we have looked at *how* to gather samples, but not at some of the problems that inevitably arise in sampling. Remember, the goal of sampling is to collect information about a population through a sample.

DEFINITION

If the results of the sample are not representative of the population, then the sample has **bias**.

There are three sources of bias in sampling:

1. Sampling bias
2. Nonresponse bias
3. Response bias

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.5 Objective 1 - Explain the Sources of Bias in Sampling

October 3, 2016 12:31 PM

Sampling Bias

Sampling bias means that the technique used to obtain the sample's individuals tends to favor one part of the population over another.

Any convenience sample has sampling bias because the individuals are not chosen through a random sample.

Undercoverage

Sampling bias also results due to **undercoverage**, which occurs when the proportion of one segment of the population is lower in a sample than it is in the population.

Undercoverage can result if the frame used to obtain the sample is incomplete or not representative of the population.

Sampling Bias Can Lead to Incorrect Predictions

The magazine *Literary Digest* predicted that Alfred M. Landon would defeat Franklin D. Roosevelt in the 1936 presidential election.

The *Literary Digest* conducted a poll based on a list of its subscribers, telephone directories, and automobile owners.

On the basis of the results, the *Literary Digest* predicted that Landon would win with 57% of the popular vote.

However, Roosevelt won the election with about 62% of the popular vote.

Gaining Access to a Complete List of Individuals in a Population

Public-opinion polls often use random telephone surveys, which implies that the frame is all households with telephones.

This method of sampling excludes households without telephones, as well as homeless people.

Nonresponse Bias

Nonresponse bias exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do.

Nonresponse can occur because individuals selected for the sample do not wish to respond or the interviewer was unable to contact them.

Nonresponse Bias

The federal government's Current Population Survey has a response rate of about 92%, but it varies depending on the age of the individual.

For example, the response rate for 20- to 29-year-olds is 85%, and for individuals 70 and older, it is 99%.

Response rates in random digit dialing (RDD) telephone surveys are typically around 70%.

E-mail survey response rates hover around 40%.

Mail surveys can have response rates as high as 60%.

Nonresponse Bias Can Be Controlled Using Callbacks

If a mailed questionnaire was not returned, a callback might mean phoning the individual to conduct the survey.

If an individual was not at home, a callback might mean returning to the home at other times in the day.

Using Rewards or Incentives

I received \$1 with a survey regarding my satisfaction with a recent purchase.

The \$1 "payment" was meant to make me feel guilty enough to fill out the questionnaire.

As another example, a city may send out questionnaires to households and state in a cover letter that the responses to the questionnaire will be used to decide pending issues within the city.

Literary Digest Poll

The *Literary Digest* mailed out more than 10 million questionnaires and 2.3 million people responded.

The rather low response rate (23%) contributed to the incorrect prediction.

After all, Roosevelt was the incumbent president and only those who were unhappy with his administration were likely to respond.

Response Bias

Response bias exists when the answers on a survey do not reflect the true feelings of the respondent.

Interviewer Error

A trained interviewer is essential to obtain accurate information from a survey.

A skilled interviewer can elicit responses from individuals and make the interviewee feel comfortable enough to give truthful responses.

For example, a good interviewer can obtain truthful answers to questions as sensitive as “Have you ever cheated on your taxes?”

Interviewer Error

Do not be quick to trust surveys conducted by poorly trained interviewers.

Do not trust survey results if the sponsor has a vested interest in the results of the survey.

Would you trust a survey conducted by a car dealer that reports 90% of customers say they would buy another car from the dealer?

Misrepresented Errors

Some survey questions result in responses that misrepresent facts or are flat-out lies.

For example, a survey of recent college graduates may find that self-reported salaries are inflated.

Also, people may overestimate their abilities.

Wording of Questions

The way a question is worded can lead to response bias in a survey, so questions must always be asked in balanced form.

For example, “Do you oppose the reduction of estate taxes?” should be written “Do you favor or oppose the reduction of estate taxes?”

Wording of Questions

Consider the following report based on studies from Schuman and Presser (*Questions and Answers in Attitude Surveys*, 1981, p. 277), who asked the following two questions:

(A) Do you think the United States should forbid public speeches against democracy?

(B) Do you think the United States should allow public speeches against democracy?

Wording of Questions

For those respondents presented with question A, 21.4% gave “yes” responses, while for those given question B, 47.8% gave “no” responses.

Wording of Questions

Another consideration in wording a question is not to be vague.

The question “How much do you study?” is too vague.

Does the researcher mean how much do I study for all my classes or just for statistics? Does the researcher mean per day or per week?

The question should be written “How many hours do you study statistics each week?”

Ordering of Questions or Words

Many surveys will rearrange the order of the questions within a questionnaire so that responses are not affected by prior questions.

Ordering of Questions or Words

(A) Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?

(B) Do you think a Communist country such as Russia should let American newspaper reporters come in and send back to America the news as they see it?

Ordering of Questions or Words

For surveys conducted in 1980 in which the questions appeared in the order (A, B), 54.7% of respondents answered “yes” to A and 63.7% answered “yes” to B.

If the questions were ordered (B, A), then 74.6% answered “yes” to A and 81.9% answered “yes” to B.

Ordering of Questions or Words

For example, the Gallup Organization routinely asks the following question of 1017 adults aged 18 years or older:

Do you [rotated: approve (or) disapprove] of the job Barack Obama is doing as president?

Type of Question

One of the first considerations in designing a question is determining whether the question should be *open* or *closed*.

Open Question

An **open question** allows the respondent to choose his or her response:

What is the most important problem facing America's youth today?

Closed Question

A **closed question** requires the respondent to choose from a list of predetermined responses:

What is the most important problem facing America's youth today?

- (a) Drugs
- (b) Violence
- (c) Single-parent homes
- (d) Promiscuity
- (e) Peer pressure

Open Questions

An open question should be phrased so that the responses are similar. (You don't want a wide variety of responses.)

Closed Questions

Closed questions limit the number of respondent choices and, therefore, the results are much easier to analyze.

The limited choices, however, do not always include a respondent's desired choice.

Type of Question

Survey designers recommend conducting pretest surveys with open questions and then using the most popular answers as the choices on closed-question surveys.

Another issue to consider in the closed-question design is the number of possible responses.

The option "no opinion" should be omitted, because this option does not allow for meaningful analysis.

Data Entry Error

Although not technically a result of response bias, data-entry error will lead to results that are not representative of the population.

Can a Census Have Bias?

Yes!

The discussion so far has focused on bias in samples, but bias can also occur when conducting a census.

How?

A question on a census form could be misunderstood, thereby leading to response bias in the results.

We also mentioned that it is often difficult to contact each individual in a population. For example, the U.S. Census Bureau faces challenges in counting each homeless person in the country, so the census data published by the U.S. government likely suffers from nonresponse bias.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Sampling Error versus Nonsampling Error

Nonresponse bias, response bias, and data-entry errors are types of *nonsampling error*.

However, when a sample is used to learn information about a population, *sampling error* is also likely to occur.

DEFINITION

Nonsampling errors result from under-coverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a census.

Sampling error results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

In Other Words

When sampling results in incomplete information, we mean that the individuals in the sample cannot reveal all the information about the population. Suppose we want to determine the average age of students enrolled in an introductory statistics course. To do this, we find a simple random sample of four students and ask them to write their age on a sheet of paper and turn it in. The average age of these four students is 23.3 years. Assume that no students lied about their age or misunderstood the question and the sampling was done appropriately. If the actual average age of all 30 students in the class (the population) is 22.9 years, then the sampling error is $23.3 - 22.9 = 0.4$ year. Now suppose the same survey is conducted again, but this time one student lies about his age. The results of that survey will also have non-sampling error.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.6 - The Design of Experiments

October 3, 2016 03:27 PM

Three Observational Studies

1. Cross-sectional
2. Case-control (retrospective)
3. Cohort (prospective)

In observational studies, we cannot make statements of *causality* between the explanatory variable(s) and the response variable.

In a smoking observational study:

Response: Contracts cancer or not (qualitative)
Explanatory: Smoker or not (qualitative)

Explanatory and response variables can be qualitative or quantitative. The type of explanatory/response variables in the study determine the type of inference we perform.

1.6 Objective 1 - Describe the characteristics of an experiment

October 4, 2016 01:04 PM

An **experiment** is a controlled study conducted to determine the effect of varying one or more explanatory variables or **factors** has on a response variable. Any combination of the values of the factors is called a **treatment**.

Two factors combined into four levels of a treatment:

Nonsmoker, Low Protein
Nonsmoker, High Protein
Smoker, Low Protein
Smoker, High Protein

The **experimental unit** (or **subject**) is a person, object or some other well-defined item upon which a treatment is applied.

A **control group** serves as a baseline treatment that can be used to compare to other treatments.

A **placebo** is an innocuous medication, such as a sugar tablet, that looks, tastes, and smells like the experimental medication.

Blinding refers to nondisclosure of the treatment an experimental unit is receiving.

A **single-blind** experiment is one in which the experimental unit (or subject) does not know which treatment he or she is receiving.

A **double-blind** experiment is one in which neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.

EXAMPLE 1 The Characteristics of an Experiment

Problem

Lipitor is a cholesterol-lowering drug made by Pfizer. In the Collaborative Atorvastatin Diabetes Study (CARDS), the effect of Lipitor on cardiovascular disease was assessed in 2838 subjects, ages 40 to 75, with type 2 diabetes, without prior history of cardiovascular disease. In this placebo-controlled, double-blind experiment, subjects were randomly allocated to either Lipitor 10 mg daily (1428 subjects) or placebo (1410 subjects) and were followed for 4 years. The response variable was whether there was an occurrence of any major cardiovascular event or not.

Lipitor significantly reduced the rate of major cardiovascular events (83 events in the Lipitor group versus 127 events in the placebo group). There were 61 deaths in the Lipitor group versus 82 deaths in the placebo group.

Video Solution



1.6 Objective 2 - Explain the steps in designing an experiment

October 4, 2016 01:04 PM

Steps in Designing an Experiment

Step 1 *Identify the Problem to Be Solved.* The statement of the problem should be as explicit as possible and should provide the experimenter with direction. The statement must also identify the response variable and the population to be studied. Often, the statement is referred to as the *claim*.

Step 2 *Determine the Factors That Affect the Response Variable.* The factors are usually identified by an expert in the field of study. In identifying the factors, ask, "What things affect the value of the response variable?" After the factors are identified, determine which factors to fix at some predetermined level, which to manipulate, and which to leave uncontrolled.

Step 3 *Determine the Number of Experimental Units.* As a general rule, choose as many experimental units as time and money allow. Techniques exist for determining sample size, provided certain information is available.

Step 4 *Determine the Level of Each Factor.* There are two ways to deal with the factors, control or randomize.

- **1. Control:** There are two ways to control the factors.
 - **(a)** Set the level of a factor at one value throughout the experiment (if you are not interested in its effect on the response variable).
 - **(b)** Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.
- **2. Randomize:** Randomly assign the experimental units to various treatment groups so that the effect of factors whose levels cannot be controlled is minimized. The idea is that randomization averages out the effects of uncontrolled factors (explanatory variables). It is difficult, if not impossible, to identify all factors in an experiment. This is why randomization is so important. It mutes the effect of variation attributable to factors not controlled.

Step 5 *Conduct the Experiment.*

- **(a)** Randomly assign the experimental units to the treatments. Replication occurs when each treatment is applied to more than one experimental unit. Using more than one experimental unit for each treatment ensures the effect of a treatment is not due to some characteristic of a single experimental unit. It is a good idea to assign an equal number of experimental units to each treatment.
- **(b)** Collect and process the data. Measure the value of the response variable for each replication. Then organize the results. The idea is that the value of the response variable for each treatment group is the same before the experiment because of randomization. Then any difference in the value of the response variable among the different treatment groups is a result of differences in the level of the treatment.

Step 6 *Test the Claim.* This is the subject of **inferential statistics**. Inferential statistics is a process in which generalizations about a population are made on the basis of results obtained from a sample. Provide a statement regarding the level of confidence in the generalization.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

To help understand the steps in designing an experiment, let's review [Example 1](#).

Step 1 Identify the Problem to Be Solved The problem to be solved is to determine whether 10 mg of Lipitor daily reduces the likelihood of having a major cardiovascular event in 40 to 75 year old subjects with type 2 diabetes.

Step 2 Determine the Factors That Affect the Response Variable Some factors that may affect whether one has a cardiovascular event are diet, exercise, family history, and level of cholesterol.

Step 3 Determine the Number of Experimental Units There were 2838 subjects in the study.

Step 4 Determine the Level of Each Factor The factor of interest is the drug, which was set at two levels: placebo and 10 mg of Lipitor. Although not stated, the researchers likely fixed the diet of the subjects and fixed an exercise regimen. Family history cannot be controlled, so the random assignment of the subjects to two groups will average out a bad family history of heart disease. For example, we would not expect all subjects with a poor history of heart health to end up in the placebo (control) group.

Step 5 Conduct the Experiment The subjects were randomly assigned to either the placebo or Lipitor group. There were 2838 replications of the experiment.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.6 Objective 3 - Explain the completely randomized design

October 4, 2016 01:05 PM

We now concentrate on the simplest type of experiment.

Definition

A **completely randomized design** is one in which each experimental unit is randomly assigned to a treatment.

The study from Example 1 is a completely randomized design.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Problem

A farmer wishes to determine the optimal level of a new fertilizer on his soybean crop. Design an experiment that will assist him.

Approach

Follow the [steps](#) for designing an experiment.

Solution

Step 1 The farmer wants to identify the optimal level of fertilizer for growing soybeans. We define *optimal* as the level that maximizes yield. **So the response variable will be crop yield.**

Step 2 Some **factors** that affect crop yield are **fertilizer, precipitation, sunlight, method of tilling the soil, type of soil, plant, and temperature.**

Step 3 In this experiment, we will plant 60 soybean plants (experimental units).

Step 4 We list the factors and their levels.

- **Fertilizer.** This factor is the explanatory variable of interest. So, it will be controlled and set at three levels. We wish to measure the effect of varying the level of this variable on the response variable, yield. We will set the treatments (level of fertilizer) as follows:
 - Treatment A: 20 soybean plants receive no fertilizer.
 - Treatment B: 20 soybean plants receive 2 teaspoons of fertilizer per gallon of water every 2 weeks.
 - Treatment C: 20 soybean plants receive 4 teaspoons of fertilizer per gallon of water every 2 weeks.
- **Precipitation.** We cannot control the amount of rainfall, but we can control the amount of watering we do, so that each plant receives the same amount of precipitation.
- **Sunlight.** This uncontrollable factor will be roughly the same for each plant.
- **Method of tilling.** We can control this factor and will use the round-up ready method of tilling for each plant.
- **Type of soil.** We can control certain aspects of the soil such as level of acidity. In addition, each plant will be planted within a 1-acre area, so it is reasonable to assume that the soil conditions for each plant are equivalent.
- **Plant.** There may be variation (in terms of ability to generate yield) from plant to plant. To account for this, we randomly assign the plants to a treatment.
- **Temperature.** This factor is uncontrollable, but will be the same for each plant.

Step 5

- **(a)** We need to assign each plant to a treatment group. First, we will number the plants from 1 to 60 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment

A. Next we number the remaining plants 1 to 40 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment B. The remaining plants get treatment C. Now we till the soil, plant the soybean plants, and fertilize according to the schedule prescribed.

- **(b)** At the end of the growing season, we determine the crop yield for each plant.

Step 6 We determine any differences in yield among the three treatment groups.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Figure 2 illustrates the experimental design from Example 2 on the previous screen.



An experimental design labeled “Random assignment of plants to treatments” with 3 arrows pointing to “Group 1 receives 20 plants”, “Group 2 receives 20 plants”, and “Group 3 receives 20 plants”. Respectively, each of the above points to “Treatment A: No fertilizer”, “Treatment B: 2 teaspoons”, “Treatment C: 4 teaspoons”. The last 3 all point to “Compare yield”.

From <https://media.pearsoncmg.com/aw/aw_sullivanwoodbury_online_16/longdesc/Section01_6_45_02.html>

Figure 2

Example 2 is a completely randomized design because the experimental units (the plants) were randomly assigned to the treatments. It is the most popular experimental design because of its simplicity, but it is not always the best.

1.6 The Design of Experiments

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

1.6 Objective 4 - Explain the matched-pairs design

October 4, 2016 01:05 PM

Another type of experimental design is called a *matched-pairs design*.

Definition

A **matched-pairs design** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.

In matched-pairs design, one matched individual will receive one treatment and the other receives a different treatment. The matched pair is randomly assigned to the treatment using a coin flip or a random-number generator. We then look at the difference in the results of each matched pair. One common type of matched-pairs design is to measure a response variable on an experimental unit before and after a treatment is applied. In this case, the individual is matched against itself. These experiments are sometimes called before–after or pretest–posttest experiments.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Problem

An educational psychologist wants to determine whether listening to music has an effect on a student's ability to learn. Design an experiment to help the psychologist answer the question.

Approach

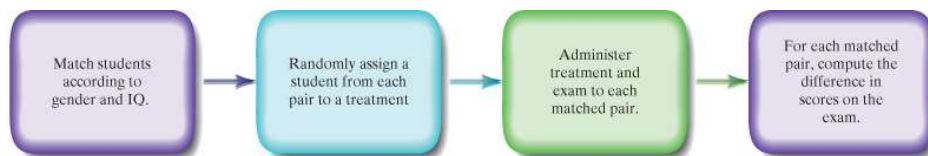
We will use a matched-pairs design by matching students according to IQ and gender (just in case gender plays a role in learning with music).

Solution

We match students according to IQ and gender. For example, we match two females with IQs in the 110 to 115

range. For each pair of students, we flip a coin to determine which student is assigned the treatment of a quiet room or a room with music playing in the background.

Each student will be given a statistics textbook and asked to study Section 1.1. After 2 hours, the students will enter a testing center and take a short quiz on the material in the section. We compute the difference in the scores of each matched pair. Any differences in scores will be attributed to the treatment. Figure 3 illustrates the design.



4 boxes each connected with arrows to the next. Boxes labeled: “Match students according to gender and IQ”, “Randomly assign a student from each pair to a treatment”, “Administer treatment and exam to each matched pair”, and “For each matched pair, compute the difference in scores on the exam”.

Figure 3

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

2.1 Organizing Qualitative Data

October 6, 2016 09:10 AM

2.1 Objective 1

October 6, 2016 09:12 AM

When [qualitative data](#) are collected, we often first determine the number of individuals within each category.

DEFINITION

A **frequency distribution** lists each category of data and the number of occurrences for each category of data.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 1 Organizing Qualitative Data into a Frequency Distribution

Problem

A physical therapist wants to determine types of rehabilitation required by her patients. To do so, she obtains a simple random sample of 30 of her patients and records the body part requiring rehabilitation. See Table 1. Construct a frequency distribution of location of injury.

Video Solution



Technology Step-By-Step



TABLE 1

Back	Back	Hand
Wrist	Back	Groin
Elbow	Back	Back
Back	Shoulder	Shoulder
Hip	Knee	Hip
Neck	Knee	Knee
Shoulder	Shoulder	Back
Back	Back	Back
Knee	Knee	Back
Hand	Back	Wrist

Data from Krystal Catton, student at Joliet Junior College

Approach

To construct a frequency distribution, create a list of the body parts (categories) and tally each occurrence. Then, add up the number of tallies (observations) to determine the frequency.

Solution

Table 2 shows that the back is the most common body part requiring rehabilitation, with a total frequency of 12.

Body Part	Tally	Frequency
Back	 	12
Wrist		2
Elbow		1
Hip		2
Shoulder		4
Knee	 	5
Hand		2
Groin		1
Neck		1



Even though the qualitative data has been counted into a frequency, it is still qualitative

In any frequency distribution, it is a good idea to add up the frequency column to make sure that it equals the number of observations.

In [Example 1](#), the frequency column totals to 30 as it should because there are 30 body parts (observations).

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Often, we want to know the *relative frequency* of the categories rather than the frequency.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Definition: Relative Frequency

DEFINITION

The **relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Sum of all frequencies}}$$

A **relative frequency distribution** lists each category of data together with the relative frequency.

A frequency distribution lists the of occurrences of each category of data, while a relative frequency distribution lists the of occurrences of each category of data.

✓ Fantastic!



A frequency distribution lists the number of occurrences of each category of data.

A relative frequency distribution lists the proportion of occurrences of each category of data.

Next Question

EXAMPLE 2 Constructing a Relative Frequency Distribution of Qualitative Data

Problem

Using the summarized data in Table 2, construct a relative frequency distribution.

Video Solution



Approach

Add all the frequencies and then use

Technology Step-By-Step



$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Sum of all frequencies}}$$

to compute the relative frequency of each category of data.

Solution

The sum of all the values in the frequency column in Table 2 is 30. We now compute the relative frequency of each category. For example, the relative frequency of the category *Back* is $12/30 = 0.4$. The relative frequencies are shown in column 3 of Table 3. From the distribution, the most common body part for rehabilitation is the back.

TABLE 3

Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} = 0.4$
Wrist	2	$\frac{2}{30} \approx 0.0667$
Elbow	1	0.0333
Hip	2	0.0667
Shoulder	4	0.1333
Knee	5	0.1667
Hand	2	0.0667
Groin	1	0.0333
Neck	1	0.0333
Total	30	1

It is a good idea to add up the relative frequencies to be sure they sum to 1. In fraction form, the sum should be exactly 1. In decimal form, the sum may differ slightly from 1 due to rounding.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

2.1 Objective 2 - Construct Bar Graphs

October 6, 2016 09:39 AM

Once raw data are organized in a table, we can create graphs. Just as "a picture is worth a thousand words," pictures of data result in a more powerful message than do tables.

Try the following exercise: Open a newspaper or news website and look at a table and a graph. Study each one. Then put the paper away and close your eyes. What do you see in your mind's eye? Can you recall information more easily from the table or the graph? In general, people are more likely to recall information obtained from a graph than from a table.

A common device for graphically representing qualitative data is a *bar graph*.

DEFINITION

A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#question3>>

EXAMPLE 3 Constructing a Frequency and Relative Frequency Bar Graph

Problem

Use the data summarized in Table 3 to construct a frequency bar graph and relative frequency bar graph.

Video Solution



Approach

Use a horizontal axis to indicate the categories of the data (body parts) and a vertical axis to represent the frequency or relative frequency. Draw rectangles of equal width to the height that is the frequency or relative frequency for each category. The bars do not touch each other.

Technology Step-By-Step



Solution

Figure 1(a) shows the frequency bar graph, and Figure 1(b) shows the relative frequency bar graph.

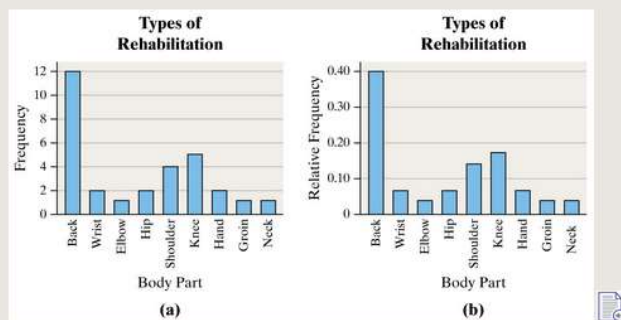


Figure 1

TABLE 3

Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} = 0.4$
Wrist	2	$\frac{2}{30} \approx 0.0667$
Elbow	1	0.0333
Hip	2	0.0667
Shoulder	4	0.1333
Knee	5	0.1667
Hand	2	0.0667
Groin	1	0.0333
Neck	1	0.0333
Total	30	1

Both bar graphs are labeled “Types of Rehabilitation”. The bar graph with the x-axis labeled “Body Part (a)” and the y-axis labeled “Frequency” is summarized below:

Body part	Frequency
Back	12
Wrist	2
Elbow	1
Hip	2
Shoulder	4
Knee	5
Hand	2
Groin	1
Neck	1

Bar graph with the x-axis labeled “Body Part (b)” and the y-axis labeled “Relative Frequency” is summarized below:

Body part	Relative frequency
Back	0.40
Wrist	0.07
Elbow	0.04
Hip	0.07
Shoulder	0.14
Knee	0.18
Hand	0.07

Groin	0.04
Neck	0.04

All data are approximate.

From <https://media.pearsoncmg.com/aw/aw_sullivanwoodbury_online_16/longdesc/Section02_1_39_01.html>

In bar graphs, the order of the categories does not usually matter. However, bar graphs that have categories arranged in decreasing order of frequency help prioritize information for decision-making purposes.

Definition

A **Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

Figure 2 illustrates a relative frequency Pareto chart for the data in [Table 3](#).

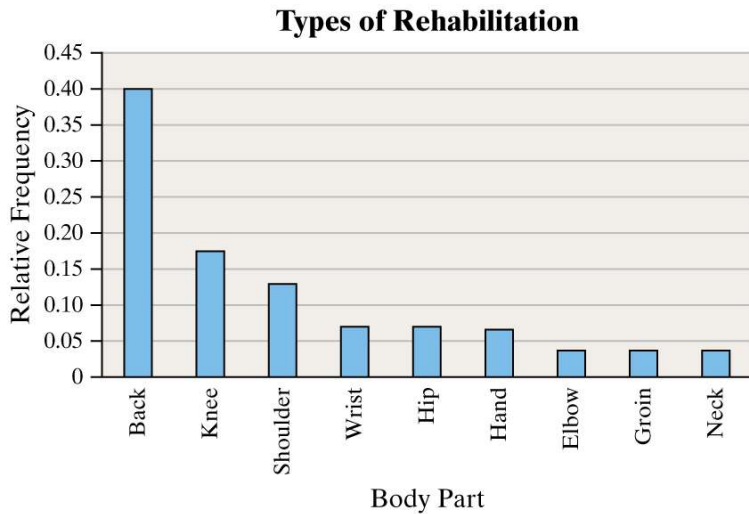


Figure 2

A bar graph labeled “Types of Rehabilitation”. The x-axis labeled “Body Part” and y-axis labeled “Relative Frequency” is summarized below:

Body Part	Relative Frequency
Back	0.40
Knee	0.17
Shoulder	0.13
Wrist	0.06
Hip	0.06
Hand	0.06
Elbow	0.04
Groin	0.04
Neck	0.04

All data are approximate.

From <https://media.pearsoncmg.com/aw/aw_sullivanwoodbury_online_16/longdesc/Section02_1_42_01.html>

Side-by-Side Bar Graphs

Suppose we want to know whether more people finished college in 2012 than in 1990.

We could draw a **side-by-side bar graph** to compare the data for the two different years.

When comparing data sets, it is best to use relative frequencies because different sample or population sizes make comparisons using frequencies difficult or misleading.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#question3>>

EXAMPLE 4 Comparing Two Data Sets

Problem

The frequency data in Table 4 represent the educational attainment (level of education) in 1990 and 2012 of adults 25 years and older who are U.S. residents. The data are in thousands. So 39,344 represents 39,344,000.

Video Solution



Technology Step-By-Step



TABLE 4

Educational Attainment	1990	2012
Not a high school graduate	39,344	25,276
High school diploma	47,643	62,113
Some college, no degree	29,780	34,163
Associate's degree	9,792	19,737
Bachelor's degree	20,833	40,561
Graduate or professional degree	11,478	22,730
Totals	158,870	204,580

Data from U.S. Census Bureau

Screen clipping taken: 06-Oct-16 09:51 AM

Draw a side-by-side relative frequency bar graph of the data.

Approach

First, determine the relative frequencies of each category for each year. To construct side-by-side bar graphs, draw two bars for each category of data—one for 1990, the other for 2012.

Solution

Table 5 shows the relative frequency of each category (by the year). The side-by-side bar graph is shown in Figure 3.

TABLE 5

Educational Attainment	Relative Frequency in 1990	Relative Frequency in 2012
Not a high school graduate	0.2476	0.1236
High school diploma	0.2999	0.3036
Some college, no degree	0.1874	0.1670
Associate's degree	0.0616	0.0965
Bachelor's degree	0.1311	0.1983
Graduate or professional degree	0.0722	0.1111

Data from U.S. Census Bureau

Educational Attainment in 1990 versus 2012

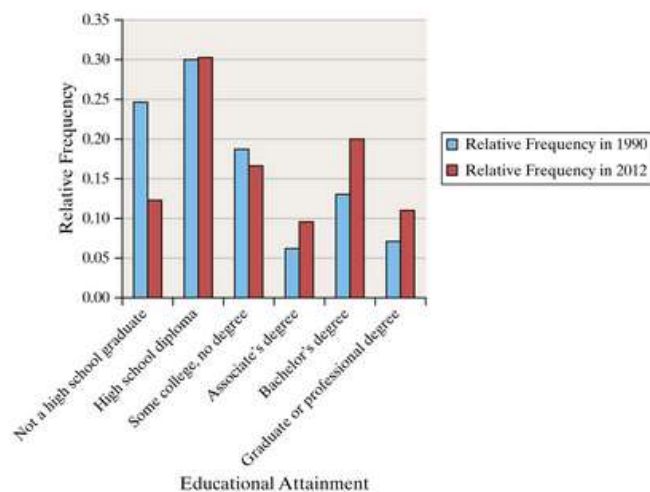


Figure 3

Screen clipping taken: 06-Oct-16 09:51 AM

The side-by-side relative frequency bar graph (Figure 3) shows additional information that was not easy to identify from the frequency table in Table 4. Comment on the interesting features of the side-by-side relative frequency bar graph.

Solution

The side-by-side bar graph illustrates that the proportion of Americans 25 years and older who had some college but no degree was higher in 1990. This information is not clear from the frequency table (Table 4) because the total population sizes are different. The increase in the number of Americans who did not complete a degree is due partly to the increases in the size of the population. In addition, the number of individuals with a Bachelor's

total population sizes are different. The increase in the number of Americans who did not complete a degree is due partly to the increases in the size of the population. In addition, the number of individuals with a Bachelor's degree almost doubled (20,833 to 40,561). However, from the side-by-side bar graph, we see that the proportion of Americans 25 years and older who had a Bachelor's degree did not double.

HIDE SOLUTION

Screen clipping taken: 06-Oct-16 09:52 AM

Horizontal Bars

So far we have only looked at bar graphs with vertical bars. However, the bars may also be horizontal. Horizontal bars are preferable when category names are lengthy. For example, Figure 4 uses horizontal bars to display the same data as in [Figure 3](#).

Educational Attainment in 1990 versus 2012

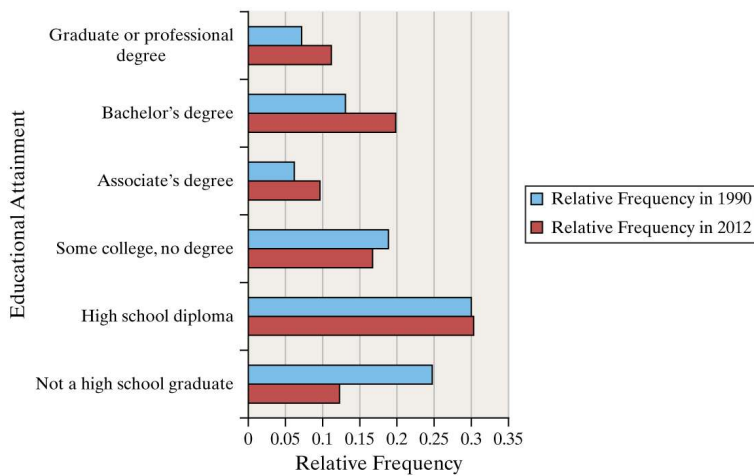


Figure 4

From <https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#question3>

2.1 Objective 3 - Construct Pie Charts

October 6, 2016 10:08 AM

Pie charts are typically used to present the relative frequency of qualitative data. In most cases, the data are nominal, but ordinal data can also be displayed in a pie chart.

DEFINITION

A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#question3>>

EXAMPLE 5 Constructing a Pie Chart

Problem

The frequency data presented in Table 6 represent the educational attainment of U.S. residents 25 years and older in 2012. The data are in thousands so 25,276 represents 25,276,000. Construct a pie chart of the data.

Video Solution



Technology Step-By-Step



TABLE 6

Educational Attainment	2012
Not a high school graduate	25,276
High school diploma	62,113
Some college, no degree	34,163
Associate's degree	19,737
Bachelor's degree	40,561
Graduate or professional degree	22,730
Total	204,580

Data from U.S. Census Bureau

Screen clipping taken: 06-Oct-16 10:12 AM

Approach

The pie chart will have one part, or sector, corresponding to each category of data. The area of each sector is proportional to the frequency of each category. For example, from Table 5, the proportion of all U.S. residents 25 years and older who are not high school graduates in 2012 is 0.1236. The category "not a high school graduate" will make up 12.36% of the area of the pie chart. Because a circle has 360 degrees, the degree measure of the sector for this category will be $(0.1236)360^\circ \approx 44^\circ$. Use a protractor to measure each angle.

Solution

We use the same approach for the remaining categories to obtain Table 7. To construct a pie chart by hand, we use a protractor to approximate the angles for each sector. See Figure 5.

TABLE 7

Educational Attainment	Frequency	Relative Frequency	Degree Measure of Each Sector
Not a high school graduate	25,276	0.1236	44
High school diploma	62,113	0.3036	109
Some college, no degree	34,163	0.1670	60
Associate's degree	19,737	0.0965	35
Bachelor's degree	40,561	0.1983	71
Graduate or professional degree	22,730	0.1111	40

Data from U.S. Census Bureau

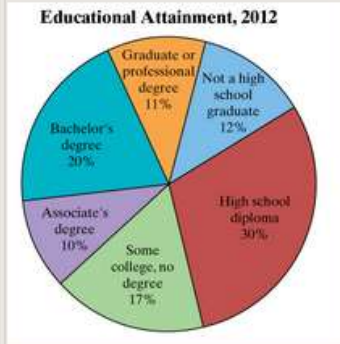


Figure 5

Screen clipping taken: 06-Oct-16 10:13 AM

To make a pie chart, we need all the categories of the variable under consideration. For example, using Example 1, we could create a bar graph that lists the proportion of patients requiring back, shoulder, or knee rehabilitation, but it would not make sense to construct a pie chart for this situation. Do you see why? Only 70% of the data would be represented (missing data for wrist, elbow, and so on).

From <https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#question3>

2.2 Organizing Quantitative Data: The Popular Displays

October 9, 2016 12:05 PM

2.2 Objective 1 - Organize Discrete Data into Tables



October 9, 2016 12:08 PM

We use the values of a discrete variable to create the classes when the number of distinct data values is small. The approach to summarizing the data is similar to that of constructing frequency or relative frequency distributions from qualitative data where the categories of data are determined by the actual observations.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 1 Constructing Frequency and Relative Frequency Distributions from Discrete Data

Problem
The manager of a Wendy's® fast-food restaurant wants to know the typical number of customers who arrive during the lunch hour. The data in Table 8 represent the number of customers who arrive at Wendy's for 40 randomly selected 15-minute intervals of time during lunch. For example, during one 15-minute interval, seven customers arrived. Construct a frequency and relative frequency distribution.

Video Solution
 



Technology Step-By-Step


TABLE 8 

Number of Arrivals at Wendy's

7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

Approach

The number of people arriving could be 0, 1, 2, 3, Table 8 shows that there are 11 categories of data from this study: 1, 2, 3, . . . , 11. We tally the number of observations for each category, count each tally, and create the frequency and relative frequency distributions.

Solution

The two distributions are shown in Table 9. From the relative frequencies, for example, 27.5% of the 15-minute intervals had six customers arrive at Wendy's during the lunch hour.

TABLE 9

Number of Customers	Tally	Frequency	Relative Frequency
1		1	$\frac{1}{40} = 0.025$
2		6	0.15
3		1	0.025
4		4	0.1
5		7	0.175
6		11	0.275
7		5	0.125
8		2	0.05
9		2	0.05
10		0	0.0
11		1	0.025

2.2 Objective 2 - Construct Histograms of Discrete Data

October 9, 2016 12:21 PM

DEFINITION

A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same, and the rectangles touch each other.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 2 Drawing a Histogram of Discrete Data

Problem

Construct a frequency histogram and a relative frequency histogram using the data in Table 9. Recall that this table summarizes the data for the number of customers who arrive at Wendy's for 40 randomly selected 15-minute intervals of time during lunch.

Video Solution



Technology Step-By-Step



Approach

On the horizontal axis, we place the value of each category of data (number of customers). The vertical axis is the frequency or relative frequency of each category. We draw rectangles of equal width centered at the value of each category. For example, the first rectangle is centered at 1. For the frequency histogram, the height of the rectangle is the frequency of the category; for the relative frequency histogram, the height is the relative frequency of the category. Remember that in a histogram, the rectangles touch.

Solution

Figures 6(a) and (b) show the frequency and relative frequency histograms, respectively.

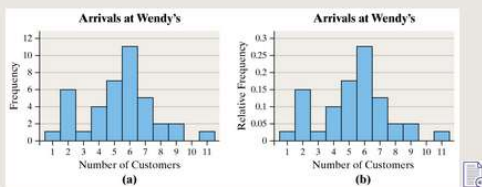


Figure 6

2.2 Objective 3 - Organize Continuous Data in Tables

October 9, 2016 12:27 PM

When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, we must create classes by using intervals of numbers.

Table 10 is a typical frequency distribution created from continuous data. The data represent the number of U.S. residents, ages 25 to 74, who had a bachelor's degree or higher in 2012.

Age	Number (in thousands)
25–34	14,064
35–44	13,871
45–54	13,417
55–64	11,862
65–74	6,334

Data from U.S. Census Bureau

Notice that the data are categorized, or grouped, by intervals of numbers. Each interval represents a class. For example, the first class is 25- to 34-year-old U.S. residents who have a bachelor's degree or higher. We read this interval as follows: “The number of U.S. residents, ages 25 to 34, with a bachelor's degree or higher was 14,064,000

in 2012.” There are five classes in the table, each with a **lower class limit** (the smallest value within the class) and an **upper class limit** (the largest value within the class). The lower class limit for the first class in Table 10 is 25; the upper class limit is 34. The **class width** is the difference between consecutive lower class limits. In Table 10, the class width is $35 - 25 = 10$. Also in Table 10, the data are continuous. So the class 25–34 actually represents 25–34.999..., or 25 up to every value less than 35.

Notice that the classes in Table 10 do not overlap. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

One exception to the requirement of equal class widths occurs in open-ended tables. A table is **open-ended** if the first class has no lower class limit or the last class has no upper class limit. The data in Table 11 represent the number of births to unmarried mothers in 2012 in the United States. The last class in the table, “40 and older,” is open-ended.

TABLE 11

Age	Number of Births (in thousands)
10–19	274
20–29	988
30–39	320
40 and older	27

Data from National Vital Statistics Report, Vol. 62, No. 3

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 3 Organizing Continuous Data into a Frequency and Relative Frequency Distribution

Problem

Suppose you are considering investing in a Roth IRA. You collect the data in Table 12, which represent the five-year rate of return (in percent, adjusted for sales charges) for a simple random sample of 40 large-blend mutual funds. Construct a frequency and relative frequency distribution of the data.

Video Solution



Technology Step-By-Step

**TABLE 12****Five-Year Rate of Return of Mutual Funds (in percent)**

3.27	3.53	3.45	5.98	4.55	3.54	4.91	4.75
3.30	10.87	3.25	3.98	5.78	4.43	4.44	10.90
5.38	4.37	4.27	3.33	8.56	11.70	3.54	5.93
4.04	3.22	4.86	3.28	11.74	6.64	3.25	3.57
4.19	4.91	12.03	3.24	4.18	4.10	3.28	3.23

Data from Morningstar.com



Approach

To construct a frequency distribution, first create classes of equal width. Table 12 has 40 observations that range from 3.22 to 12.03, so we decide to create the classes such that the lower class limit of the first class is 3 (a little smaller than the smallest data value) and the class width is 1. There is nothing magical about the choice of 1 as a class width. We could have selected a class width of 3 or any other class width. Choose a class width that you think will summarize the data nicely. The second class has a lower class limit of $3 + 1 = 4$. The classes cannot overlap, so the upper class limit of the first class is 3.99. Continuing in this fashion, we obtain the following classes:

$$3 - 3.99$$

$$4 - 4.99$$

.

.

.

$$12 - 12.99$$

This gives us ten classes. Tally the number of observations in each class, count the tallies, and create the frequency distribution. By dividing each class's frequency by 40, the number of observations, we create the relative frequency distribution.

Solution

We tally the data in Table 12 as shown in the second column of Table 13. The third column shows the frequency of each class. From the frequency distribution, we conclude that a five-year rate of return between 3% and 3.99% occurs with the most frequency. The fourth column shows the relative frequency of each class. So 40% of the large-blended mutual funds had a five-year rate of return between 3% and 3.99%. One mutual fund had a five-year rate of return between 12% and 12.99%. We might consider this mutual fund worthy of our investment. This type of information would be more difficult to obtain from the raw data.

TABLE 13

Class (5-year rate of return)	Tally	Frequency	Relative Frequency
3–3.99		16	$\frac{16}{40} = 0.4$
4–4.99		13	$\frac{13}{40} = 0.325$
5–5.99		4	0.1
6–6.99		1	0.025
7–7.99		0	0
8–8.99		1	0.025
9–9.99		0	0
10–10.99		2	0.05
11–11.99		2	0.05
12–12.99		1	0.025

2.2 Objective 4 - Construct Histograms with Continuous Data

October 9, 2016 12:42 PM

EXAMPLE 4 Drawing a Histogram of Continuous Data

Problem

Construct a frequency and relative frequency histogram of the five-year rate of return data discussed in Example 3.

Video Solution



Approach

To draw the frequency histogram, use the frequency distribution in Table 13. First, label the lower class limits of each class on the horizontal axis. Then for each class, draw a rectangle whose width is the class width and whose height is the frequency. For the relative frequency histogram, the height of the rectangle is the relative frequency.

Technology Step-By-Step



Solution

Figures 7(a) and (b) show the frequency and relative frequency histograms, respectively.

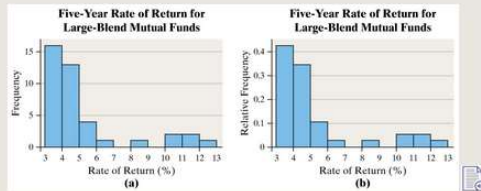


Figure 7

TABLE 13

Class (5-year rate of return)	Tally	Frequency	Relative Frequency
3-3.99		16	$\frac{16}{40} = 0.4$
4-4.99		13	$\frac{13}{40} = 0.325$
5-5.99		4	0.1
6-6.99		1	0.025
7-7.99		0	0
8-8.99		1	0.025
9-9.99		0	0
10-10.99		2	0.05
11-11.99		2	0.05
12-12.99		1	0.025

Two histograms both titled “Five-Year Rate of Return for Large-Blended Mutual Funds” with the x-axis labeled “Rate of Return (%) (a)” and y-axis labeled “Frequency” is summarized below:

Rate of Return (%)	Frequency
3-4	16
4-5	13
5-6	4
6-7	2
7-8	0
8-9	2
9-10	0
10-11	3
11-12	3
12-13	2

The x-axis labeled “Rate of Return (%) (b)” and y-axis labeled “Relative Frequency” is summarized below:

Rate of Return (%)	Relative Frequency
3-4	0.4
4-5	0.32
5-6	0.1
6-7	0.02
7-8	0.0
8-9	0.02
9-10	0.0
10-11	0.05
11-12	0.05
12-13	0.02

All data are approximate.

From <https://media.pearsoncmg.com/aw/aw_sullivanwoodbury_online_16/longdesc/Section02_2_55_01.html>

Constructing Histograms Is Somewhat of an Art Form

In Examples 3 and 4, the choices of the lower class limit of the first class and the class width were rather arbitrary. Although formulas and procedures exist for creating frequency distributions from raw data, they do not necessarily provide better summaries.

There is no one correct frequency distribution for a particular set of data. However, some frequency distributions better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form. Use the distribution that seems to provide the best overall summary of the data.

Next, you will use an applet to explore how changing the class width and the lower class limit of the first class affects the appearance of a histogram. As you use the applet, remember: The goal is to design a distribution that is best for revealing the patterns within the data.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Based on the applet activity, we can see that there is no “right” frequency distribution, but there are bad ones. The goal in constructing a frequency distribution is to reveal interesting features of the data. With that said, we typically want the number of classes to be between 5 and 20. When the data set is small, we usually want fewer classes. When the data set is large, we usually want more classes. The larger the class width, the fewer the classes in a frequency distribution. Use the following guidelines to help determine an appropriate lower class limit of the first class and class width.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Guidelines for Determining the Lower Class Limit of the First Class and Class Width

Choosing the Lower Class Limit of the First Class

Choose the smallest observation in the data set or a convenient number slightly smaller than the smallest observation in the data set. For example, in Table 12, the smallest observation is 3.22. A convenient lower class limit of the first class is 3.

Determining the Class Width

- Decide on the number of classes. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer the classes. For example, we might choose ten classes for the data in Table 12.
- Determine the class width by computing

$$\text{Class width} \approx \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}$$

- Round the value up to a convenient number. For example, using the data in Table 12, we obtain class width $\approx \frac{12.03 - 3.22}{10} = 0.881$. Round this up to 1 because this is an easy number to work with. Rounding up may result in fewer classes than were originally intended.



2.2 Objective 5 - Draw Stem-and-Leaf Plots

October 9, 2016 12:49 PM

A **stem-and-leaf plot** is another way to represent quantitative data graphically. In a stem-and-leaf plot (or *stem plot*), we use the digits to the left of the rightmost digit to form the **stem**. Each rightmost digit forms a **leaf**.

For example, a data value of 147

would have 14 as the stem and 7 as the leaf.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Table 13

State	Percent	State	Percent	State	Percent
Alabama	15.8	Kentucky	16.9	North Dakota	10.7
Alaska	10.8	Louisiana	21.1	Ohio	15.2
Arizona	18.1	Maine	13.1	Oklahoma	16.0
Arkansas	19.4	Maryland	9.6	Oregon	13.9
California	16.4	Massachusetts	10.9	Pennsylvania	13.2
Colorado	12.5	Michigan	14.3	Rhode Island	13.5
Connecticut	10.2	Minnesota	10.0	South Carolina	17.8
Delaware	13.6	Mississippi	19.7	South Dakota	13.7
District of Columbia	19.1	Missouri	15.3	Tennessee	17.4
Florida	15.1	Montana	15.0	Texas	17.2
Georgia	18.3	Nebraska	11.2	Utah	11.0
Hawaii	13.0	Nevada	15.6	Vermont	11.4
Idaho	15.1	New Hampshire	7.9	Virginia	11.0
Illinois	13.4	New Jersey	10.4	Washington	12.1
Indiana	15.4	New Mexico	21.3	West Virginia	17.1
Iowa	10.4	New York	16.6	Wisconsin	12.2
Kansas	14.2	North Carolina	16.3	Wyoming	10.2

EXAMPLE 5 Constructing a Stem-and-Leaf Plot

Problem

The data in Table 14 represent the two-year average percentage of persons living in poverty, by state, for the years 2011–2012. Draw a stem-and-leaf plot of the data.

Video Solution



Technology Step-By-Step



Approach

Step 1 We will treat the integer portion of the number as the stem and the decimal portion as the leaf. For example, the stem of Alabama will be 15 and the leaf will be 8. The stem of 15 will include all data from 15.0 to 15.9.

Step 2 Write the stems vertically in ascending order and then draw a vertical line to the right of the stems.

Step 3 Write the leaves corresponding to the stem.

Step 4 Within each stem, rearrange the leaves in ascending order. Title the plot and include a legend to indicate what the values represent.

Solution

Step 1 The stem from Alabama is 15, and the corresponding leaf is 8. The stem from Alaska is 10 and its leaf is 8, and so on.

Step 2 Because the lowest data value is 7.9 and the highest data value is 21.3, let the stems range from 7 to 21. Write the stems vertically in Figure 8(a), along with a vertical line to the right of the stem.

7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
(a)

Figure 8(a)

Step 3 Write the leaves corresponding to each stem. See Figure 8(b).

7	7	9
8	8	
9	9	6
10	10	8 2 4 9 0 4 7 2
11	11	2 0 4 0
12	12	5 1 2
13	13	6 0 4 1 9 2 5 7
14	14	2 3
15	15	8 1 1 4 3 0 6 2
16	16	4 9 6 3 0
17	17	8 4 2 1
18	18	1 3
19	19	4 1 7
20	20	
21	21	1 3

(a) (b)

Figure 8(b)

Step 4 Rearrange the leaves in ascending order, give the plot a title, and add a legend. See Figure 8(c).

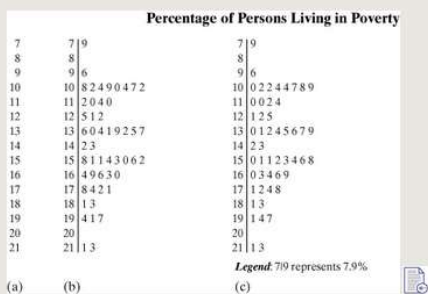


Figure 8(c)

Construction of a Stem-and-Leaf Plot

Step 1 The stem of a data value will consist of the digits to the left of the rightmost digit. The leaf of a data value will be the rightmost digit.

Step 2 Write the stems in a vertical column in increasing order. Draw a vertical line to the right of the stems.

Step 3 Write each leaf corresponding to the stems to the right of the vertical line.

Step 4 Within each stem, rearrange the leaves in ascending order, title the plot, and include a legend to indicate what the values represent.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

NOTE

Sometimes the steps listed for creating stem-and-leaf plots must be modified to meet the needs of the data. This will be illustrated in Example 6.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Stem-and-Leaf Plots versus Histograms

Notice that a stem-and-leaf plot looks much like a histogram turned on its side. The stem serves as the class. For example, the stem 10 contains all data from 10.0 to 10.9.

The leaves represent the frequency (height of the rectangle). Therefore, it is important to space the leaves evenly.

One advantage of the stem-and-leaf plot over frequency distributions and histograms is that the raw data can be retrieved from the stem-and-leaf plot. So, from a stem-and-leaf plot we can determine the maximum observation. We cannot learn this information immediately from a histogram. Refer to Figure 9, which shows a histogram of the poverty data drawn in StatCrunch. We can see that the largest observation is between 21

and 21.9, but we don't know that the largest value is 21.3, which is clear from the stem-and-leaf plot in [Figure 8\(c\)](#).

Percentage of Persons Living in Poverty

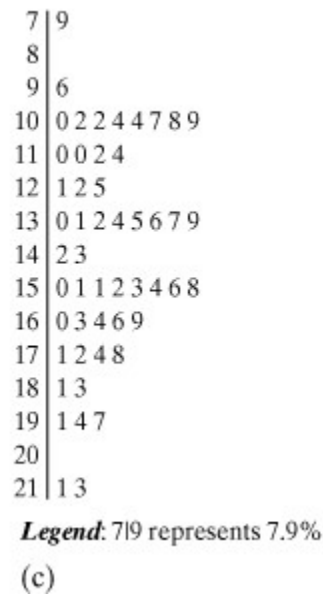


Figure 8(c)

On the other hand, stem-and-leaf plots lose their usefulness when data sets are large or consist of a large range of values.

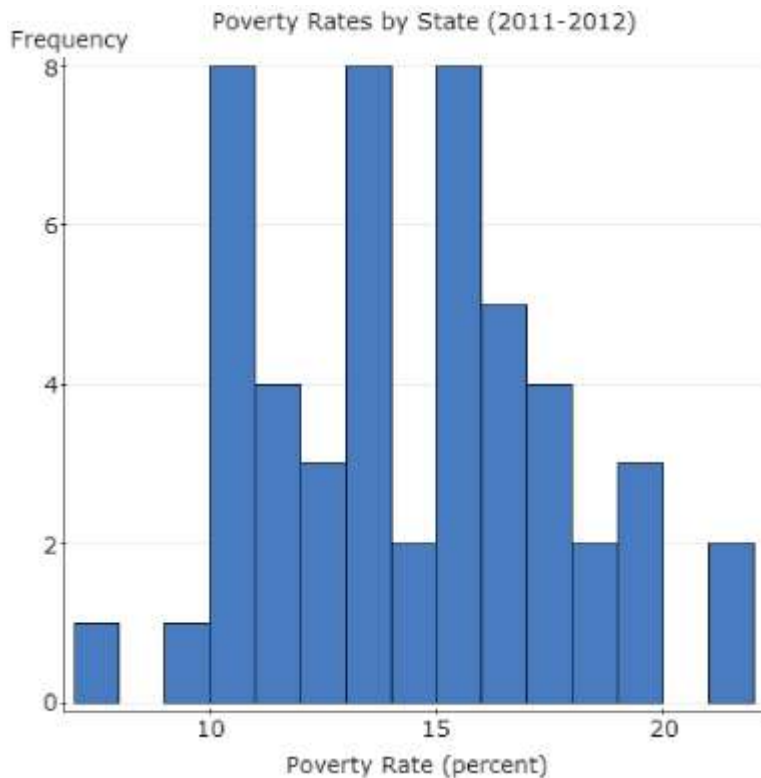


Figure 9

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

TABLE 12

Five-Year Rate of Return of Mutual Funds (in percent)

3.27	3.53	3.45	5.98	4.55	3.54	4.91	4.75
3.30	10.87	3.25	3.98	5.78	4.43	4.44	10.90
5.38	4.37	4.27	3.33	8.56	11.70	3.54	5.93
4.04	3.22	4.86	3.28	11.74	6.64	3.25	3.57
4.19	4.91	12.03	3.24	4.18	4.10	3.28	3.23

Data from Morningstar.com

Problem

Construct a stem-and-leaf plot of the five-year rate of return data listed in [Table 12](#).

Approach

Step 1 If we use the integer portion as the stem and the decimals as the leaves, the stems will be 3,4,5,...,12,

but the leaves will be two digits (such as 27 and 30

). This is not acceptable because each leaf must be a single digit. To solve this problem, round the data to the nearest tenth.

Step 2 Create a vertical column of the integer stems in increasing order.

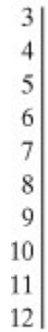
Step 3 Write the leaves corresponding to each stem.

Step 4 Rearrange the leaves in ascending order, title the plot, and include a legend.

Solution

Step 1 Round the data from Table 12 to the nearest tenth as shown in [Table 15](#).

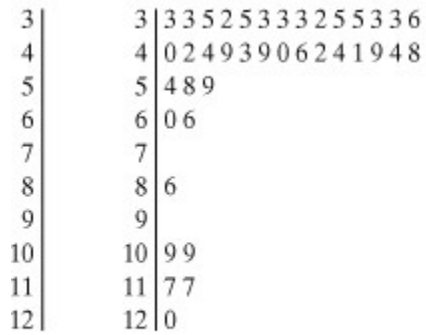
Step 2 Write the stems vertically in ascending order as shown in Figure 10(a).



(a)

Figure 10(a)

Step 3 Write the leaves corresponding to each stem as shown in Figure 10(b).



(a)

(b)



Figure 10(b)

Step 4 Rearrange the leaves in ascending order, title the plot, and include a legend as shown in Figure 10(c).

Five-Year Rate of Return of Mutual Funds

3	3	3 3 5 2 5 3 3 3 2 5 5 3 3 6 2	3	2 2 2 3 3 3 3 3 3 3 5 5 5 6
4	4	0 2 4 9 3 9 0 6 2 4 1 9 4 8	4	0 0 1 2 2 3 4 4 4 6 8 9 9 9
5	5	4 8 9	5	4 8 9
6	6	0 6	6	0 6
7	7		7	
8	8	6	8	6
9	9		9	
10	10	9 9	10	9 9
11	11	7 7	11	7 7
12	12	0	12	0

Legend: 12|0 represents 12.0%

(a) (b)

(c)

Figure 10(c)

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Altering the data to construct the stem-and-leaf plot in Figure 10(c) means that we cannot retrieve the original data. A second limitation appearing in Example 6 is that we are effectively forced to use a class width of 1.0 even though a larger class width may be more desirable. This illustrates that we must weigh the advantages against the disadvantages when choosing which type of graph to use to summarize data.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Split Stems

The data in Table 16 range from 11 to 48.

Figure 11(a) shows a stem-and-leaf plot using the tens digit as the stem and the ones digit as the leaf. The data appear rather bunched. To resolve this problem, we can use **split stems**. For example, rather than using one stem for the class of data 10–19, we can use two stems, one for the 10–14 interval and the second for the 15–19 interval. We do this in Figure 11(b).

TABLE 16

27	17	11	24	36
13	29	22	18	17
23	30	12	46	17
32	48	11	18	23
18	32	26	24	38
24	15	13	31	22
18	21	27	20	16
15	37	19	19	29

```

1 | 11233556777888899
2 | 01223344467799
3 | 0122678
4 | 68

```

Legend: 111 represents 11

(a)

```

1 | 11233
1 | 556777888899
2 | 012233444
2 | 67799
3 | 0122
3 | 678
4 |
4 | 68

```

Legend: 111 represents 11

(b)



Figure 11

The stem-and-leaf plot shown in Figure 11(b) reveals a better distribution of the data. As with the determination of class intervals in the creation of frequency histograms, judgment plays a major role in how you present data in a stem-and-leaf plot. There is no such thing as a correct stem-and-leaf plot. However, a quick comparison of Figures 11(a) and (b) shows that some plots are better than others.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

2.2 Objective 6 - Draw Dot Plots

October 9, 2016 01:13 PM

We draw a **dot plot** by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 7 Drawing a Dot Plot

Problem

Draw a dot plot for the data from Table 8.

Video Solution



Approach

The smallest observation in the data set is 1, and the largest is 11. Write the numbers 1 through 11 horizontally. Each time that a specific observation occurred, place a dot above the value of the observation. Remember to give the graph a title and label the horizontal axis.

Technology Step-By-Step



Solution

Figure 12 shows the dot plot.

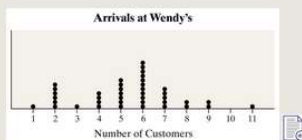


Figure 12

TABLE 8

Number of Arrivals at Wendy's

7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

2.2 Objective 7 - Identify the Shape of a Distribution

October 9, 2016 01:15 PM

One way that a variable is described is through the shape of its distribution. Distribution shapes are typically classified as *symmetric*, *skewed left*, or *skewed right*. Figure 13 displays various histograms and the shape of the distribution.

Figures 13(a) and (b) show symmetric distributions. They are symmetric because if we split the histogram down the middle, the right and left sides are mirror images. Figure 13(a) is a **uniform distribution** because the frequency of each value of the variable is evenly spread across the values of the variable. Figure 13(b) displays a **bell-shaped distribution** because the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle. The distribution in Figure 13(c) is **skewed right**. Notice that the tail to the right of the peak is longer than the tail to the left of the peak. Finally, Figure 13(d) illustrates a distribution that is **skewed left** because the tail to the left of the peak is longer than the tail to the right of the peak.

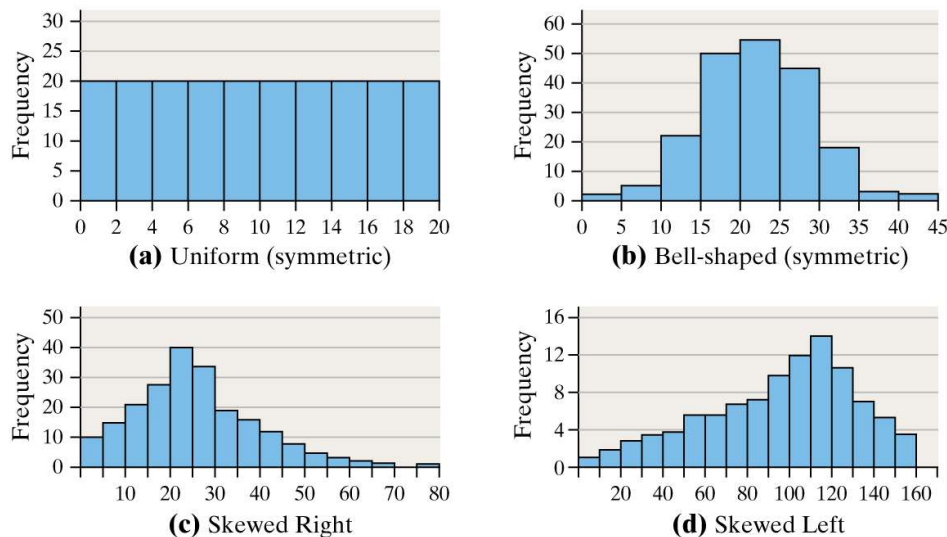


Figure 13

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 8 Identifying the Shape of a Distribution

Problem

Figure 14 displays the histogram obtained in Example 4 for the five-year rate of return for large-blended mutual funds. Describe the shape of the distribution.

Video Solution

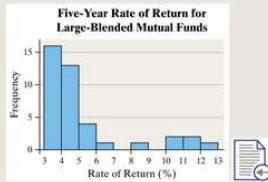


Figure 14

Approach

Compare the shape of the distribution displayed in Figure 14 with Figure 13.

Solution

Because the histogram looks most like Figure 13(c), the distribution is skewed right.

2.3 Additional Displays of Quantitative Data

October 10, 2016 08:49 AM

2.3 Objective 1 - Construct Frequency Polygons

October 10, 2016 08:53 AM

DEFINITION

A **frequency polygon** is a graph that uses points, connected by line segments, to represent the frequencies for the classes. It is constructed by plotting a point above each **class midpoint** (the sum of consecutive lower class limits divided by 2) on a horizontal axis at a height equal to the frequency of the class.

From <https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>

EXAMPLE 1 Constructing a Frequency Polygon

Problem

Draw a frequency polygon of the five-year rate of return data summarized in Table 13.

Video Solution



Approach

Begin by calculating the class midpoints of each class. Plot points above each class midpoint at a height equal to the frequency of the class. Next, draw line segments connecting the points. Draw two additional line segments connecting each end of the graph with the horizontal axis. Remember to label your axes and title your graph.

Technology Step-By-Step



Solution

The class midpoints of each class are shown in Table 17. Now plot points with the class midpoints as the x -coordinates and the frequencies as the y -coordinates. Connect these points with line segments. Then determine the midpoint of the class preceding the first class (2.5) and the midpoint of the class after the last class (13.5). Finally, connect each end of the graph with the horizontal axis at (2.5, 0) and (13.5, 0), respectively, to create Figure 15.

TABLE 17

Class (5-year rate of return)	Class Midpoint	Frequency	Relative Frequency
3–3.99	$\frac{3+4}{2} = 3.5$	16	0.4
4–4.99	$\frac{4+5}{2} = 4.5$	13	0.325
5–5.99	5.5	4	0.1
6–6.99	6.5	1	0.025
7–7.99	7.5	0	0
8–8.99	8.5	1	0.025
9–9.99	9.5	0	0
10–10.99	10.5	2	0.05
11–11.99	11.5	2	0.05
12–12.99	12.5	1	0.025



2.3 Objective 2 Create Cumulative Frequency and Relative Frequency Distributions

October 10, 2016 09:14 AM

Because quantitative data can be ordered (written in ascending or descending order), they can be summarized in a *cumulative frequency distribution* and a *cumulative relative frequency distribution*.

DEFINITION

A **cumulative frequency** distribution displays the aggregate frequency of the category. In other words, it displays the total number of observations less than or equal to the upper class limit of the class.

A **cumulative relative frequency** distribution displays the proportion (or percentage) of observations less than or equal to the upper class limit of the class.

So the cumulative frequency for the second class is the sum of the frequencies of classes 1 and 2; the cumulative frequency for the third class is the sum of the frequencies of classes 1, 2, and 3; and so on.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Class (5-year rate of return)	Tally	Frequency	Relative Frequency
3–3.99		16	$\frac{16}{40} = 0.4$
4–4.99		13	$\frac{13}{40} = 0.325$
5–5.99		4	0.1
6–6.99		1	0.025
7–7.99		0	0
8–8.99		1	0.025
9–9.99		0	0
10–10.99		2	0.05
11–11.99		2	0.05
12–12.99		1	0.025

EXAMPLE 2 Constructing a Cumulative and Cumulative Relative Frequency Distribution**Problem**

Obtain a cumulative frequency distribution and a cumulative relative frequency distribution for the five-year rate of return data summarized in Table 13.

Video Solution

**Approach**

For the cumulative frequency distribution, determine the total number of observations less than or equal to each class. For the cumulative relative frequency distribution, determine the proportion of observations less than or equal to each class.

Solution

Table 18 displays the cumulative frequency and cumulative relative frequency of the data summarized from Table 13. Table 18 shows that 34 of the 40 mutual funds had five-year rates of return of 7.99% or less. The cumulative relative frequency distribution is shown in the fifth column. We see that 85% of the mutual funds had a five-year rate of return of 7.99% or less. Also, A mutual fund with a five-year rate of return of 11% or higher outperformed 92.5% of its peers. Notice that the last class (12– 12.99) has a cumulative relative frequency of 1—this will always be the case.

TABLE 18

Class (5-year rate of return)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
3–3.99	16	0.4	16	0.4
4–4.99	13	0.325	29	0.725
5–5.99	4	0.1	33	0.825
6–6.99	1	0.025	34	0.85
7–7.99	0	0	34	0.85
8–8.99	1	0.025	35	0.875
9–9.99	0	0	35	0.875
10–10.99	2	0.05	37	0.925
11–11.99	2	0.05	39	0.975
12–12.99	1	0.025	40	1

2.3 Objective 3 Construct Frequency and Relative Frequency Ogives

October 10, 2016 09:20 AM

DEFINITION

An **ogive** (read as “oh jive”) is a graph that represents the cumulative frequency or cumulative relative frequency for the class. It is constructed by plotting points whose x -coordinates are the upper class limits and whose y -coordinates are the cumulative frequencies or cumulative relative frequencies of the class. Then line segments are drawn connecting consecutive points. An additional line segment is drawn connecting the first point to the horizontal axis at a location representing the upper limit of the class that would precede the first class (if it existed).

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 3 Constructing Ogives

Problem

Draw a relative frequency ogive of the 5-year rate of return data summarized in Table 18.

Video Solution



Approach

A relative frequency ogive is drawn by plotting points whose x -coordinates are the upper class limit of each class and whose y -coordinates are the cumulative relative frequencies of each class. Then connect the points with line segments. Also, an additional line segment is drawn connecting the first point to the horizontal axis at a location representing the upper limit of the class that would precede the first class (if it existed).

Technology Step-By-Step



HIDE APPROACH

Solution

See Figure 16. Notice how 85% of the mutual funds had a 5-year rate of return less than or equal to 7.99%. Ogives do not have a line segment drawn from the last point to the horizontal axis because ogives represent the number or proportion of observations less than or equal to the x -coordinate of the point. Note the height of the last point in a relative frequency ogive is always 1.

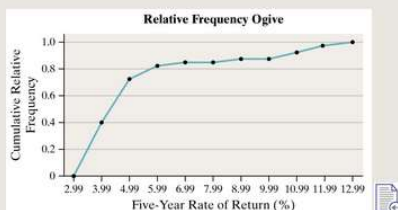


Figure 16

HIDE SOLUTION

2.3 Objective 4 Draw Time-Series Graphs

October 10, 2016 09:25 AM

If the value of a variable is measured at different points in time, then the data are referred to as **time-series data**. The closing price of Cisco Systems stock at the end of each year for the past 12 years is an example of time-series data.

DEFINITION


A **time-series plot** is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis. Line segments are then drawn connecting the points.


Time-series plots are very useful in identifying trends in the data over time.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 4 Drawing a Time-Series Plot

Problem
A housing permit is authorization from a governing body to construct a privately owned housing unit. The data in Table 19 represent the number of housing permits (in thousands) issued from 2000 to 2012 in the United States. Construct a time-series plot of the data. What was the percentage change in housing permits from 2008 to 2009?

Video Solution


Technology Step-By-Step


Year	Housing Permits (thousands)
2000	1592.3
2001	1636.7
2002	1747.7
2003	1889.2
2004	2070.1
2005	2155.3
2006	1838.9
2007	1398.4
2008	905.4
2009	583.0
2010	604.6
2011	624.1
2012	829.7

Approach

Step 1 Plot points for each year, with the date on the horizontal axis and the number of housing permits on the vertical axis.

Step 2 Connect the points with line segments.

Solution

Figure 17 shows the time-series plot. The overall trend is dismal if you are in the home-building business. Housing permits peaked in 2005 and declined sharply until 2009. Since then, permits have stabilized. The percentage change in housing permits issued from 2008 to 2009 is

$$\begin{aligned} \text{Percentage change in permits issued} &= \frac{583.0 - 905.4}{905.4} \quad \% \text{ change} = \frac{P_2 - P_1}{P_1} \\ &\approx -0.356 \\ &= -35.6\% \end{aligned}$$

So housing permits issued declined 35.6% from 2008 to 2009.



Figure 17

2.4 Graphical Misrepresentation of Data

October 11, 2016 12:08 PM

2.4 Objective 1 Describe What can make a graph misleading or deceptive

October 11, 2016 12:09 PM

Statistics: *The only science that enables different experts using the same figures to draw different conclusions.*—Evan Esar

Statistics often gets a bad rap for having the ability to manipulate data to support any position. One method of distorting the truth is through graphics. We mentioned in Section 2.1 how visual displays send more powerful messages than do raw data or tables of data. Because graphs are so powerful, care must be taken in constructing graphs and interpreting their messages. Graphs may *mislead* or *deceive*.

- Graphs mislead if they unintentionally create an incorrect impression.
- Graphs deceive if they purposely create an incorrect impression.

In either case, a reader's incorrect impression can have serious consequences. Therefore, it is important to be able to recognize misleading and deceptive graphs

The most common graphical misrepresentations of data involve the scale of the graph, an inconsistent scale, or a misplaced origin. Increments between tick marks should be consistent, and scales for comparative graphs should be the same. Also, because readers usually assume that the baseline, or zero point, is at the bottom of the graph, a graph that begins at a higher or lower value can be misleading.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Misrepresentation of Data

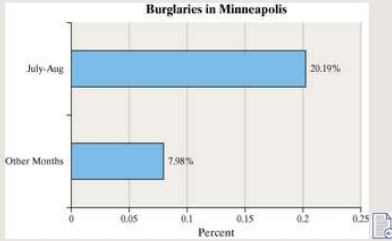
A home security company located in Minneapolis, MN develops a summer ad campaign with the slogan "When you leave for vacation, burglars leave for work."

According to the City of Minneapolis, roughly 20% of home burglaries occur during the peak vacation months of July and August. The advertisement contains the graphic shown. Explain what is wrong with the graphic.

EXAMPLE 1 Misrepresentations of Data

Problem

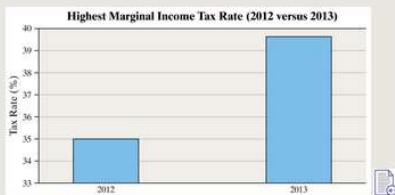
A home security company located in Minneapolis, Minnesota develops a summer ad campaign with the slogan "When you leave for vacation, burglars leave for work." According to the city of Minneapolis, roughly 20% of home burglaries occur during the peak vacation months of July and August. The advertisement contains the graphic shown. Explain what is wrong with the graphic.



EXAMPLE 2 Misrepresentations of Data by Manipulating the Vertical Scale

Problem

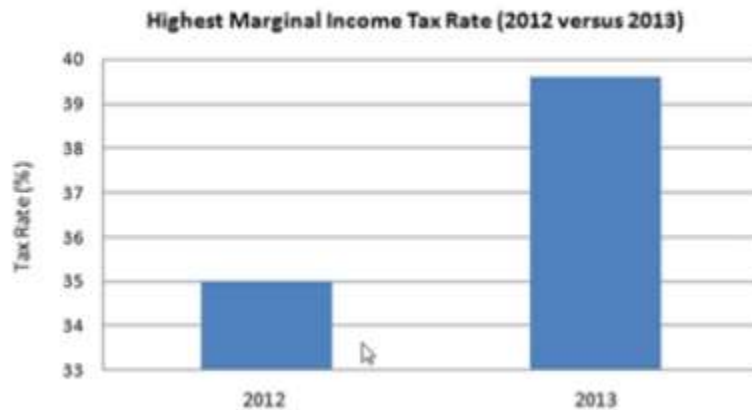
A national news organization developed the following graphic to illustrate the change in the highest marginal tax rate effective January 1, 2013. Why might this graph be considered misleading?



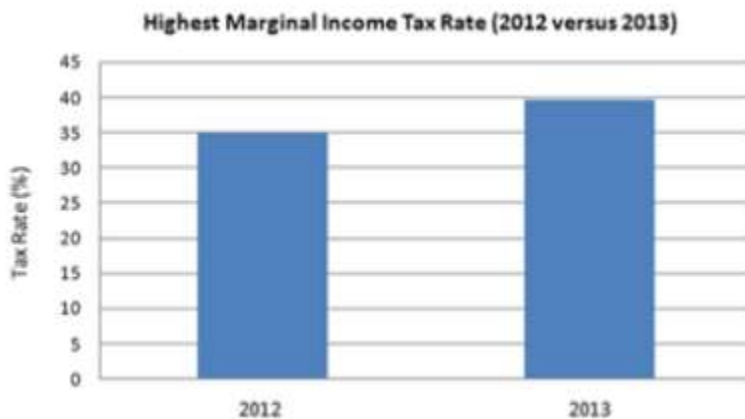
Example – Misrepresentation of Data by Manipulating the Vertical Scale

A national news organization developed the following graphic to illustrate the change in the highest marginal tax rate effective January 1, 2013.

Why might this graph be considered misleading?



Starts at 33

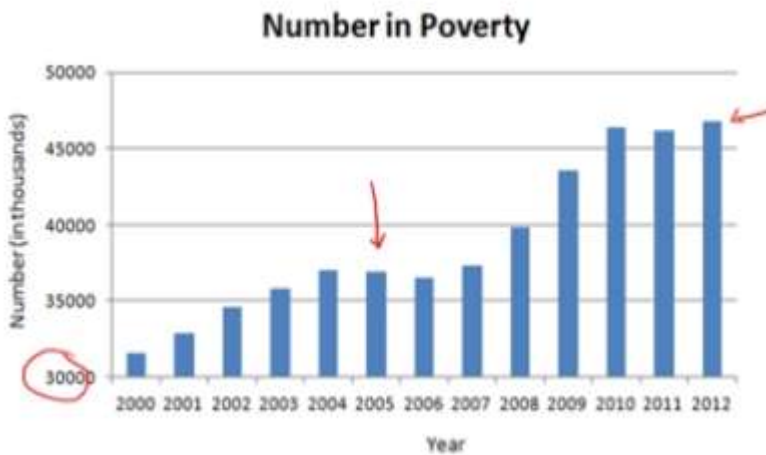
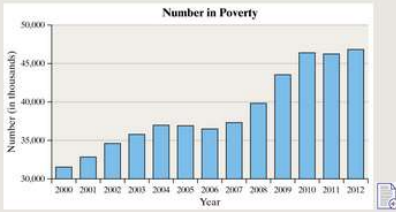


Starts at 0, where it should

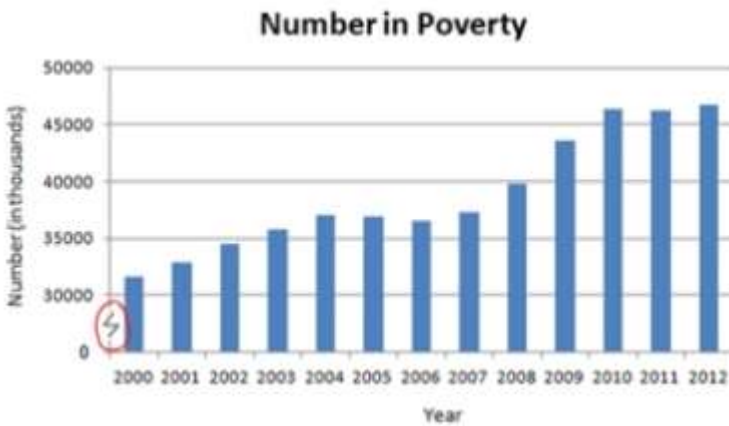
EXAMPLE 3 Misrepresentations of Data by Manipulating the Vertical Scale

Problem

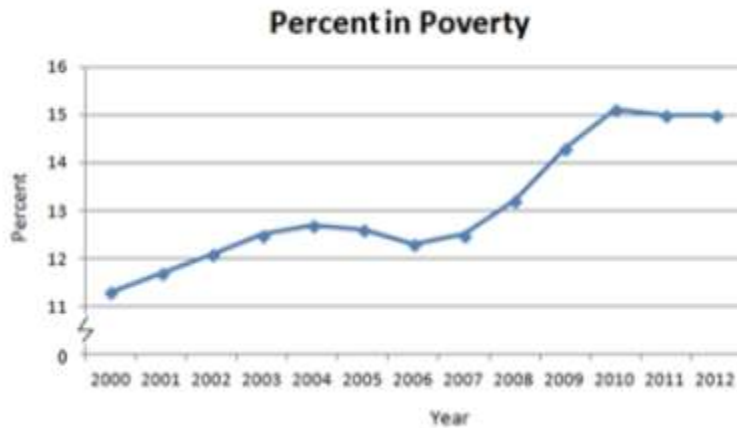
The graph depicts the number of residents in the United States living in poverty. Why might this graph be considered misrepresentative?



Incorrect



Correct



Time-Series plot

EXAMPLE 4 Misrepresentations of Data

Problem

The bar graph shown is a *USA Today*-type graph. A survey was conducted by Impulse Research in which individuals were asked how they would flush a toilet when the facilities are not sanitary. What is wrong with the graphic?



Example – Misrepresentation of Data

The following bar graph is a *USA Today* type graph. A survey was conducted by Impulse Research for Quilted Northern Confidential in which individuals were asked how they would flush a toilet when the facilities are not sanitary. What's wrong with the graphic?



Newspapers, magazines, and Internet sites often go for a "wow" factor when displaying graphs. The graph designer may be more interested in catching the reader's eye than making the data stand out. The two most commonly used tactics are 3-D graphs and *pictograms* (graphs that use pictures to represent the data). The use of 3-D effects is strongly discouraged because such graphs are often difficult to read, add little value to the graph, and distract the reader from the data.

When comparing bars, our eyes are really comparing the *areas* of the bars. That is why we emphasize that the bars or classes should have the same width. Uniform width ensures that the area of the bar is proportional to its height so that we can simply compare the heights of the bars. However, when we use two-dimensional pictures in place of bars, as with pictograms, it is not possible to obtain a uniform width. To avoid distorting the picture when values increase or decrease, both the height and width of the picture must be adjusted. This often leads to misleading graphs.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 5 Misrepresentations of Data by Manipulating Dimension

Problem

Soccer continues to grow in popularity as a sport in the United States. In 1991, there were approximately 10 million participants in the United States aged 7 years and older. By 2009, this number had climbed to 14 million. To illustrate this increase, we could create a graphic like the one shown below. Describe how the graph may be misleading. *Source:* U.S. Census Bureau; National Sporting Goods Association



Example – Misleading Graphs

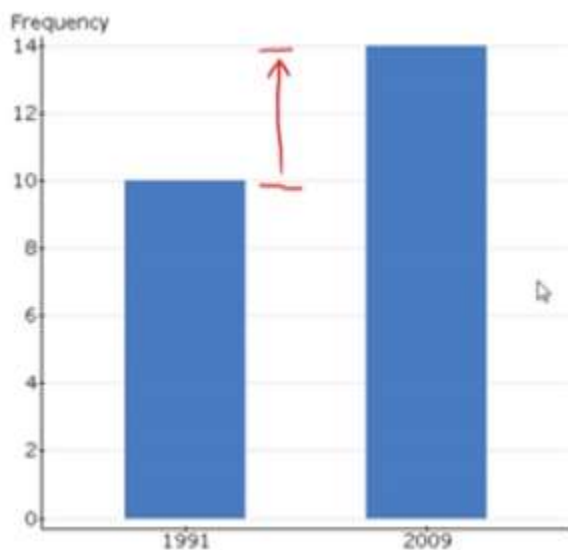
Soccer continues to grow in popularity as a sport in the United States.

In 1991 there were approximately 10 million participants in the United States aged 7 years or older.

By 2009 this number had climbed to 14 million.




(Source: U.S. Census Bureau, National Sporting Goods Association.)



Soccer Participation

1991 

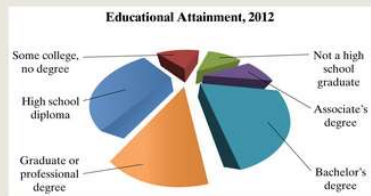
2009 

 = 1 million participants

EXAMPLE 6 Misrepresentations of Data: Three-Dimensional Scale

Problem

The figure represents the educational attainment (level of education) in 2012 of adults 25 years and older who are U.S. residents. Why might this graph be considered misrepresentative?

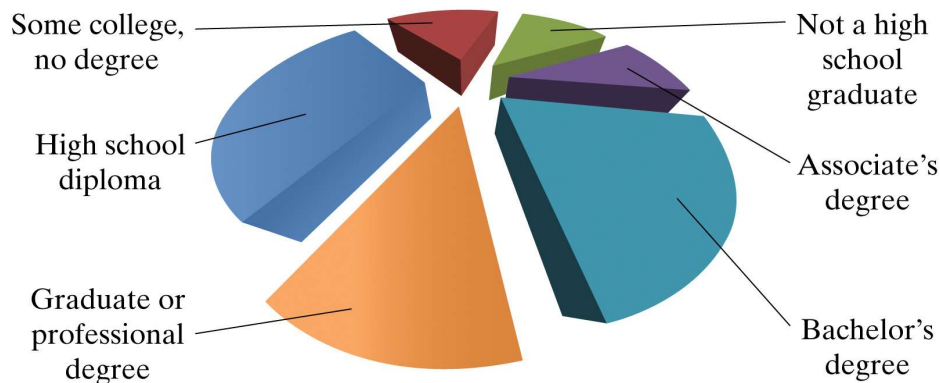


Example – Misleading Graphs

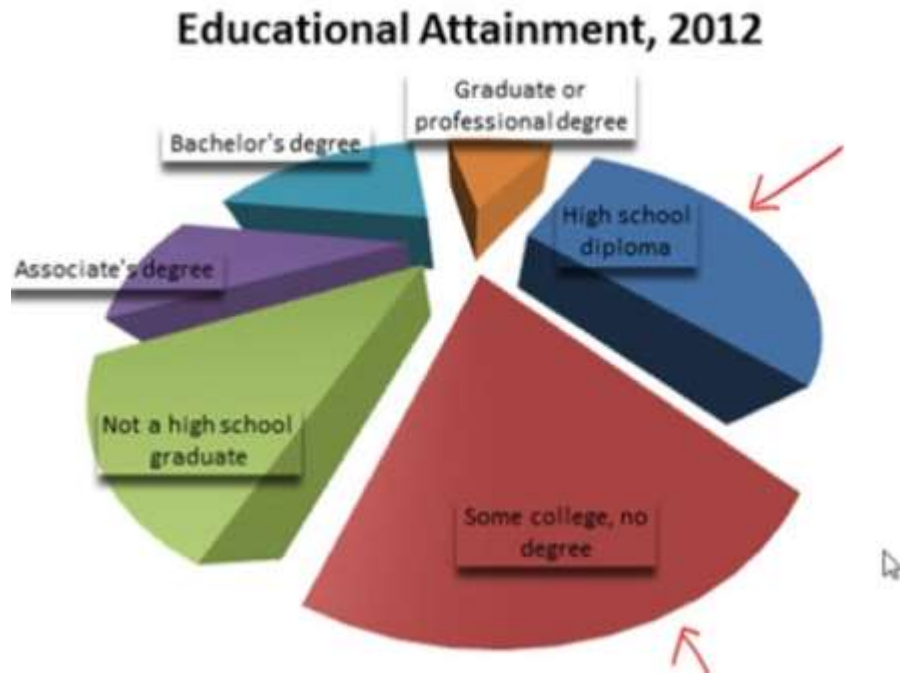
The figure below represents the educational attainment (level of education) in 2012 of adults 25 years and older who are U.S. residents.

Why might this graph be considered misrepresentative?

Educational Attainment, 2012



Educational Attainment	2012
Not a high school graduate	25,276
High school diploma	62,133
Some college, no degree	34,163
Associate's degree	19,737
Bachelor's degree	40,561
Graduate or professional degree	22,730
Totals	204,580



Graph rotated 180°

The material presented in this section is by no means all-inclusive. There are many ways graphs mislead or deceive. Two popular texts written about ways that graphs mislead or deceive are *How to Lie with Statistics* (W. W. Norton & Company, Inc., 1982) by Darrell Huff and *The Visual Display of Quantitative Information* (Graphics Press, 2001) by Edward Tufte.

Below are some guidelines for constructing good graphics.

Guidelines for Constructing Good Graphics

- • Label and name the axes clearly, providing explanations if needed. Include units of measurement and a data source when appropriate.
- • Include a meaningful title on the graph
- • Avoid distortion. Never lie about the data.
- • Minimize the amount of white space in the graph. Use the available space to let the data stand out. If you truncate the scales, clearly indicate this to the reader.
- • Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader from the data.
- • Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- • Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion to draw attention to that area. Don't try to force the reader to a specific part of the

graph. Let the data speak for themselves.

- • Avoid relative graphs that do not contain data or scales.

One final point to make. When reading graphs, look at the source of the data represented in the graphic. Often, a group with an agenda will conduct allegedly unbiased studies and report the results that support their position. Always "consider the source" and any possible hidden agendas they may have when reading graphics.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

3.1 Measures of Central Tendency

October 13, 2016 09:11 AM

An Introduction to Measures of Central Tendency

A **measure of central tendency** numerically describes the average (or typical) data value. We hear the word *average* in the news all the time:

- The average miles per gallon of gasoline for the 2013 Chevrolet Corvette Z06 in highway driving is 24.
- According to the U.S. Census Bureau, the national average commute time to work in 2010 was 25.3 minutes.
- According to the U.S. Census Bureau, the average household income in 2010 was \$49,455.
- The average American woman is 5' 4" tall and weighs 142 pounds.

In this chapter, we discuss the three most widely used measures of central tendency: the *mean*, the *median*, and the *mode*.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

3.1 Objective 1 Determine the Arithmetic Mean of a Variable from Raw Data

October 13, 2016 09:17 AM

In everyday language, the word *average* often represents the arithmetic mean. To compute the arithmetic mean of a set of data, the data must be [quantitative](#).

DEFINITION

The **arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.

The **population arithmetic mean**, μ

(pronounced "mew"), is a [parameter](#) that is computed using data from all the individuals in a population.

The **sample arithmetic mean**, \bar{x}

(pronounced "x-bar"), is a [statistic](#) that is computed using data from individuals in a sample.

While other types of means exist, the arithmetic mean is generally referred to as **the mean**. We will follow this practice for the remainder of the course.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

We usually use [Greek letters](#) to represent parameters and Roman letters (such as x or s) to represent statistics. The notation used below (for the arithmetic mean) may look intimidating. It is important to understand the notation in a formula because it is then easier to remember and use it.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Population Mean

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Remember: Capital n (N) equals Population

Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Remember: Lower case n (n) equals sample size

Σ

Greek symbol for "sigma" is the notation for "sum"

$$\frac{\sum x_i}{N}$$

This formula states **Take the sum (Σ) of all "x" values and divide by the population (N) to get the population mean**

$$\frac{\sum x_i}{n}$$

This formula states **Take the sum (Σ) of all x values and divide by the sample size (n) to get the sample mean**

Population Mean

If x_1, x_2, \dots, x_N are the N observations of a variable from a population, then the population mean, μ (pronounced "mew"), is

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum x_i}{N}$$

Sample Mean

If x_1, x_2, \dots, x_n are n observations of a variable from a sample, then the sample mean, is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$$

μ = Population Mean

x-bar (x with horizontal bar above) = Sample Mean

N = Population

n = Sample

Σ = Sum

X_i = Dataset

M = Median (not covered in this chapter, but still useful to know)

Throughout this course, we agree to round the mean to one more decimal place than that in the raw data.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 1 Computing a Population Mean and a Sample Mean

Problem

Table 1 shows the first exam scores of the ten students enrolled in Introductory Statistics.

Video Solution



Technology Step-By-Step



TABLE 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Part A

Part B

Part C

Compute the population mean, μ .

Approach

To compute the population mean, μ , add all the data values (test scores) and then divide by the number of individuals in the population.

Solution

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + \cdots + x_{10} \\ &= 82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88 \\ &= 790\end{aligned}$$

Divide this result by 10, the number of students in the class.

$$\mu = \frac{\sum x_i}{N} = \frac{790}{10} = 79$$

Although it was not necessary in this problem, we will agree to round the mean to one more decimal place than that in the raw data.

Find a **simple random sample** of size $n = 4$ students.

Approach

Recall that a calculator with a random number generator or computer software can be used to obtain simple random samples. We will use a TI-84 Plus graphing calculator.

Solution

To find a simple random sample of size $n = 4$ from a population of size $N = 10$, we will use the TI-84 Plus random number generator with a **seed** of 54. Figure 1 shows the students in the sample: Bilal (90), Ryanne (77), Pam (71), and Michelle (82).



Figure 1

Compute the sample mean, \bar{x} , of the sample found in part (B).

Approach

To find the sample mean, \bar{x} , add the data values from the individuals in the **sample** and then divide by $n = 4$, the sample size.

Solution

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + x_4 \\ &= 90 + 77 + 71 + 82 \\ &= 320\end{aligned}$$

Divide this result by 4, the number of individuals in the sample.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{320}{4} = 80$$

Student	Test Score
Bilal	90
Ryanne	77
Pam	71
Michelle	82

Seed:

The **seed** gives the calculator its starting point to generate the list of random numbers. The choice of the

seed is up to the individual obtaining the simple random sample

Visualizing the Mean as the Center of Gravity: An Animation

It helps to think of the mean as the center of gravity. In other words, the mean is the value such that a histogram of the data is perfectly balanced, with equal weight on each side of the mean. Figure 2 shows a histogram of the data in [Table 1](#). Recall, the mean of the data in Table 1 is 79 points. Play with the fulcrum (triangle) to verify that the mean is the balancing point of the data.

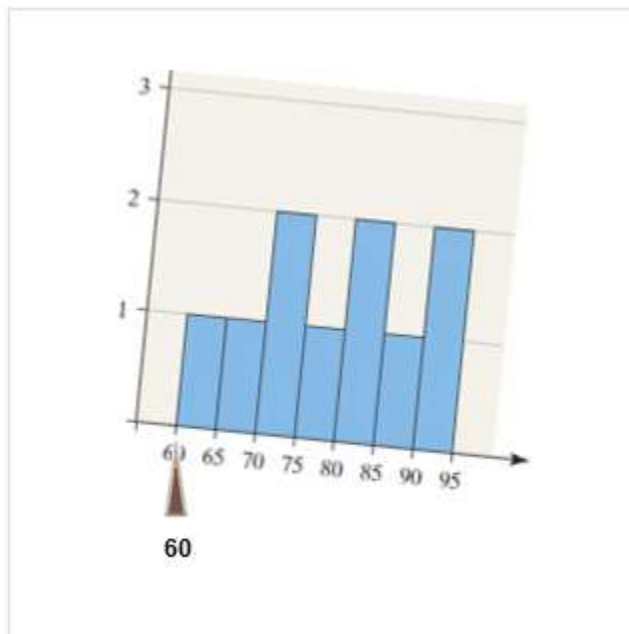


Figure 2

TABLE 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

3.1.21 Question

October 13, 2016 09:37 AM

✓ 3.1.21

1 of 1 Point

☰ Question Help



The following data represent the pulse rates (beats per minute) of nine students enrolled in a statistics course. Treat the nine students as a population. Complete parts (a) to (c).

Student	Pulse
Perpectual Bempah	66
Megan Brooks	73
Jeff Honeycutt	90
Clarice Jefferson	75
Crystal Kurtenbach	72
Janette Lantka	63
Kevin McCarthy	81
Tammy Ohm	75
Kathy Wojdya	90

(a) Determine the population mean pulse.

The population mean pulse is approximately 76.1 beats per minute. (Type an integer or decimal rounded to the nearest tenth as needed.)

(b) Determine the sample mean pulse of the following two simple random samples of size 3.

Sample 1: {Perpectual, Clarice, Tammy}

Sample 2: {Janette, Kevin, Perpectual}

The mean pulse of sample 1 is approximately 72 beats per minute. (Round to the nearest tenth as needed.)

The mean pulse of sample 2, is approximately 70 beats per minute. (Round to the nearest tenth as needed.)

(c) Determine if the means of samples 1 and 2 overestimate, underestimate, or are equal to the population mean.

The mean pulse rate of sample 1 underestimates the population mean.

The mean pulse rate of sample 2 underestimates the population mean.

3.1 Objective 2 Determine the Media of a Variable from Raw Data

October 13, 2016 09:39 AM

A second measure of central tendency is the *median*. To compute the median of a set of data, the data must be [quantitative](#).

DEFINITION

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

The next slide shows the steps for finding the median, M , of a data set by hand. It is important to understand how to find the median by hand, before using technology.

From <https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>

Steps in Finding the Median of a Data Set

Step 1 Arrange the data in ascending order.

Step 2 Determine the number of observations, n .

Step 3 Determine the observation in the middle of the data set.

- If the number of observations is odd, then the median is the data value exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{n+1}{2}$ position.
- If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the $\frac{n}{2}$ position and the $\frac{n}{2} + 1$ position.

EXAMPLE 2 Determining the Median of a Data Set (Odd Number of Observations)

Problem

Table 2 shows the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

Video Solution



Technology Step-By-Step



TABLE 2

Song Name	Length
"Sister Golden Hair"	201
"Black Water"	257
"Free Bird"	284
"The Hustle"	208
"Southern Nights"	179
"Stayin' Alive"	222
"We Are Family"	217

"Stayin' Alive"	222
"We Are Family"	217
"Heart of Glass"	206
"My Sharona"	240

Approach

Follow the [steps](#) for finding the median, M .

Solution

Step 1 Arrange the data in ascending order:

$$179, 201, 206, 208, 217, 222, 240, 257, 284$$

Step 2 There are $n = 9$ observations.

Step 3 Because n is odd, the median is the observation exactly in the middle of the data set with the data written in ascending order. This value lies in the $\frac{n+1}{2} = \frac{9+1}{2} = 5$ th position. The median appears in blue with four observations on each side of the median. So $M = 217$.

$$179, 201, 206, 208, \mathbf{217}, 222, 240, 257, 284$$

EXAMPLE 3 Determining the Median of a Data Set (Even Number of Observations)

Problem

Find the median score of the data in [Table 1](#).

Video Solution



Approach

Follow the [steps](#) for finding the median, M .

Solution

Step 1 Arrange the data in ascending order:

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$

Step 2 There are $n = 10$ observations.

Step 3 Because n is even, the median is the mean of the two middle observations, the fifth ($\frac{n}{2} = \frac{10}{2} = 5$) and sixth ($\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$) observations with the data written in ascending order. So the median is the mean of 77 and 82:

$$M = \frac{77 + 82}{2} = 79.5$$

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$



$$M = 79.5$$

Notice that there are five observations on each side of the median. We conclude that 50% (or half) of the students scored less than 79.5 and 50% (or half) scored above 79.5.

TABLE 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

3.1 Objective 3 Explain what It Means for a Statistic to be Resistant.

October 13, 2016 09:49 AM

You may be asking yourself, "Why would I ever compute the mean?" After all, the mean and median are close in value for symmetric data, and the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of statistical inference is based on the mean.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Based on Activity 1, we notice that the median is not affected by extreme observations, but the mean is affected by extreme observations. Because extreme values do not affect the value of the median, we say that the median is *resistant*.

DEFINITION

A numerical summary of data is said to be **resistant** if values that are extreme (very large or small) relative to the data do not affect its value substantially.

So the median is resistant, whereas the mean is not resistant.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

The ideas explored in Activity 2 are presented in Table 3 and Figure 3. A word of caution is in order. The relation between the mean, median, and skewness are guidelines. These guidelines tend to hold up well for continuous data, but when the data are discrete the rules can be easily violated.

TABLE 3

Relation among the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

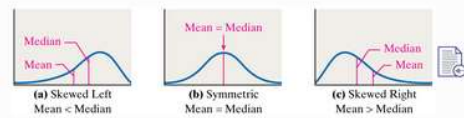
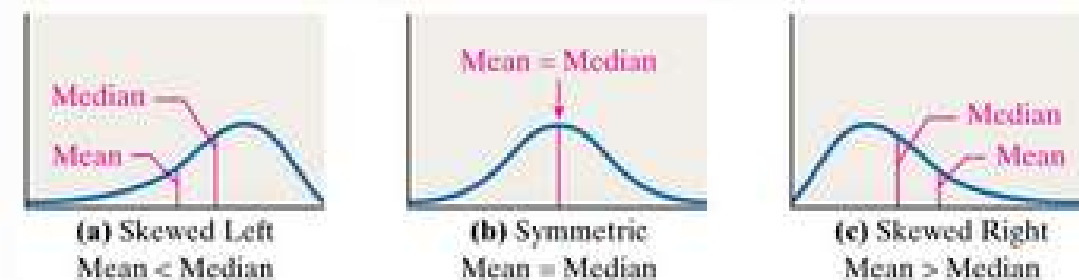


Figure 3 Mean and Median versus Skewness



EXAMPLE 4 Describing the Shape of a Distribution**Problem**

The data in Table 4 represent the birth weights (in pounds) of 50 randomly sampled babies.

- (a) Find the mean and median birth weight.
- (b) Describe the shape of the distribution.
- (c) Which measure of central tendency best describes the average birth weight?

TABLE 4

5.8	7.4	9.2	7.0	8.5	7.6
7.9	7.8	7.9	7.7	9.0	7.1
8.7	7.2	6.1	7.2	7.1	7.2
7.9	5.9	7.0	7.8	7.2	7.5
7.3	6.4	7.4	8.2	9.1	7.3
9.4	6.8	7.0	8.1	8.0	7.5
7.3	6.9	6.9	6.4	7.8	8.7
7.1	7.0	7.0	7.4	8.2	7.2
7.6	6.7				

Solution

Video Solution



HIDE SOLUTION

3.1 Objective 4 Determine the Mode of a Variable from Raw Data

October 13, 2016 10:13 AM

A third measure of central tendency is the mode, which can be computed for either [quantitative](#) or [qualitative](#) data.

DEFINITION

The **mode** of a variable is the observation of the variable that occurs most frequently in the data set.

- To compute the mode, tally the number of observations that occur for each data value.
- The data value that occurs most often is the mode.
- If no observation occurs more than once, we say that the data have **no mode**.
- A set of data can have no mode, one mode, or more than one mode

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 5 Finding the Mode of Quantitative Data

Problem

The following data represent the number of O-ring failures on the shuttle *Columbia* for the 17 flights prior to its fatal flight:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3

Find the mode number of O-ring failures.

Approach

Tally the number of times each data value occurs. The data value with the highest frequency is the mode.

Solution

The mode is 0 because it occurs most frequently (11 times).

EXAMPLE 6 Finding the Mode of Quantitative Data

Problem

Find the mode of the exam score data listed in [Table 1](#).

Approach

Tally the number of times each data value occurs. The data value with the highest frequency is the mode.

Solution

Because each data value occurs only once, there is no mode.

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Data Sets with More than One Mode

A data set can have more than one mode. For example, if the data in [Table 1](#) had two scores each of 77 and 88, then the data would have two modes: 77 and 88.

In this case, we would say that the data are **bimodal**. If a data set has three or more modes, then we say that the data are **multimodal**. The mode is usually not reported for multimodal data because it is not representative of a typical value. Figure 4(a) shows a distribution with one mode. Figure 4(b) shows a bimodal distribution.

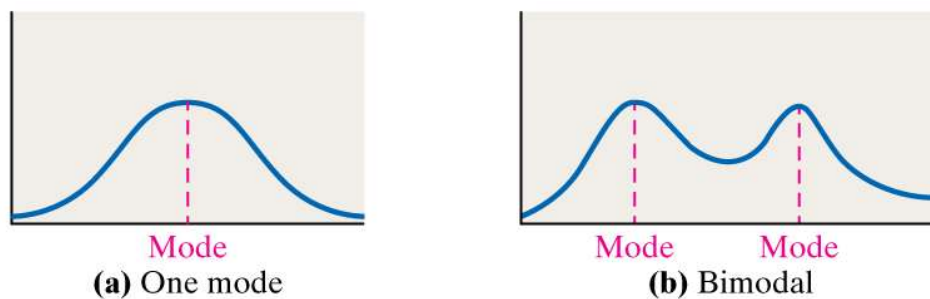


Figure 4

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 7 Finding the Mode of Qualitative Data**Problem**

The data in Table 5 represent the location of injuries that required rehabilitation by a physical therapist. Determine the mode location of injury.

TABLE 5

Back	Back	Hand	Neck	Knee	Knee
Wrist	Back	Groin	Shoulder	Shoulder	Back
Elbow	Back	Back	Back	Back	Back
Back	Shoulder	Shoulder	Knee	Knee	Back
Hip	Knee	Hip	Hand	Back	Wrist

Data from Krystal Catton, student at Joliet Junior College

**Approach**

Determine the location of injury that occurs with the highest frequency.

Solution

The mode location of injury is the back, with 12 instances.

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\sum x_i}{N}$ Sample mean: $\bar{x} = \frac{\sum x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and divide the data set in half	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or skewed right
Mode	Tally data to determine most frequent observation	Most frequent observation	When the most frequent observation is the desired measure of central tendency or the data are qualitative

3.2 Measures of Dispersion

October 15, 2016 08:34 AM

Example – Comparing Two Sets of Data

- The data in the following tables represent the IQ scores of a random sample of 100 students from two different universities.

University A

73	103	91	93	136	108	92	104	90	78
108	93	91	78	81	130	82	86	111	93
102	111	125	107	80	90	122	101	82	115
103	110	84	115	85	83	131	90	103	106
71	69	97	130	91	62	85	94	110	85
102	109	105	97	104	94	92	83	94	114
107	94	112	113	115	106	97	106	85	99
102	109	76	94	103	112	107	101	91	107
107	110	106	103	93	110	125	101	91	119
118	85	127	141	129	60	115	80	111	79

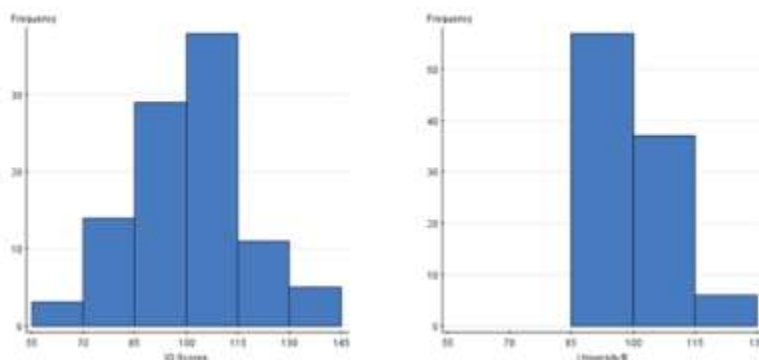
University B

86	91	107	94	105	107	89	96	102	96
92	109	103	106	98	95	97	95	109	109
93	91	92	91	117	108	89	95	103	109
110	88	97	119	90	99	96	104	98	95
87	105	111	87	103	92	103	107	106	97
107	108	89	96	107	107	96	95	117	97
98	89	104	99	99	87	91	105	109	108
116	107	90	98	98	92	119	96	118	98
97	106	114	87	107	96	93	99	89	94
104	88	99	97	106	107	112	97	94	107

Example – Comparing Two Sets of Data

For each university, compute the mean IQ score and draw a histogram, using a lower class limit of 55 for the first class and a class width of 15. Comment on the results.

Histograms



Summary of Results

Both universities have the same mean IQ.

The two histograms are quite different.

The IQs at University A are more dispersed, while the IQs at University B are grouped more closely to the mean.

3.2 Objective 1 Determine the Range of a Variable from Raw Data

October 15, 2016 08:35 AM

The simplest measure of dispersion is the *range*. To compute the range, the data must be [quantitative](#).

DEFINITION

The **range, R** , of a variable is the difference between the largest and smallest data value. That is,
Range = R = largest data value – smallest data value

R = Range

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Computing the Range of a Set of Data

Student	Score
1. Michelle	82
2. RYanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

The data in the table represent the first exam scores of 10 students enrolled in Introductory Statistics. Compute the range.

Solution

Student	Score
1. Michelle	82
2. RYanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

$R = \text{largest data value} - \text{smallest data value}$

$$R = 94 - 62 = 32$$

Range is Not Resistant

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62 28
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

$$R = 94 - 28 = 66$$

3.2 Objective 2 Determine the Standard Deviation of a Variable from Raw Data

October 15, 2016 08:43 AM

DEFINITION

The **population standard deviation** of a variable is the **square root** of the sum of squared deviations about the population mean divided by the number of observations in the population, N .

In other words, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

where x_1, x_2, \dots, x_N are the N observations in the population and μ is the population mean.

EXAMPLE 3 Computing a Population Standard Deviation

Problem

Compute the population standard deviation of the test scores in Table 6.

Video Solution



Technology Step-By-Step



TABLE 6

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Approach

Step 1 Create a table with four columns. Enter the population data in Column 1. In Column 2, enter the population mean. When computing the population standard deviation, be sure to use μ with as many decimal places as possible to avoid round-off error.

Step 2 In Column 3, compute the deviation about the mean for each data value, $x_i - \mu$.

Step 3 In Column 4, enter the squares of the values in Column 3.

Step 4 Sum the entries in Column 4 and divide this result by the size of the population, N .

Step 5 Determine the square root of the value found in Step 4.

Solution

Step 1 See Table 7. Column 1 lists the observations in the data set, and Column 2 lists the population mean.

TABLE 7

Score, x_i	Population Mean, μ
82	79
77	79
90	79
71	79
62	79
68	79
74	79
84	79
94	79
88	79

Step 2 Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Michelle is $82 - 79 = 3$. It is a good idea to add the entries in this column to make sure they sum to 0.

Step 2 Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Michelle is $82 - 79 = 3$. It is a good idea to add the entries in this column to make sure they sum to 0.

TABLE 7

Score, x_i	Population Mean, μ	Deviation about the Mean, $x_i - \mu$
82	79	$82 - 79 = 3$
77	79	$77 - 79 = -2$
90	79	11
71	79	-8
62	79	-17
68	79	-11
74	79	-5
84	79	5
94	79	15
88	79	9
		$\sum (x_i - \mu) = 0$

Step 3 Column 4 shows the squared deviations about the mean.

TABLE 7

Score, x_i	Population Mean, μ	Deviation about the Mean, $x_i - \mu$	Squared Deviation about the Mean, $(x_i - \mu)^2$
82	79	$82 - 79 = 3$	$3^2 = 9$
77	79	$77 - 79 = -2$	$(-2)^2 = 4$
90	79	11	121
71	79	-8	64
62	79	-17	289
68	79	-11	121
74	79	-5	25
84	79	5	25
94	79	15	225
88	79	9	81
		$\sum (x_i - \mu) = 0$	$\sum (x_i - \mu)^2 = 964$

Step 4 Sum the entries in Column 4 to obtain the **numerator** of the **formula** for the population standard deviation. Divide this sum by the number of students, **10**:

$$\frac{\sum (x_i - \mu)^2}{N} = \frac{964}{10} = 96.4 \text{ points}^2$$

Step 5 The square root of the result in Step 4 is the population standard deviation.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{96.4 \text{ points}^2} \approx 9.8 \text{ points}$$

Look at [Table 7](#). The further an observation is from the mean, 79, the larger the squared deviation. For example, because the second observation, 77, is not “far” from 79, the squared deviation, 4, is not large. However, the fifth observation, 62, is further from 79, so the squared deviation, 289, is much larger.

If a data set has many observations that are “far” from the mean, then the sum of the squared deviations will be large and the standard deviation will be large.

Score, x_i	Population Mean, μ	Deviation about the Mean, $x_i - \mu$	Squared Deviation about the Mean, $(x_i - \mu)^2$
82	79	$82 - 79 = 3$	$3^2 = 9$
77	79	$77 - 79 = -2$	$(-2)^2 = 4$
90	79	11	121
71	79	-8	64
62	79	-17	289
68	79	-11	121
74	79	-5	25
84	79	5	25
94	79	15	225
88	79	9	81
		$\sum (x_i - \mu) = 0$	$\sum (x_i - \mu)^2 = 964$

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

DEFINITION

The **sample standard deviation**, s , of a variable is the square root of the sum of squared deviations about the sample mean divided by $n - 1$, where n is the sample size.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

where x_1, x_2, \dots, x_n are the n observations in the sample and \bar{x} is the sample mean.

x-bar (x with horizontal bar above) = Sample Mean
n = Sample

Σ = Sum

X_i = Dataset

Why Do We Divide by $n - 1$? Understanding Degrees of Freedom

To find the sample standard deviation, we divide by $n - 1$. Showing why we divide by $n - 1$ is beyond the scope of the course. However, the following explanation has intuitive appeal. We already know that the sum of the deviation about the mean, $\Sigma (x_i - \bar{x})$, must equal zero. Therefore, if the sample mean is known and the first $n - 1$ observations are known, then the n th observation must be the value that causes the sum of the deviations to equal zero. For example, suppose $\bar{x} = 4$ is based on a sample of size $n = 3$. If $x_1 = 2$ and $x_2 = 3$, then we can determine x_3 as follows:

$$\begin{aligned}\frac{x_1 + x_2 + x_3}{n} &= \bar{x} \\ \frac{2 + 3 + x_3}{3} &= 4 \\ 5 + x_3 &= 12 \\ x_3 &= 7\end{aligned}$$

We call $n - 1$ the **degrees of freedom** because the first $n - 1$ observations have freedom to be any value, but the n th observation has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

IN
OTHER
WORDS



Simple Random Sample:

Student	Score
Bilal	90
Ryanne	77
Pam	71
Michelle	82

Sample Mean:

$$\sum x_i = x_1 + x_2 + x_3 + x_4$$

$$= 90 + 77 + 71 + 82$$

$$= 320$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{320}{4} = 80$$

EXAMPLE 4 Computing a Sample Standard Deviation

Problem

In a [previous lesson](#) we obtained a simple random sample of exam scores and computed a sample mean of 80. Compute the sample standard deviation of the sample of test scores for that data.

Video Solution



Technology Step-By-Step



Approach

Follow the same approach used to compute the population standard deviation, but this time use the sample data:

Bilal (90), Ryanne (77), Pam (71), and Michelle (82)

When computing the sample standard deviation, be sure to use \bar{x} with as many decimal places as possible to avoid round-off error. However, report the standard deviation to one more decimal place than the original data. For this example, report the standard deviation to the tenths place.

Solution

Step 1 Create a table with four columns. Enter the sample data in Column 1. In Column 2, enter the sample mean. See Table 8.

TABLE 8

Score, x_i	Sample Mean, \bar{x}
90	80
77	80
71	80
82	80

Step 2 Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Bilal is $90 - 80 = 10$. It is a good idea to add the entries in this column to make sure they sum to 0.

TABLE 8

Score, x_i	Sample Mean, \bar{x}	Deviation about the Mean, $x_i - \bar{x}$
90	80	$90 - 80 = 10$
77	80	-3
71	80	-9
82	80	2

$$\sum (x_i - \bar{x}) = 0$$

Step 3 Column 4 shows the squared deviations about the mean.

TABLE 8

Score, x_i	Sample Mean, \bar{x}	Deviation about the Mean, $x_i - \bar{x}$	Squared Deviation about the Mean, $(x_i - \bar{x})^2$
90	80	$90 - 80 = 10$	$10^2 = 100$
77	80	-3	9
71	80	-9	81
82	80	2	4
		$\sum (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 194$

Step 4 Sum the entries in Column 4 to obtain the numerator of the formula for the sample standard deviation. Divide the sum of the entries in Column 4 by $n - 1 = 4 - 1 = 3$.

$$\frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{194}{4 - 1} = 64.7 \text{ points}^2$$

Step 5 Find the square root of the result in Step 4 to obtain the sample standard deviation.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{64.7 \text{ points}^2} \approx 8.0 \text{ points}$$

Remember, round the sample standard deviation to one more decimal place than the raw data.

Standard Deviation: Not Resistant

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

The data in the table represent the first exam scores of 10 students enrolled in Introductory Statistics.

A random sample of 4 students (Sam, Pam, Bilal, and Michelle) is selected.

If Michelle's score is incorrectly recorded as 28 instead of 82, find the standard deviation for this sample.

Interpretations of the Standard Deviation

The standard deviation is used along with the mean to describe symmetric distributions numerically. The mean measures the *center* of the distribution, whereas the standard deviation measures the *spread* of the distribution. So how does the value of the standard deviation relate to the spread of the distribution?

If we are comparing two populations, **the larger the standard deviation, the greater the dispersion, or spread, of the distribution** as long as the variable of interest from the two populations has the same

unit of measure. The units of measure must be the same so that we are comparing “apples with apples.” For example, \$100 is not the same value as 100 Japanese yen (because recently \$1 was equivalent to about 102 yen). So a standard deviation of \$100 is substantially higher than a standard deviation of 100 yen.

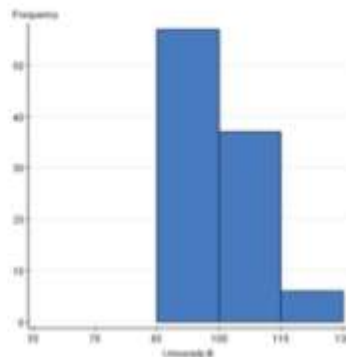
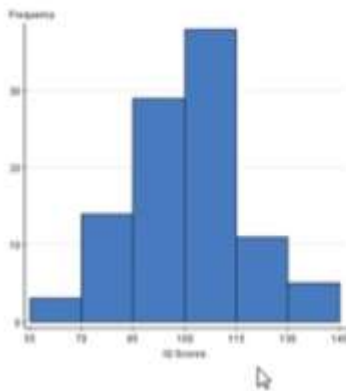
From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Example – Comparing the Standard Deviation of Two Data Sets

The data in the following tables represent the IQ scores of a random sample of 100 students from two different universities.

Use the standard deviation to determine whether University A or University B has more dispersion in the IQ scores of its students.

Histograms



3.2 Objective 3 Determine the Variance of a Variable from Raw Data

October 15, 2016 09:37 AM

Up to this point, we have studied two measures of dispersion – the **range** and the **standard deviation**. A third measure of dispersion is called the **variance**.

DEFINITION

The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 , and the **sample variance** is s^2 .

The units of measure in variance are squared values. So if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult.

EXAMPLE 6 Determining the Variance of a Variable for a Population and a Sample

In previous examples, we considered **population data** of exam scores in a statistics class. For this data, we computed a population mean of $\mu = 79$ points and a population standard deviation of $\sigma = 9.8$ points. Then, we obtained a **simple random sample** of exam scores. For this data, we computed a sample mean of $\bar{x} = 80$ points and a sample standard deviation of $s = 8.0$ points. Use the population standard deviation exam score and the sample standard deviation exam score to determine the population and sample variance of scores on the statistics exam.

Approach

The population variance is found by squaring the population standard deviation. The sample variance is found by squaring the sample standard deviation.

Solution

The population standard deviation is $\sigma = 9.8$ points, so the population variance is $\sigma^2 = (9.8 \text{ points})^2 = 96.04 \text{ points}^2$. The sample standard deviation is $s = 8.0$ points, so the sample variance is $s^2 = (8.0 \text{ points})^2 = 64.0 \text{ points}^2$.

NOTE

Be sure to include the units *squared* when reporting the variance.

Population Data

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Simple Random Sample

Student	Test Score
Bilal	90
Ryanne	77
Pam	71
Michelle	82

Bias in the Variance and Standard Deviation

The sample variance is obtained using the formula $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. What if we divided by n instead of $n - 1$ to obtain the sample variance, as one might expect? Then the sample variance would consistently underestimate the population variance. Whenever a statistic consistently underestimates a parameter, it is said to be biased. To obtain an unbiased estimate of the population variance, we divide the sum of the squared deviations about the sample mean by $n - 1$.

Let's look at an example of a biased estimator. Suppose you work for a carnival, guessing people's ages. After 20 people have come to your booth, you notice that you have a tendency to underestimate people's ages, or guess too low. What could you do to correct this? You could adjust your guesses higher to avoid underestimating. In other words, originally your guesses were biased. To remove the bias, you increase your guesses. This is what dividing by $n - 1$ in the sample variance formula accomplishes.

Unfortunately, the sample standard deviation given by the formula $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ is not an unbiased estimate of the population standard deviation. In fact, it is not possible to provide an unbiased estimator of the population standard deviation for all distributions. The explanation is beyond the scope of this class (it has to do with the shape of the square root function). However, for the applications in this text, the bias is minor and does not impact results.

3.2 Objective 4 Use the Empirical Rule to Describe Data That Are Bell-Shaped

October 15, 2016 09:43 AM

If data have a distribution that is bell-shaped, the *Empirical Rule* can be used to determine the percentage of data that will lie within k standard deviations of the mean.

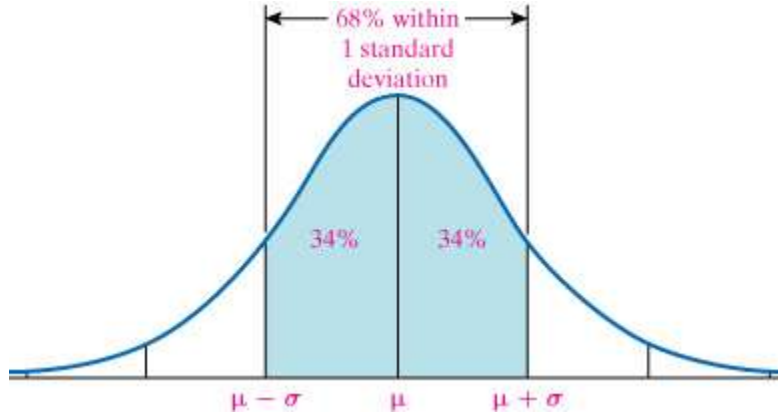
The Empirical Rule

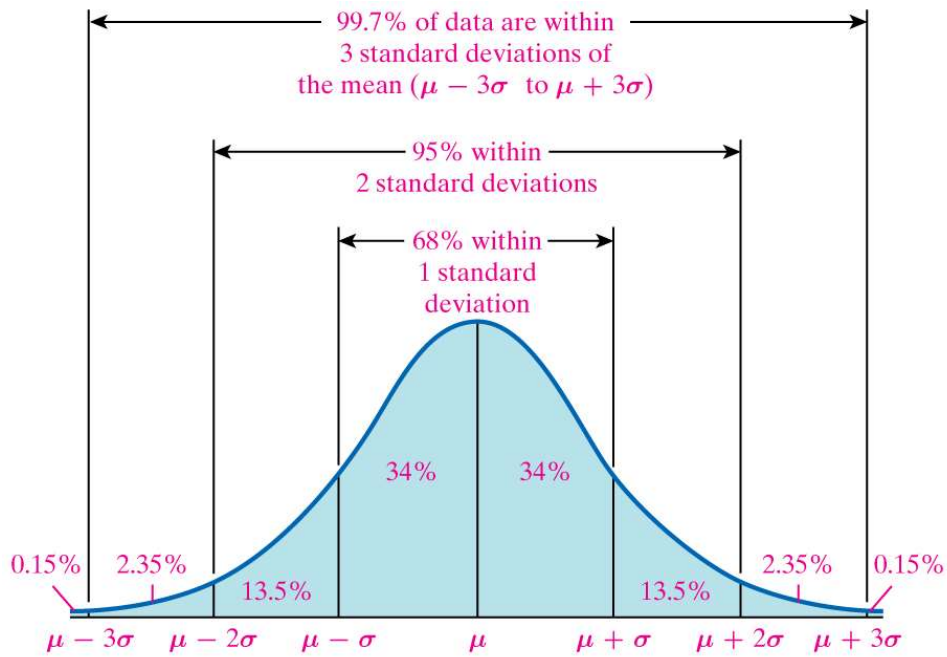
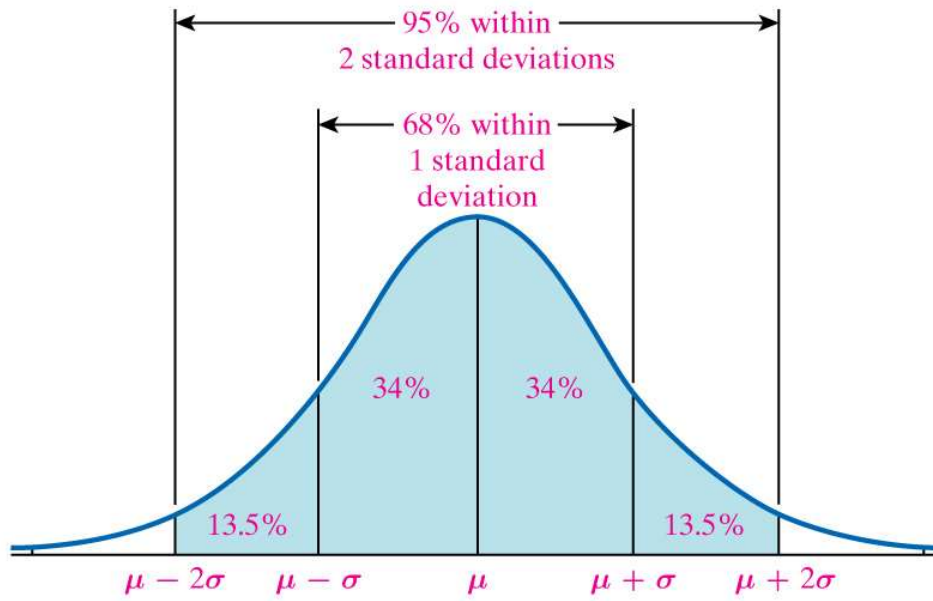
If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data will lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

NOTE

The Empirical Rule can also be used based on sample data with \bar{x} in place of μ and s in place of σ .





EXAMPLE 7 Using the Empirical Rule

Problem

Table 9 represents the IQs of a random sample of 100 students at a university.

Video Solution



- Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.
- Determine the percentage of students who have IQ scores between 67.8 and 132.2 according to the Empirical Rule.
- Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.
- According to the Empirical Rule, what percentage of students have IQ scores between 116.1 and 148.3?

TABLE 9

73	103	91	93	136	108	92	104	90	78
108	93	91	78	81	130	82	86	111	93
102	111	125	107	80	90	122	101	82	115
103	110	84	115	85	83	131	90	103	106
71	69	97	130	91	62	85	94	110	85
102	109	105	97	104	94	92	83	94	114
107	94	112	113	115	106	97	106	85	99
102	109	76	94	103	112	107	101	91	107
107	110	106	103	93	110	125	101	91	119
118	85	127	141	129	60	115	80	111	79



Approach

Problems (a) through (d) can all be answered using the Empirical Rule provided a histogram of the data is roughly bell-shaped. Therefore, we begin by drawing a histogram of the data to be sure it satisfies the bell-shaped criterion.

The histogram of the IQ scores in Figure 6 is roughly bell-shaped. The mean IQ scores of the students is 100, and the standard deviation is 16.1.

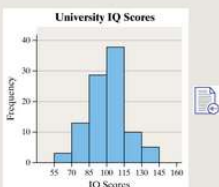


Figure 6

Solution

Part A

Part B

Part C

Part D

To help organize our thoughts and make the analysis easier, we draw a bell-shaped curve like the one in Figure 5 with $\bar{x} = 100$ and $s = 16.1$. See Figure 7.

According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean [that is, greater than or equal to $100 - 3(16.1) = 51.7$ and less than or equal to $100 + 3(16.1) = 148.3$].

Part A Part B Part C Part D

Because 67.8 is exactly 2 standard deviations below the mean [$100 - 2(16.1) = 67.8$] and 132.2 is exactly 2 standard deviations above the mean [$100 + 2(16.1) = 132.2$], the Empirical Rule tells us that approximately 95% of all IQ scores lie between 67.8 and 132.2. See Figure 7.

Part A Part B Part C Part D

Of the 100 IQ scores listed in Table 9, a total of 96, or 96%, are between 67.8 and 132.2. This is very close to the Empirical Rule's approximation.

Part A Part B Part C Part D

Based on Figure 7, approximately $13.5\% + 2.35\% = 15.85\%$ of the students have IQ scores between 116.1 and 148.3.

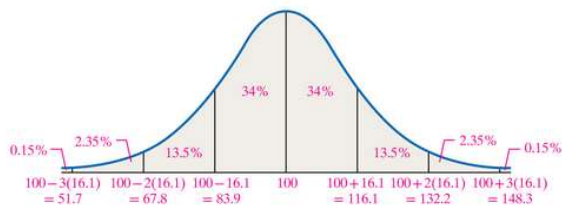


Figure 7

3.2.35x - Chebyshev's inequality

October 15, 2016 10:37 AM

At one point the average price of regular unleaded gasoline was \$3.55 per gallon. Assume that the standard deviation price per gallon is \$0.09 per gallon and use Chebyshev's inequality to answer the following.

- (a) What percentage of gasoline stations had prices within 2 standard deviations of the mean?
(b) What percentage of gasoline stations had prices within 1.5 standard deviations of the mean? What are the gasoline prices that are within 1.5 standard deviations of the mean?
(c) What is the minimum percentage of gasoline stations that had prices between \$3.19 and \$3.91?

(a) To answer this question, use Chebyshev's inequality, which says that for any data set, regardless of the shape of the distribution, at least $\left(1 - \frac{1}{k^2}\right)$ 100% of the observations will lie within k standard deviations of the mean, where k is any number greater than 1.

So use Chebyshev's inequality with $k = 2$.

$$\left(1 - \frac{1}{k^2}\right) 100\% = \left(1 - \frac{1}{2^2}\right) 100\% = 75\%$$

(Round to the nearest hundredth as needed.)

Therefore, at least 75% of gasoline stations had prices within 2 standard deviations of the mean.

(b) Again, to find the percentage of gasoline stations that had prices within 1.5 standard deviations of the mean, use Chebyshev's inequality, this time with $k = 1.5$.

$$\left(1 - \frac{1}{k^2}\right) 100\% = \left(1 - \frac{1}{1.5^2}\right) 100\% = 55.56\%$$

(Round to the nearest hundredth as needed.)

Now find the gasoline prices that are within 1.5 standard deviations of the mean. To find the minimum price, calculate $\mu - 1.5\sigma$.

$$\mu - 1.5\sigma = 3.55 - 1.5(0.09) = \$ 3.415$$

To find the maximum price, calculate $\mu + 1.5\sigma$.

$$\mu + 1.5\sigma = 3.55 + 1.5(0.09) = \$ 3.685$$

Therefore, at least 55.56% of gasoline stations had prices within 1.5 standard deviations of the mean, which corresponds to prices from \$3.415 to \$3.685.

(c) To find the minimum percentage of gasoline stations that had prices between \$3.19 and \$3.91, first determine how many standard deviations these prices are from the mean, \$3.55.

To determine the number of standard deviations between \$3.19 and \$3.55, take the difference between the two values, $3.55 - 3.19 = 0.36$, and divide by the standard deviation.

$$0.36/0.09 = 4 \text{ standard deviations}$$

Similarly, the difference between \$3.55 and \$3.91 is \$0.36. Thus, \$3.91 is also 4 standard deviations from the mean.

Now use Chebyshev's inequality with $k = 4$ to find the percent of observations that lie within k standard deviations of the mean.

$$\left(1 - \frac{1}{k^2}\right) 100\% = \left(1 - \frac{1}{4^2}\right) 100\% = 93.75\%$$

(Round to the nearest hundredth as needed.)

Therefore, 93.75% is the minimum percentage of gasoline stations that had prices between \$3.19 and \$3.91.

3.3 Measures of Central Tendency and Dispersion from Grouped Data

October 16, 2016 11:03 AM

3.3 Objective 1 Approximate the Mean of a Variable from Grouped Data

October 16, 2016 11:24 AM

Because raw data cannot be retrieved from a frequency table, we assume that within each class, the mean of the data values is equal to the **class midpoint**. We then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. We repeat the process for each class and add the results. This sum approximates the sum of all the data. The **video** explains the formulas.

DEFINITIONS Approximate the Mean of a Variable from a Frequency Distribution

Population Mean

$$\begin{aligned} \mu &= \frac{\sum x_i f_i}{\sum f_i} \\ &= \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} \end{aligned}$$

Sample Mean

$$\begin{aligned} \bar{x} &= \frac{\sum x_i f_i}{\sum f_i} \\ &= \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} \end{aligned}$$

where: x_i is the midpoint or value of the i th class
 f_i is the frequency of the i th class
 n is the number of classes

In each formula, $x_1 f_1$ approximates the sum of all the data values in the first class, $x_2 f_2$ approximates the sum of all the data values in the second class, and so on. Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.

EXAMPLE 1 Approximating the Mean for Continuous Quantitative Data from a Frequency Distribution

Problem

The frequency distribution in Table 10 represents the five-year rate of return of a random sample of 40 large-blend mutual funds. Approximate the mean five-year rate of return.

Solution

Video Solution Technology Step-by-Step



HIDE SOLUTION

Result

The approximate mean 5-year rate of return is 5.2%. The mean 5-year rate of return from the **raw data** is 5.194%. Click the link to view a complete **by-hand solution**.

HIDE RESULT

TABLE 10

Class (five-year rate of return)	Frequency
3–3.99	16
4–4.99	13
5–5.99	4
6–6.99	1
7–7.99	0
8–8.99	1
9–9.99	0
10–10.99	2
11–11.99	2
12–12.99	1

11-11.99	2
12-12.99	1

Example – Approximating the Mean for Continuous Quantitative Data from a Frequency Distribution

The following frequency distribution represents the 5-year rate of return of a random sample of 40 large-blended mutual funds.

Approximate the mean 5-year rate of return.

Class (5-year rate of return)	Frequency
3 - 3.99	16
4 - 4.99	13
5 - 5.99	4
6 - 6.99	1
7 - 7.99	0
8 - 8.99	1
9 - 9.99	0
10 - 10.99	2
11 - 11.99	2
12 - 12.99	1

Solution

Class	Frequency	Midpoint	$x_i f_i$
3 - 3.99	16	$\frac{3+4}{2} = 3.5$	$16(3.5) = 56$
4 - 4.99	13	4.5	58.5
5 - 5.99	4	5.5	22
6 - 6.99	1		
7 - 7.99	0		
8 - 8.99	1		
9 - 9.99	0		
10 - 10.99	2		
11 - 11.99	2		
12 - 12.99	1		

Class	Frequency	Midpoint	$x_i f_i$
3 - 3.99	16	3.5	56
4 - 4.99	13	4.5	58.5
5 - 5.99	4	5.5	22
6 - 6.99	1	6.5	6.5
7 - 7.99	0	7.5	0
8 - 8.99	1	8.5	8.5
9 - 9.99	0	9.5	0
10 - 10.99	2	10.5	21
11 - 11.99	2	11.5	23
12 - 12.99	1	12.5	<u>12.5</u>

Solution

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

$$\bar{x} = \frac{208}{40} = 5.2$$

3.3 Objective 2 Compute the Weighted Mean

October 16, 2016 11:35 AM

OBJECTIVE 2 Compute the Weighted Mean

When data values have different importance, or *weights*, associated with them, we compute the *weighted mean*. For example, grade point average is a weighted mean, with weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value. The [video](#) explains the formula for obtaining a weighted mean.

DEFINITION

The **weighted mean**, \bar{x}_w , of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}$$

where w_i is the weight of the i th observation
 x_i is the value of the i th observation

EXAMPLE 2 Computing the Weighted Mean

Problem

Marissa just completed her first semester in college. She earned an A in her 4-hour statistics course, a B in her 3-hour sociology course, an A in her 3-hour psychology course, a C in her 5-hour computer programming course, and an A in her 1-hour drama course. Determine Marissa's grade point average.

Video Solution



Technology Step-By-Step



Approach

Assign a point value to each grade. Let an A equal 4 points, a B equal 3 points, and a C equal 2 points. The number of credit hours for each course determines its weight. So a 5-hour course gets a weight of 5, a 4-hour course gets a weight of 4, and so on. Multiply the weight of each course by the points earned in the course, add these products, and divide this sum by the sum of the weights, number of credit hours.

Solution

$$\text{GPA} = \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{4(4) + 3(3) + 3(4) + 5(2) + 1(4)}{4 + 3 + 3 + 5 + 1} = \frac{51}{16} = 3.19$$

Marissa's grade-point average for her first semester is **3.19**.

A = 4 points

B = 3 points

C = 2 points

3.3 Objective 3 Approximate the Standard Deviation from a Frequency Distribution

October 16, 2016 11:47 AM

The procedure for approximating the standard deviation from grouped data is similar to that of finding the mean from grouped data. Because we do not have access to the original data, the standard deviation is approximate. The [video](#) explains the formulas.

DEFINITIONS Approximate Standard Deviation of a Variable from a Frequency Distribution

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2 f_i}{\sum f_i}}$$

Sample Standard Deviation

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

where: x_i is the midpoint or value of the i th class
 f_i is the frequency of the i th class
 n is the number of classes

EXAMPLE 3 Approximating the Standard Deviation from a Frequency Distribution

Problem

The frequency distribution in Table 11 represents the five-year rate of return of a random sample of 40 large-blend mutual funds. Approximate the standard deviation five-year rate of return.

Solution

Video Solution Technology Step-by-Step



HIDE SOLUTION

Result

The approximate standard deviation of the five-year rate of return is 2.55%. The standard deviation of the five-year rate of return from the [raw data](#) is 2.64%. The approximation is close to the standard deviation from the raw data. Click the link to view a complete [by-hand solution](#).

HIDE RESULT

TABLE 11

Class (five-year rate of return)	Frequency
3–3.99	16
4–4.99	13
5–5.99	4
6–6.99	1
7–7.99	0
8–8.99	1
9–9.99	0
10–10.99	2
11–11.99	2
12–12.99	1

Example – Approximating the Standard Deviation from a Frequency Distribution

The following frequency distribution represents the 5-year rate of return of a random sample of 40 large-blended mutual funds.

Approximate the standard deviation of the 5-year rate of return.

Class (5-year rate of return)	Frequency
3 - 3.99	16
4 - 4.99	13
5 - 5.99	4
6 - 6.99	1
7 - 7.99	0
8 - 8.99	1
9 - 9.99	0
10 - 10.99	2
11 - 11.99	2
12 - 12.99	1

Class	Midpt. x_i	Freq. f_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
3 - 3.99	3.5	16	5.2	$3.5 - 5.2 = -1.7$	
4 - 4.99	4.5	13	5.2	$4.5 - 5.2 = -0.7$	
5 - 5.99	5.5	4	5.2		
6 - 6.99	6.5	1	5.2		
7 - 7.99	7.5	0	5.2		
8 - 8.99	8.5	1	5.2		
9 - 9.99	9.5	0	5.2		
10 - 10.99	10.5	2	5.2		
11 - 11.99	11.5	2	5.2		
12 - 12.99	12.5	1	5.2		

Class	Midpt. x_i	Freq. f_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
3 - 3.99	3.5	16	5.2	-1.7	$(-1.7)^2 \cdot 16$
4 - 4.99	4.5	13	5.2	-0.7	$(-0.7)^2 \cdot 13$
5 - 5.99	5.5	4	5.2	0.3	$(0.3)^2 \cdot 4$
6 - 6.99	6.5	1	5.2	1.3	$(1.3)^2 \cdot 1$
7 - 7.99	7.5	0	5.2	2.3	$(2.3)^2 \cdot 0$
8 - 8.99	8.5	1	5.2	3.3	$(3.3)^2 \cdot 1$
9 - 9.99	9.5	0	5.2	4.3	$(4.3)^2 \cdot 0$
10 - 10.99	10.5	2	5.2	5.3	$(5.3)^2 \cdot 2$
11 - 11.99	11.5	2	5.2	6.3	$(6.3)^2 \cdot 2$
12 - 12.99	12.5	1	5.2	7.3	$(7.3)^2 \cdot 1$

Class	Midpt. x_i	Freq. f_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
3 - 3.99	3.5	16	5.2	-1.7	$(-1.7)^2 \cdot 16$ †
4 - 4.99	4.5	13	5.2	-0.7	$(-0.7)^2 \cdot 13$ †
5 - 5.99	5.5	4	5.2	0.3	$(0.3)^2 \cdot 4$
6 - 6.99	6.5	1	5.2	1.3	$(1.3)^2 \cdot 1$
7 - 7.99	7.5	0	5.2	2.3	$(2.3)^2 \cdot 0$
8 - 8.99	8.5	1	5.2	3.3	$(3.3)^2 \cdot 1$
9 - 9.99	9.5	0	5.2	4.3	$(4.3)^2 \cdot 0$
10 - 10.99	10.5	2	5.2	5.3	$(5.3)^2 \cdot 2$
11 - 11.99	11.5	2	5.2	6.3	$(6.3)^2 \cdot 2$
12 - 12.99	12.5	<u>1</u>	5.2	7.3	<u>$(7.3)^2 \cdot 1$</u> =
		40			254.4

Solution

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot f_i}{(\sum f_i) - 1}}$$

$$s = \sqrt{\frac{254.4}{40 - 1}}$$

$$s \approx 2.55$$

4

3.3.11

October 16, 2016 01:06 PM

Charlene and Gary want to make confetti. In order to get the right balance of ingredients for their tastes they bought 3 pounds of paper hearts at \$4.78 per pound, 5 pounds of sparkling stars for \$2.84 per pound, and 2 pounds of shiny coils for \$3.73 per pound. Determine the cost per pound of the confetti.

Finding the cost per pound of confetti is equivalent to finding the weighted mean of cost per pound over all the pounds of confetti.

The weighted mean of a variable is found by multiplying each value of the variable by its corresponding weight, summing these products, and dividing the result by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{x_1 w_1 + x_2 w_2 + \dots + x_n w_n}{w_1 + w_2 + \dots + w_n}$$

where x_i is the midpoint or value of the i th class, w_i is the weight on the i th class, and n is the number of classes.

Fill the second row with cost per pound. Complete the third row with number of pounds.

Course	paper hearts	sparkling stars	shiny coils
Cost per pound, x_i	\$ 4.78	\$ 2.84	\$ 3.73
Number of pounds, w_i	3	5	2

Now, total the number of pounds in the third row.

$$\sum w_i = 3 + 5 + 2 = 10$$

For the weighted mean formula, multiply each x_i by the corresponding w_i .

$$\begin{aligned}\sum x_i w_i &= (\$4.78 \cdot 3) + (\$2.84 \cdot 5) + (\$3.73 \cdot 2) \\ &= \$ 14.34 + \$ 14.2 + \$ 7.46 \text{ (Round to the nearest cent as needed.)}\end{aligned}$$

Find the sum of the $x_i w_i$.

$$\sum x_i w_i = \$ 36 \text{ (Round to the nearest cent as needed.)}$$

Now use the formula given earlier to compute the weighted mean.

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \$ 3.6 \text{ (Round to the nearest cent as needed.)}$$

Therefore, the confetti costs \$3.60 per pound.

Question is complete.

3.4 Measures of Position

October 17, 2016 11:03 AM

3.4 Objective 1 Determine and Interpret z-Scores

October 17, 2016 11:03 AM

Z-score

Population z-score	Sample z-score
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$

For the 2013 baseball season, the Boston Red Sox led the American League with 853 runs scored, whereas the St. Louis Cardinals led the National League with 783 runs scored. It appears that the Red Sox were the better run-producing team. However, this comparison is unfair because the teams play in different leagues. The Red Sox play in the American League, where the designated hitter bats for the pitcher, whereas the Cardinals play in the National League, where the pitcher must bat (pitchers are typically poor hitters). To compare the teams' runs scored, we must determine their relative standings within their respective leagues. We can do this using a *z*-score. The [video](#) explains the formula.

DEFINITION

The *z*-score represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation.

Population z-score

$$z = \frac{x - \mu}{\sigma}$$

Sample z-score

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

The *z*-score is unitless. It has mean 0 and standard deviation 1.

If a data value is larger than the mean, the *z*-score is positive. If a data value is smaller than the mean, the *z*-score is negative. If the data value equals the mean, the *z*-score is zero. A *z*-score measures the number of standard deviations an observation is above or below the mean. For example, a *z*-score of 1.24 means the data value is 1.24 standard deviations above the mean. A *z*-score of -2.31 means the data value is 2.31 standard deviations below the mean.

NOTE

Round *z*-scores to the nearest hundredth.

μ = Population Mean

\bar{x} (x with horizontal bar above) = Sample Mean

N = Population

n = Sample

Σ = Sum

X_i = Dataset

M = Median (not covered in this chapter, but still useful to know)

EXAMPLE 1 Comparing z -scores

Problem

Determine whether the Boston Red Sox or the St. Louis Cardinals had a relatively better run-producing season. The Red Sox scored 853 runs and play in the American League, where the mean number of runs scored was $\mu = 701.7$ and the standard deviation was $\sigma = 60.5$ runs. The Cardinals scored 783 runs and play in the National League, where the mean number of runs scored was $\mu = 648.7$ and the standard deviation was $\sigma = 75.7$ runs.

Video Solution



Approach

The word "relatively" suggests we are comparing two data sets. Therefore, compute each team's z -score. The team with the higher z -score had the better season. Because we know the values of the population parameters, compute the population z -score.

Solution

Compute each team's z -score, rounded to two decimal places.

$$\text{Red Sox's } z\text{-score} = \frac{x - \mu}{\sigma} = \frac{853 - 701.7}{60.5} = 2.50$$

$$\text{Cardinals' s } z\text{-score} = \frac{x - \mu}{\sigma} =$$

So the Red Sox had a run production 2.50 standard deviations above the mean, whereas the Cardinals had a run production 1.77 standard deviations above the mean. Therefore, the Red Sox had a relatively better year at scoring runs than did the Cardinals.

Interpreting Negative z -Scores

Example 1 dealt with positive z -scores. With negative z -scores, we need to be careful when deciding the better outcome. For example, suppose Bob and Mary ran a marathon. Bob finished the marathon in 213 minutes, where the mean finishing time among all men was 241 minutes with a standard deviation of 57 minutes. Mary finished the marathon in 241 minutes, where the mean finishing time among all women was 273 minutes with a standard deviation of 52 minutes.

Who did better in the race?

$$\text{Bob's } z\text{-score is } z_{\text{Bob}} = \frac{213 - 241}{57} = -0.49$$

$$\text{Mary's } z\text{-score is } z_{\text{Mary}} =$$

Therefore, Mary did better. Even though Bob's z -score is larger, Mary did better because her time is more standard deviations below the mean.

3.4 Objective 2 Interpret Percentiles

October 17, 2016 11:13 AM

Recall that the **median** divides the lower 50% of a data set from the upper 50%. The median is a special case of a general concept called the *percentile*.

Definition

The k th percentile, denoted P_k , of a set of data is a value such that k percent of the observations are less than or equal to the value.

Percentiles divide a set of data, written in ascending order, into 100 parts such that 99 percentiles can be determined. For example, P_1 divides the bottom 1% of the observations from the top 99%, P_2 divides the bottom 2% of the observations from the top 98%, and so on. Figure 8 displays the 99 possible percentiles.



Figure 8

Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT, use percentiles to let students know how they scored on the exam in relation to all others who took the exam.



EXAMPLE 2 Interpreting a Percentile

Problem

Jennifer just received the results of her SAT exam. Her math score of 600 is at the 74th percentile. Interpret this result.

Approach

The fact that an observation is at the k th percentile means that k percent of the observations are less than or equal to the observation.

Interpretation

A percentile rank of 74 means that 74% of the SAT math scores are less than or equal to 600 and 26% of the scores are greater than 600. So 26% of the students who took the exam scored better than Jennifer.

3.4 Objective 3 Determine and Interpret Quartiles

October 17, 2016 11:17 AM

The most common percentiles are **quartiles**, which divide data sets into fourths, or four equal parts.



- The first quartile, denoted Q_1 , divides the bottom 25% of the data from the top 75%.
- The second quartile, Q_2 , divides the bottom 50% of the data from the top 50%.
- The third quartile, Q_3 , divides the bottom 75% of the data from the top 25%.

Figure 9 illustrates the concept of quartiles.

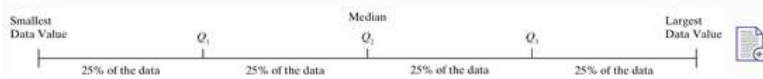


Figure 9

Finding Quartiles

Step 1 Arrange the data in ascending order.

Step 2 Determine the median, M , or second quartile, Q_2 .

Step 3 Divide the data set into two halves: the observations less than M and the observations greater than M . The first quartile, Q_1 , is the median of the bottom half, and the third quartile, Q_3 , is the median of the top half. Do not include M in these halves.

EXAMPLE 3 Finding and Interpreting Quartiles

Problem

The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. Table 12 represents a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute for 2007 models. Find and interpret the first, second, and third quartiles for collision coverage claims.

TABLE 12

\$6751	\$9908	\$3461
\$2336	\$21,147	\$2332
\$189	\$1185	\$370
\$1414	\$4668	\$1953
\$10,034	\$735	\$802
\$618	\$180	\$1657

Video Solution



Technology Step-By-Step



Approach

Follow the [steps](#) for finding quartiles.

Solution

Step 1 Write the data in ascending order:

\$180	\$189	\$370	\$618	\$735	\$802	\$1185	\$1414	\$1657
\$1953	\$2332	\$2336	\$3461	\$4668	\$6751	\$9908	\$10,034	\$21,147

Step 2 There are $n = 18$ observations, so the median, or second quartile, Q_2 , is the mean of the 9th and 10th observations. Therefore, $M = Q_2 = \frac{\$1657 + \$1953}{2} = \$1805$.

Step 3 The median of the bottom half of the data is the first quartile, Q_1 , which is the 5th observation, so $Q_1 = \$735$.

\$180 \$189 \$370 \$618 \$735 \$802 \$1185 \$1414 \$1657
 ↑
 Q_1

The median of the top half of the data is the third quartile, Q_3 , which is the 5th observation, so $Q_3 = \$4668$.

\$1953 \$2332 \$2336 \$3461 \$4668 \$6751 \$9908 \$10,034 \$21,147
 ↑
 Q_3

Interpretation

- For the first quartile, 25% of the collision claims are less than or equal to \$735, and 75% of the collision claims are greater than \$735.
- For the second quartile, 50% of the collision claims are less than or equal to \$1805, and 50% of the collision claims are greater than \$1805.
- For the third quartile, 75% of the collision claims are less than or equal to \$4668, and 25% of the collision claims are greater than \$4668.

3.4 Objective 4 Determine and Interpret the Interquartile Range

October 17, 2016 11:39 AM

So far we have discussed three measures of dispersion: [range](#), [standard deviation](#), and [variance](#), all of which are not [resistant](#). Quartiles, however, are resistant. For this reason, quartiles are used to define a resistant measure of dispersion.

DEFINITION

The **interquartile range**, IQR, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using this formula $IQR = Q_3 - Q_1$.

The interpretation of the interquartile range is the range of the middle 50% of the data. The more spread a set of data has, the higher the interquartile range will be. The interquartile range, IQR, is a resistant measure of dispersion.

EXAMPLE 4 Finding and Interpreting the Interquartile Range

Problem

Determine and interpret the interquartile range of the collision claim data from [Table 12](#) in [Example 3](#).

Video Solution



Approach

Using the quartiles found in [Example 3](#), find the interquartile range, IQR, by computing the difference between the first and third quartiles. It represents the range of the middle 50% of the observations.

Solution

The interquartile range is

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= \$4668 - \$735 \\ &= \$3933 \end{aligned}$$

Interpretation

The IQR, or range of the middle 50% of the observations, for the collision claim data is **\$3933**. The IQR is used as the measure of dispersion for data sets that are skewed or contain extreme observations because it is a resistant measure of dispersion.

Deciding Which Measure of Central Tendency and Dispersion to Report

Let's compare the measures of central tendency and dispersion discussed for the collision claim data.

- The median of \$1805 is more representative than the mean of the "center" because the data are skewed to the right (only 5 of the 18 observations are greater than the mean, which is \$3874.4).
- The range is $\$21,147 - \$180 = \$20,967$. The standard deviation is \$5301.6 and the interquartile range is \$3933. The values of the range and standard deviation are affected by the extreme claim of \$21,147. In fact, if this claim were \$120,000 (let's say the claim was for a totaled Mercedes S-class AMG), then the range and standard deviation would increase to \$119,820 and \$27,782.5, respectively. The interquartile range would not be affected.

Therefore, when the distribution of data is highly skewed or contains extreme observations, it is best to use the median as the measure of central tendency and the interquartile range as the measure of dispersion because these measures are resistant.

SUMMARY: WHICH MEASURES TO REPORT

Shape Of Distribution	Measure Of Central Tendency	Measure Of Dispersion
Symmetric	Mean	Standard Deviation
Skewed Left Or Skewed Right	Median	Interquartile Range

For the remainder of this course, the direction **describe the distribution** will mean to describe its shape (skewed left, skewed right, or symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

3.5 Objective 5 Check a Set of Data for Outliers

October 17, 2016 11:42 AM

When analyzing data, we must check for extreme observations, called **outliers**. Outliers can occur by chance, because of errors in the measurement of a variable, during data entry, or from errors in sampling. For example, in the 2000 presidential election, a precinct in New Mexico accidentally recorded 610 absentee ballots for Al Gore (the Democratic nominee) as 110. Workers in the Gore camp discovered the data entry error through an analysis of vote totals.

Outliers aren't always due to error or chance. Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European cars. If our sample included a Ferrari F430 Spider (approximately \$175,000), it probably would be an outlier because this car costs much more than the typical European car. The value of this car would be considered *unusual* because it is not a typical value from the data set.

Checking for Outliers by Using Quartiles

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. Fences serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

TABLE 12

\$6751	\$9908	\$3461
\$2336	\$21,147	\$2332
\$189	\$1185	\$370
\$1414	\$4668	\$1953
\$10,034	\$735	\$802
\$618	\$180	\$1657

EXAMPLE 5 Checking for Outliers

Problem

Check the data in Table 12 on collision coverage claims for outliers.

Video Solution



Approach

Follow the steps for checking for outliers. Any data value that is less than the lower fence or greater than the upper fence will be considered an outlier.

Solution

Step 1 The quartiles found in Example 3 are $Q_1 = \$735$ and $Q_3 = \$4668$.

Step 2 The interquartile range, IQR, is

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= \$4668 - \$735 \\ &= \$3933 \end{aligned}$$

Step 3 The lower fence, LF, is

$$\begin{aligned} \text{LF} &= Q_1 - 1.5 (\text{IQR}) \\ &= \$735 - 1.5 (\$3933) \\ &= -\$5164.5 \end{aligned}$$

The upper fence, UF, is

$$\begin{aligned} \text{UF} &= Q_3 + 1.5 (\text{IQR}) \\ &= \$4668 + 1.5 (\$3933) \\ &= \$10,567.5 \end{aligned}$$

Step 4 There are no observations below the lower fence. However, there is an observation above the upper fence. The claim of \$21,147 is an outlier.

3.4 Interactive Assignment

October 17, 2016 11:52 AM

3.4 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell Date: 10/17/16	Instructor: Matthew Nabity Course: MTH 243: Introduction to Probability and (4)	Assignment: 3.4 Interactive Assignment
---	--	---

Violent crimes include rape, robbery, assault, and homicide. The following is a summary of the violent-crime rate (violent crimes per 100,000 population) for all states of a country in a certain year. Complete parts (a) through (d).

$$Q_1 = 271.8, Q_2 = 387.9, Q_3 = 529.7$$

(a) Provide an interpretation of these results.

The k th percentile of a set of data is a value such that k percent of the observations are less than or equal to the value.

The most common percentiles are quartiles. Quartiles divide data sets into fourths, or four equal parts. The first quartile, denoted Q_1 , divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile. The second quartile divides the bottom 50% of the data from the top 50%, so the second quartile is equivalent to the 50th percentile, which is equivalent to the median. Finally, the third quartile divides the bottom 75% of the data from the top 25%, so the third quartile is equivalent to the 75th percentile.

Provide an interpretation of these results. Choose the correct answer below.

- A.** 25% of the states have a violent-crime rate that is 271.8 crimes per 100,000 population or more. 50% of the states have a violent-crime rate that is 387.9 crimes per 100,000 population or more. 75% of the states have a violent-crime rate that is 529.7 crimes per 100,000 population or more.
- B.** 25% of the states have a violent-crime rate that is 271.8 crimes per 100,000 population or less. 50% of the states have a violent-crime rate that is 387.9 crimes per 100,000 population or less. 75% of the states have a violent-crime rate that is 529.7 crimes per 100,000 population or less.
- C.** 75% of the states have a violent-crime rate that is 271.8 crimes per 100,000 population or less. 50% of the states have a violent-crime rate that is 387.9 crimes per 100,000 population or less. 25% of the states have a violent-crime rate that is 529.7 crimes per 100,000 population or less.

(b) Determine and interpret the interquartile range.

The interquartile range, denoted IQR, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using the formula below.

$$IQR = Q_3 - Q_1$$

The interquartile range is $529.7 - 271.8 =$ crimes per 100,000 population.

(Type an integer or a decimal.)

The interpretation of the interquartile range is similar to that of the range and standard deviation. That is, the more spread a set of data has, the higher the interquartile range will be.

Interpret the interquartile range. Choose the correct answer below.

- A.** All observations have a range of 257.9 crimes per 100,000 population.
- B.** The middle 50% of all observations have a range of 116.1 crimes per 100,000 population.
- C.** The middle 25% of all observations have a range of 257.9 crimes per 100,000 population.
- D.** The middle 50% of all observations have a range of 257.9 crimes per 100,000 population.

(c) The violent-crime rate in a certain state of the country in that year was 1,467. Would this be an outlier?

To check for outliers by using quartiles, first determine the first and third quartiles of the data. Then compute the interquartile range and determine the fences, using the formulas below.

$$\text{Lower fence} = Q_1 - 1.5(IQR)$$

$$\text{Upper fence} = Q_3 + 1.5(IQR)$$

First find the upper fence. Recall from part (b) that the interquartile range is 257.9.

$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

$$\text{Upper fence} = 529.7 + 1.5(257.9)$$

$$\text{Upper fence} = 916.55$$

Upper fence = (Type an integer or a decimal.)

Next find the lower fence.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Lower fence} = 271.8 - 1.5(257.9)$$

$$\text{Lower fence} = -115.05$$

Upper fence = (Type an integer or a decimal.)

Fences serve as cutoff points for determining outliers. If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

Would the violent-crime rate of 1,467 in a certain state of the country in that year be an outlier?

- Yes
 No

(d) Do you believe that the distribution of violent-crime rates is skewed or symmetric?

The shape of a distribution can be described as symmetric (in particular, bell shaped or uniform) or skewed (skewed right or skewed left).

When data are either skewed left or skewed right, there are extreme values in the tail, which tend to pull the mean in the direction of the tail. For example, in skewed-right distributions, there are large observations in the right tail.

Another way to determine whether the distribution is skewed or symmetric is to compare the difference $Q_2 - Q_1$ to the difference $Q_3 - Q_2$. If the differences are about equal, the distribution is symmetric. If the differences are not approximately equal, the distribution is skewed.

Find the difference $Q_2 - Q_1$.

$$Q_2 - Q_1 = 387.9 - 271.8$$

$$Q_2 - Q_1 = 116.1$$

$Q_2 - Q_1$ = (Type an integer or a decimal.)

Now find the difference $Q_3 - Q_2$.

$$Q_3 - Q_2 = 529.7 - 387.9$$

$$Q_3 - Q_2 = 141.8$$

$Q_3 - Q_2$ = (Type an integer or a decimal.)

Do you believe that the distribution of violent-crime rates is skewed or symmetric?

- A. The distribution of violent-crime rates is skewed left.
 B. The distribution of violent-crime rates is symmetric.
 C. The distribution of violent-crime rates is skewed right.

YOU ANSWERED: A.

3.5 The Five-Number Summary and Boxplots

October 17, 2016 11:59 AM

3.5 Objective 1 Determine the five-number summary

October 18, 2016 08:43 AM

Remember that the **median** is **resistant** to extreme values, so it is the preferred measure of central tendency when data are **skewed right** or **skewed left**.

The three measures of dispersion that are not resistant are the **range**, **standard deviation**, and **variance**. The **interquartile range** is resistant. However, the median, Q_1 , and Q_3 do not provide information about the extremes of the data. For this, we need the smallest and largest values in the data set.

The **five-number summary** of a set of data consists of the smallest data value, Q_1 , the median, Q_3 , and the largest data value. We use the five-number summary to learn information about the extremes of the data set. The summary is organized as follows:

Five-Number Summary

Minimum	Q_1	M	Q_3	
Maximum				

Five-Number summary:

1. Minimum
2. Maximum
3. Q1
4. Q2 (Median)
5. Q3

EXAMPLE 1 Obtaining the Five-Number Summary

Problem

Table 13 shows the finishing times (in minutes) of the men in the 60- to 64-year-old age group in a 5-kilometer race. Determine the five-number summary of the data.

Video Solution

Technology Step-By-Step

19.95	23.25	23.32	25.55	25.83	26.28	42.47
28.58	28.72	30.18	30.35	30.95	32.13	49.17
33.23	33.53	36.68	37.05	37.43	41.42	54.63

Data from Laura Gillogly, student at Joliet Junior College

Approach

The five-number summary includes the minimum data value, Q_1 , M (the median), Q_3 , and the maximum data value. We use the **procedures** for finding quartiles.

Solution

The data in ascending order are as follows:

19.95, 23.25, 23.32, 25.55, 25.83, 26.28, 28.58, 28.72, 30.18, 30.35, 30.95,
32.13, 33.23, 33.53, 36.68, 37.05, 37.43, 41.42, 42.47, 49.17, 54.63

The smallest number (the fastest time) in the data set is 19.95. The largest number in the data set is 54.63. The first quartile, Q_1 , is 26.06. The median, M , is 30.95. The third quartile, Q_3 , is 37.24. The five-number summary is:

19.95 26.06 30.95 37.24 54.63

Screen clipping taken: 18-Oct-16 08:48 AM

3.5 Objective 2 Draw and Interpret Boxplots

October 18, 2016 08:49 AM

Drawing a Boxplot

Step 1 Determine the lower and upper fences:

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Upper Fence} = Q_3 + 1.5(\text{IQR}) \quad \text{where } \text{IQR} = Q_3 - Q_1$$

Step 2 Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q_1 , M , and Q_3 . Enclose these vertical lines in a box.

Step 3 Label the lower and upper fences with a temporary mark.

Step 4 Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called **whiskers**.

Step 5 Plot any data values less than the lower fence or greater than the upper fence as outliers. Outliers are marked with an asterisk (*). Remove the temporary marks labeling the fences.

EXAMPLE 2 Constructing a Boxplot

Problem

Use the [results](#) of Example 1 to construct a boxplot of the finishing times of the men in the 60- to 64-year-old age group. The [data](#) is provided for convenience.

Approach

Follow the [steps](#) for constructing a boxplot.

Video Solution



Technology Step-By-Step



Solution

Step 1 In Example 1, we found that $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Therefore, the interquartile range = $IQR = Q_3 - Q_1 = 37.24 - 26.06 = 11.18$ minutes. From this, we find the lower fence and upper fence.

$$\text{Lower fence} = Q_1 - 1.5(IQR) = 26.06 - 1.5(11.18) = 9.29$$

$$\text{Upper fence} = Q_3 + 1.5(IQR) = 37.24 + 1.5(11.18) = 54.01$$

Step 2 Draw a horizontal number line with a scale that will accommodate our graph. Draw vertical lines at $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Enclose these lines in a box. See Figure 10(a).

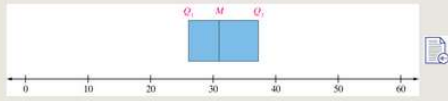


Figure 10(a)

Step 3 Temporarily mark the location of the lower and upper fences with brackets ([and]). See Figure 10(b).

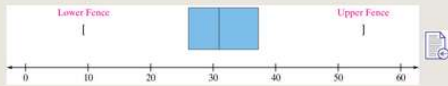


Figure 10(b)

Step 4 Draw a horizontal line from Q_1 to 19.95, the smallest data value that is larger than 9.29 (the lower fence). Draw a horizontal line from Q_3 to 49.17, the largest data value that is smaller than 54.01 (the upper fence). See Figure 10(c).

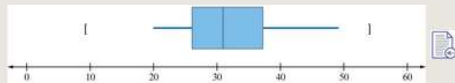


Figure 10(c)

Step 5 Plot any data values less than the lower fence or greater than the upper fence as outliers. Outliers are marked with an asterisk (*). Remove the temporary marks labeling the fences. See Figure 10(d).

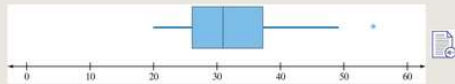


Figure 10(d)

Figure 11 suggests guidelines for identifying the distribution shape from a boxplot. Judging the shape of a distribution is a subjective practice.

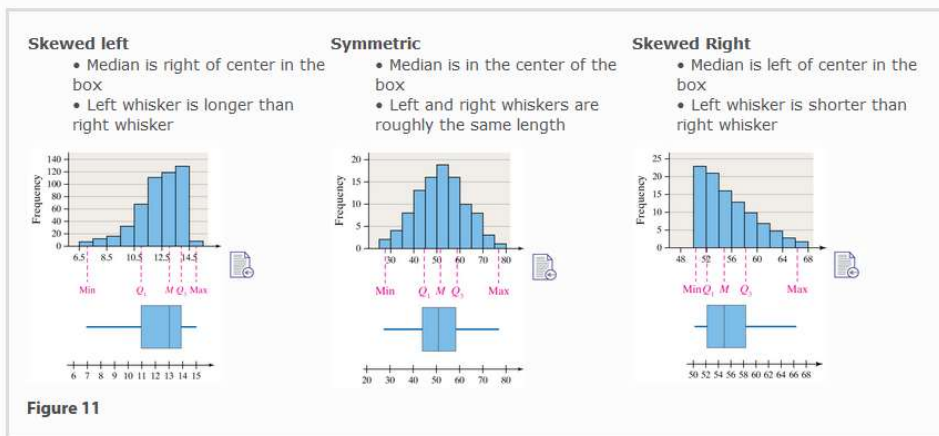


Figure 11

Using a Boxplot to Determine Skewness

The boxplot in Figure 10(d) suggests that the distribution is skewed right because the right whisker is longer than the left whisker and the median is left of center in the box. We can also assess the shape of a distribution using quartiles. The distance from \bar{M} to Q_1 is 4.89 ($= 30.95 - 26.06$), whereas the distance from \bar{M} to Q_3 is 6.29 ($= 37.24 - 30.95$). Also, the distance from \bar{M} to the minimum value is 11 ($= 30.95 - 19.95$), whereas the distance from \bar{M} to the maximum value is 23.68 ($= 54.63 - 30.95$).

When describing the shape of a distribution from a boxplot, be sure to justify your conclusion as we did above.

EXAMPLE 3 Comparing Two Distributions Using Boxplots

Problem

Table 14 shows the red blood cell mass (in millimeters) for 14 rats sent into space (flight group) and for 14 rats that were not sent into space (control group). Construct side-by-side boxplots for red blood cell mass for the flight group and control group. Does it appear that space flight affects the rats' red blood cell mass?

TABLE 14

Flight				Control			
7.43	7.21	8.59	8.64	8.65	6.99	8.40	9.66
9.79	6.85	6.87	7.89	7.62	7.44	8.55	8.70
9.30	8.03	7.00	8.80	7.33	8.58	9.88	9.94
6.39	7.54			7.14	9.14		

Data from NASA Life Sciences Data Archive

Solution

[Video Solution](#) [Technology Step-By-Step](#)



HIDE SOLUTION

Example – Comparing Two Distributions by Using Boxplots

In the Spacelab Life Sciences 2, led by Paul X. Callahan, 14 male rats were sent to space. The red blood cell mass (in millimeters) of the rats was determined when they returned.

A control group of 14 male rats was held under the same conditions (except for space flight) as the space rats, and their red blood cell mass was also measured when the space rats returned.

Flight						
7.43	7.21	8.59	8.64	9.79	6.85	6.87
7.89	9.30	8.03	7.00	8.80	6.39	7.54

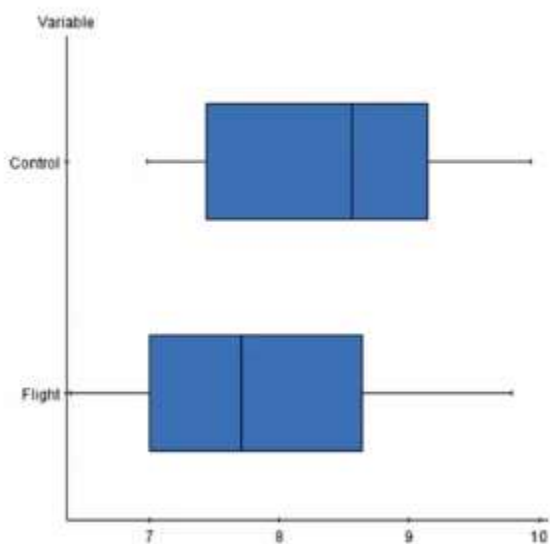
Control						
8.65	6.99	8.40	9.66	7.62	7.44	8.55
8.70	7.33	8.58	9.88	9.94	7.14	9.14

Example – Comparing Two Distributions by Using Boxplots

Construct side-by-side boxplots for red blood cell mass for the flight group and the control group.

Does it appear that spaceflight affects the rats' red blood cell mass?

Solution



Solution

It appears that spaceflight has reduced the red blood cell mass of the rats.

5.1 Probability Rules

October 20, 2016 08:56 AM

The video you just watched illustrates a basic premise of probability. **Probability is the measure of the likelihood of a random phenomenon or chance behavior occurring.** It deals with experiments that yield random short-term results or outcomes yet reveal long-term predictability. The long-term proportion in which a certain outcome is observed is the probability of that outcome. So we say that the probability of observing a head is $\frac{1}{2}$ or 50% or 0.5 because as we flip the coin more times, the proportion of heads tends toward $\frac{1}{2}$. This phenomenon is referred to as the *Law of Large Numbers*.

The Law of Large Numbers

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.

Jakob Bernoulli (a major contributor to the field of probability) believed that the Law of Large Numbers was common sense. This is evident in the following quote from his text *Ars Conjectandi*: "For even the most stupid of men, by some instinct of nature, by himself and without any instruction, is convinced that the more observations have been made, the less danger there is of wandering from one's goal."

In probability, an **experiment is any process with uncertain results that can be repeated.** The result of any single *trial* of the experiment is not known ahead of time. However, the results of the experiment over many trials produce regular patterns that allow accurate predictions. For example, an insurance company cannot know whether a particular 16-year-old driver will have an accident over the course of a year. However, based on historical records, the company can be fairly certain that about three out of every ten 16-year-old male drivers will have a traffic accident during the course of a year. Therefore, of 825,000 male 16-year-old drivers (825,000 repetitions of the experiment), the insurance company is fairly confident that about 30%, or 247,500, will have an accident. This prediction helps to establish insurance rates for any particular 16-year-old male driver.

We now introduce some terminology that will help in our study of probability.

DEFINITIONS

The **sample space, S** , of a probability experiment is the collection of all possible outcomes for that experiment.

An **event** is any collection of outcomes from a probability experiment. An event consists of one or more outcomes. We denote events with one outcome, sometimes called *simple events*, as e_i . In general, events are denoted using capital letters such as E .

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 1 Identifying Events and the Sample Space of a Probability Experiment

Problem

A probability experiment consists of rolling a single six-sided *fair* die. A fair die is one in which each possible outcome is equally likely. For example, rolling a two is just as likely as rolling a five.

Video Solution



Part (A)

Part (B)

Part (C)

Identify the outcomes of the probability experiment.

Approach

The outcomes are the possible results of the experiment.

Solution

The outcomes from rolling a single fair die are $e_1 = \text{"rolling a one"} = \{1\}$,
 $e_2 = \text{"rolling a two"} = \{2\}$, $e_3 = \text{"rolling a three"} = \{3\}$,
 $e_4 = \text{"rolling a four"} = \{4\}$, $e_5 = \text{"rolling a five"} = \{5\}$, and
 $e_6 = \text{"rolling a six"} = \{6\}$.

Part (A)

Part (B)

Part (C)

Determine the sample space.

Approach

The sample space is a list of all possible outcomes.

Solution

The set of all possible outcomes forms the sample space, $S = \{1, 2, 3, 4, 5, 6\}$. There are six outcomes in the sample space.

Part (A)

Part (B)

Part (C)

Define the event $E = \text{"roll an even number"}$.

Solution

The event $E = \text{"roll an even number"} = \{2, 4, 6\}$.

HIDE SOLUTION

5.1 Objective 1 Apply the Rules of Probability

October 20, 2016 09:19 AM

In the following probability rules, the notation $P(E)$ means "the probability that event E occurs."

Rules of Probabilities

1. The probability of any event E , $P(E)$, must be greater than or equal to 0 and less than or equal to 1. That is, $0 \leq P(E) \leq 1$.

2. The sum of the probabilities of all outcomes must equal 1. That is, if the sample space $S = \{e_1, e_2, \dots, e_n\}$, then

$$P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

IN
OTHER
WORDS



A **probability model** lists the possible outcomes of a probability experiment and each outcome's probability. A probability model must satisfy Rules 1 and 2 of the rules of probabilities.

EXAMPLE A Probability Model

In a bag of peanut M&M milk chocolate candies, the colors of the candies can be brown, yellow, red, blue, orange, or green. Suppose that a candy is randomly selected from a bag. The table shows each color and the probability of drawing that color. Verify this is a probability model.

Color	Probability
Brown	0.12
Yellow	0.15
Red	0.12
Blue	0.23
Orange	0.23
Green	0.15

EXAMPLE A Probability Model

In a bag of peanut M&M milk chocolate candies, the colors of the candies can be brown, yellow, red, blue, orange, or green. Suppose that a candy is randomly selected from a bag. The table shows each color and the probability of drawing that color. Verify this is a probability model.

Color	Probability
Brown	0.12
Yellow	0.15
Red	0.12
Blue	0.23
Orange	0.23
Green	0.15

Rule is satisfied because all probabilities are between 0 and 1, inclusive.

That is, $0 \leq P(e) \leq 1$ for all probabilities.

In a bag of peanut M&M milk chocolate candies, the colors of the candies can be brown, yellow, red, blue, orange, or green. Suppose that a candy is randomly selected from a bag. The table shows each color and the probability of drawing that color. Verify this is a probability model.

Color	Probability
Brown	0.12
Yellow	0.15
Red	0.12
Blue	0.23
Orange	0.23
Green	0.15

- Rule 1 is satisfied because all probabilities are between 0 and 1, inclusive.
- Rule 2 is satisfied because $0.12 + 0.15 + \dots + 0.15 = 1$.

Key Concepts Regarding Probabilities

- If an event is **impossible**, the probability of the event is 0.
- If an event is a **certainty**, the probability of the event is 1.
- The closer a probability is to 1, the more likely the event will occur.
- The closer a probability is to 0, the less likely the event will occur.
- For example, an event with probability 0.8 is more likely to occur than an event with probability 0.75.
- An event with probability 0.8 will occur about 80 times out of 100 repetitions of the experiment, whereas an event with probability 0.75 will occur about 75 times out of 100.



One goal of this course is to learn how probabilities can be used to identify *unusual events*.

DEFINITION

An **unusual event** is an event that has a low probability of occurring.

Typically, an event with a probability less than 0.05 (or 5%) is considered unusual, but this cutoff point is not set in stone. The researcher and the context of the problem determine the probability that separates unusual events from not so *unusual events*.

The point of the Caution video is this: Selecting a probability that separates unusual events from not so unusual events is subjective and depends on the situation. Statisticians typically use cutoff points of 0.01, 0.05, and 0.10.

Next, we introduce three methods for determining the probability of an event:

- the Empirical Method
- the Classical Method
- the Subjective Method

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

5.1 Objective 2 Compute and Interpret Probabilities Using the Empirical Method

October 20, 2016 09:27 AM

Probabilities deal with the likelihood that a particular outcome will be observed. For this reason, we begin our discussion of determining probabilities using the idea of relative frequency. Probabilities computed in this manner rely on empirical evidence, that is, evidence based on the outcomes of a probability experiment.

Approximating Probabilities Using the Empirical Approach

The probability of an event E occurring is approximately the number of times event E is observed divided by the number of repetitions (or trials) of the experiment.

$$P(E) \approx \text{relative frequency of } E = \frac{\text{frequency of } E}{\text{number of trials of experiment}}$$

When we find probabilities using the empirical approach, the result is approximate because different trials of the experiment lead to different outcomes and, therefore, different estimates of $P(E)$.

EXAMPLE 3 Using Relative Frequencies to Approximate Probabilities

An insurance agent currently insures 182 teenage drivers (ages 16 to 19). Last year, 24 of the teenagers had to file a claim on their auto policy. Based on these results, the probability that a teenager will file a claim on his or her auto policy in a given year is

$$\frac{24}{182} \approx 0.132$$

So, for every 100 insured teenage drivers, we expect about 13 to have a claim on their auto policy.

EXAMPLE Building a Probability Model

Pass the Pigs™ is a Milton-Bradley game in which pigs are used as dice. Points are earned based on the way the pig lands. There are six possible outcomes when one pig is tossed. A class of 52 students rolled pigs 3,939 times. The number of times each outcome occurred is recorded in the table at right.

(Source: <http://www.members.tpsod.com/~passpigs/prob.html>)

Outcome	Frequency
Side with no dot	1344
Side with dot	1294
Razorback	767
Trotter	365
Snouter	137
Leaning Jowler	32

- Use the results of the experiment to build a probability model for the way the pig lands.
- Estimate the probability that a thrown pig lands on the "side with dot".
- Would it be unusual to throw a "Leaning Jowler"?

EXAMPLE Building a Probability Model

Pass the Pigs™ is a Milton-Bradley game in which pigs are used as dice. Points are earned based on the way the pig lands. There are six possible outcomes when one pig is tossed. A class of 52 students rolled pigs 3,939 times. The number of times each outcome occurred is recorded in the table at right.

(Source: <http://www.members.tripod.com/~passpigs/prob.html>)

Outcome	Frequency
Side with no dot	1344
Side with dot	1294
Razorback	767
Trotter	365
Snouter	137
Leaning Jowler	32

- (a) Use the results of the experiment to build a probability model for the way the pig lands.
 (b) Estimate the probability that a thrown pig lands on the "side with dot".
 (c) Would it be unusual to throw a "Leaning Jowler"?

$$\begin{aligned} \text{Relative Freq of "side with no dot"} &\approx P(\text{side with no dot}) \\ &= \frac{1344}{3939} \\ &= 0.341 \end{aligned}$$

(a)

Outcome	Probability
Side with no dot	$\frac{1344}{3939} \approx 0.341$
Side with dot	0.329
Razorback	0.195
Trotter	0.093
Snouter	0.035
Leaning Jowler	0.008

(b) The probability a throw results in a "side with dot" is 0.329. In 1000 throws of the pig, we would expect about 329 to land on a "side with dot".

If we rolled a pig 1000 times, we would expect a "Leaning Jowler" in about 8 rolls.

Surveys are probability experiments. Why? Each time a survey is conducted, a different random sample of individuals is selected. Therefore, the results of a survey are likely to be different each time the survey is conducted because different people are included.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

5.1 Objective 3 Compute and Interpret Probabilities using the Classical Method

October 20, 2016 09:40 AM

The [empirical method](#) gives an approximate probability of an event by conducting a probability experiment. The classical method of computing probabilities does not require that a probability experiment actually be performed. Rather, it relies on counting techniques.

The classical method of computing probabilities requires *equally likely outcomes*. An experiment has **equally likely outcomes** when each outcome has the same probability of occurring. For example, when a fair die is thrown once, each of the six outcomes in the sample space, $\{1, 2, 3, 4, 5, 6\}$, has an equal chance of occurring. Contrast this situation with a loaded die in which a five or six is twice as likely to occur as a one, two, three, or four.

Computing Probability Using the Classical Method

If an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m , then the probability of E , $P(E)$, is

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{m}{n}$$

So, if S is the sample space of this experiment, then

$$P(E) = \frac{N(E)}{N(S)}$$

where $N(E)$ is the number of outcomes in E , and $N(S)$ is the number of outcomes in the sample space.

EXAMPLE 5 Computing Probabilities Using the Classical Approach

Problem

A pair of fair dice is rolled. Fair die are die where each outcome is equally likely. The possible outcomes of this experiment are shown in Figure 1.

[Video Solution](#)



Figure 1

Compute the probability of rolling a seven.

Approach

To compute probabilities using the classical method, count the number of outcomes in the sample space and count the number of ways the event can occur. Then, divide the number of outcomes by the number of ways the event can occur.

Solution

There are 36 equally likely outcomes in the sample space, as shown in Figure 1. So $N(S) = 36$. The event $E = \text{"roll a seven"} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ has six outcomes, so $N(E) = 6$. Therefore,

$$P(E) = P(\text{roll a seven}) = \frac{N(E)}{N(S)} = \frac{6}{36} = \frac{1}{6}$$

The probability of rolling a seven is $\frac{1}{6}$.

Compute the probability of rolling "snake eyes"; that is, compute the probability of rolling a two.

Approach

To compute probabilities using the classical method, count the number of outcomes in the sample space and count the number of ways the event can occur.

Solution

The event $F = \text{"roll a two"} = \{(1, 1)\}$ has one outcome, so $N(F) = 1$.

$$P(F) = P(\text{roll a two}) = \frac{N(F)}{N(S)} = \frac{1}{36}$$

The probability of rolling a two is $\frac{1}{36}$.

Comment on the likelihood of rolling a seven versus rolling a two.

Approach

Compute the ratio of the probability of rolling a seven to the probability of rolling a two.

Solution

Because $P(\text{roll a seven}) = \frac{6}{36}$ and $P(\text{roll a two}) = \frac{1}{36}$, rolling a seven is six times as likely as rolling a two. In other words, in **36** rolls of the dice, we expect to observe about **6** sevens and only **1** two.

Comparing Empirical Probabilities and Classical Probabilities

We just saw that the classical probability of rolling a seven is $\frac{1}{8} \approx 0.167$. Suppose a pit boss at a casino rolls a pair of dice 100 times and obtains 15 sevens. From this empirical evidence, we would assign the probability of rolling a seven as $\frac{15}{100} = 0.15$. If the dice are fair, we would expect the relative frequency of sevens to get closer to 0.167 as the number of rolls of the dice increases. In other words, the empirical probability will get closer to the classical probability as the number of trials of the experiment increases. If the two probabilities do not get closer, we may suspect that the dice are not fair.

In **simple random sampling**, each individual has the same chance of being selected. Therefore, we can use the classical method to compute the probability of obtaining a specific sample.

EXAMPLE 6 Computing Probabilities Using Equally Likely Outcomes

Problem

Sophia has three tickets to a concert, but Yolanda, Michael, Kevin, and Marissa all want to go to the concert with her. To be fair, Sophia wants to randomly select the two people who will go with her.

Video Solution



Determine the sample space of the experiment. In other words, list all possible simple random samples of size $n = 2$.

Solution

The sample space is listed in Table 1.

TABLE 1

Yolanda, Michael	Yolanda, Kevin	Yolanda, Marissa
Michael, Kevin	Michael, Marissa	Kevin, Marissa

HIDE SOLUTION

Compute the probability of the event "Michael and Kevin attend the concert."

Approach

Because each outcome is equally likely, the probability of an event is the number of outcomes in the event divided by the number of outcomes in the sample space.

Solution

We have $N(S) = 6$, and there is one way the event "Michael and Kevin attend the concert" can occur. Therefore, the probability that Michael and Kevin attend the concert is $\frac{1}{6}$.

Compute and interpret the probability of the event "Marissa attends the concert."

Approach

Because each outcome is equally likely, the probability of an event is the number of outcomes in the event divided by the number of outcomes in the sample space. We interpret probabilities using a relative frequency approach.

Solution

We have $N(S) = 6$, and there are three ways the event "Marissa attends the concert" can occur. The probability that Marissa will attend is

$$\frac{3}{6} = \frac{1}{2} = 0.5 = 50\%.$$

If we conducted this experiment 100 times, about 50 of the experiments would result in Marissa attending the concert.

EXAMPLE 7 Comparing the Classical Method and Empirical Method

Problem

Suppose that a survey asked 500 families with three children to disclose the gender of their children and found that 180 of the families had two boys and one girl.

Video Solution



Part A

Part B

Estimate the probability of having two boys and one girl in a three-child family, using the empirical method.

Approach

Determine the relative frequency of the event "two boys and one girl."

Solution

The empirical probability of the event $E =$ "two boys and one girl" is

$$P(E) \approx \text{relative frequency of } E = \frac{180}{500} = 0.36$$

There is about a 0.36 probability that a family with three children will have two boys and one girl.

Compute and interpret the probability of having two boys and one girl in a three-child family, using the classical method and assuming boys and girls are equally likely.

Approach

Count the number of ways the event "two boys and one girl" can occur and divide this by the number of possible outcomes for this experiment.

Solution

To determine the sample space, we construct a **tree diagram** to list the equally likely outcomes of the experiment. To construct a tree diagram for this situation, draw two branches corresponding to the two possible outcomes (boy or girl) for the first trial (the first child). Then, for the second child, draw four branches, and so on. See [Figure 2](#), where B stands for boy and G stands for girl.

The sample space S of this experiment is found by following each branch to identify all the possible outcomes of the experiment:

$$S = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$$

So $N(S) = 8$.

For the event $E = \text{"two boys and a girl"} = \{BBG, BGB, GBB\}$ we have $N(E) = 3$. Because the outcomes are equally likely (for example, BBG is just as likely as BGB), we have

$$P(E) = \frac{N(E)}{N(S)} = \frac{3}{8} = 0.375$$

There is a **0.375** probability that a family of three children will have two boys and one girl. If we repeat this experiment **1000** times and the outcomes are equally likely (having a girl is just as likely as having a boy), we would expect about **375** of the trials to result in two boys and one girl.

LAW OF LARGE NUMBERS:

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

Comparing Empirical Probabilities and Classical Probabilities

In comparing the results of Examples 7(a) and 7(b), we notice that the two probabilities are slightly different. We know that empirical probabilities and classical probabilities often differ in value, but as the number of repetitions of a probability experiment increases, the empirical probability should get closer to the classical probability according to the [Law of Large Numbers](#).

However, it is possible that the two probabilities differ because having a boy and having a girl are not equally likely events. (Maybe the probability of having a boy is 0.505 and the probability of having a girl is 0.495.) If this is the case, then the empirical probability will not get closer to the classical probability because the events "boy" and "girl" are not equally likely.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

5.1 Objective 4 Use Simulation to Obtain Data based on Probabilities

October 20, 2016 10:04 AM

Example – Simulating Probabilities

- (a)** Simulate the experiment of sampling 100 three-child families to estimate the probability that a three-child family has two boys.
- (b)** Simulate the experiment of sampling 1000 three-child families to estimate the probability that a three-child family has two boys.

StatCrunch Steps

- Data > Simulate data > Discrete Uniform
 - Set Rows to be 100 or 1000 (# of families) and Columns to be 3 (# of children).
Set Minimum to be 0 and Maximum to be 1, where 0 represents a girl and 1 represents a boy.
 - Click Compute!
 - Rename columns as Child 1, Child 2, and Child 3.
-
- Use Data > Compute expression to find the sum of each row.
 - To find the count for each possible number of boys:
Stat > Summary Stats > Columns
Select column containing sums, and Group by the same column.
Click on n .
Click Compute!

5.1 Objective 5 Recognize and Interpret Subjective Probabilities

October 20, 2016 10:11 AM

If a sports reporter is asked what he thinks the chances are for the Boston Red Sox to play in this season's World Series, the reporter would likely process information about the Red Sox (pitching staff, leadoff hitter, and so on) and then make an educated guess of the likelihood. The reporter may respond that there is a 20% chance the Red Sox will play in the World Series. This forecast is a probability, although it is not based on relative frequencies. We cannot, after all, repeat the experiment of playing a season under the same circumstances (same players, schedule, and so on) over and over. Nonetheless, the forecast of $20\% = 0.20$ does satisfy the criterion that a probability be between 0 and 1, inclusive. This forecast is known as a *subjective probability*.

DEFINITION

A **subjective probability** is a probability that is determined based on personal judgment.

Subjective probabilities are legitimate and are often the only method of assigning likelihood to an outcome. For instance, a financial reporter may ask an economist about the likelihood of the economy falling into recession next year. Again, we cannot conduct an experiment n times to find a relative frequency. The economist must use knowledge of the current conditions of the economy and make an educated guess about the likelihood of recession.

5.1.29 Question Help

October 20, 2016 09:47 AM

A bag of 100 tulip bulbs purchased from a nursery contains 45 red tulip bulbs, 20 yellow tulip bulbs, and 35 purple tulip bulbs. What is the probability that a randomly selected tulip bulb is purple?

If an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m , then the probability of E , $P(E)$ is calculated using the equation below.

$$P(E) = \frac{\text{Number of ways that } E \text{ can occur}}{\text{Number of possible outcomes}} = \frac{m}{n}$$

Let event E = "a randomly selected tulip bulb is purple." To find $P(E)$, first find m , the number of ways that event E can occur.

$$m = 35$$

Now find n , the number of possible outcomes.

$$n = 100$$

Finally, compute $P(E)$.

$$P(E) = \frac{m}{n} = .35 \text{ (Type an integer or a simplified fraction.)}$$

Therefore, the probability that a randomly selected tulip bulb is purple is $\frac{7}{20}$.

5.1.43 Question Help

October 20, 2016 09:58 AM

5.1 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 10/20/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 5.1 Interactive Assignment

Clarice, John, and Marco work for a publishing company. The company wants to send two employees to a statistics conference. To be fair, the company decides that the two individuals who get to attend will have their names drawn from a hat. This is like obtaining a simple random sample of size 2. (a) Determine the sample space of the experiment. That is, list all possible simple random samples of size $n = 2$. (b) What is the probability that Clarice and John attend the conference? (c) What is the probability that John stays home?

(a) To find the sample space of the experiment, write out all the possible pairs of employees in a table. The first two pairs are shown below.

Sample space	
CJ (Clarice, John)	CM (Clarice, Marco)

Continue the process above and find the sample space of the experiment. Choose the correct answer below. (Note that each person is represented by the first letter in their name.)

- A. CJ, CM, JM, CC, JJ, MM
- B. CJ, CM
- C. CJ, CM, JM, JC, MC, MJ
- D. CJ, CM, JM

Therefore, sample space $S = \{CJ, CM, JM\}$.

(b) If an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m , then the probability of E , $P(E)$ is calculated using the equation below.

$$P(E) = \frac{\text{Number of ways that } E \text{ can occur}}{\text{Number of possible outcomes}} = \frac{m}{n}$$

So, if S is the sample space of the experiment then $P(E)$ can be calculated using the equation below, where $N(E)$ is the number of outcomes in E , and $N(S)$ is the number of outcomes in the sample space.

$$P(E) = \frac{N(E)}{N(S)}$$

Let event $E = \text{"Clarice and John attend the conference."}$ Then $E = \{CJ\}$. What is $N(E)$?

$N(E) =$

If $S = \{CJ, CM, JM\}$, what is $N(S)$?

$N(S) =$

Now find $P(E)$.

$P(E) = \frac{N(E)}{N(S)} =$ (Type an integer or a simplified fraction.)

Therefore, the probability that Clarice and John attend the conference is $\frac{1}{3}$.

(c) Let event $E = \text{"John stays home."}$ To find $P(E)$, use the same equation used in part (a).

$$P(E) = \frac{N(E)}{N(S)}$$

To find $N(E)$, find E given that $S = \{CJ, CM, JM\}$. Choose the correct answer below for $E = \text{"John stays home."}$

- A. CJ, CM, JM
- B. \emptyset
- C. CM
- D. JC, JM

Therefore, if $E = \{CM\}$, what is $N(E)$?

$$N(E) = \boxed{1}$$

Since S is the same as in part (b), $N(S) = 3$. Now find $P(E)$.

$$P(E) = \frac{N(E)}{N(S)} = \boxed{\frac{1}{3}} \quad (\text{Type an integer or a simplified fraction.})$$

Therefore, the probability that John stays home is $\frac{1}{3}$.

YOU ANSWERED: 6

5.2 The Addition Rule and Complements

October 22, 2016 09:00 AM

5.2 Objective 1 Use the Addition Rule for Disjoint Events

October 22, 2016 09:02 AM

EXAMPLE 1 Benford's Law and the Addition Rule for Disjoint Events

Problem

Our number system consists of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9.

Because we do not write numbers such as 12 as 012, the first significant digit in any number must be 1, 2, 3, 4, 5, 6, 7, 8, or 9. Although we may think that each digit appears with equal frequency so that each digit has a $\frac{1}{9}$ probability of being the first significant digit, this is not true. In 1881, Simon Newcomb discovered that first-digits do not occur with equal frequency. The physicist Frank Benford discovered the same result in 1938. After studying a great deal of data, he assigned probabilities of occurrence for each of the first digits, as shown in Table 2. The [probability model](#) is now known as *Benford's Law* and plays a major role in identifying fraudulent data on tax returns and accounting books.

Video Solution



TABLE 2

Digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Data from The First Digit Phenomenon, T. P. Hill, American Scientist, July–August, 1998

Part (A)

Part (B)

Part (C)

Verify that Benford's Law is a probability model.

Approach

Verify that each probability is between 0 and 1, inclusive, and that the sum of all probabilities equals 1.

Solution

Each probability in Table 2 is between 0 and 1. In addition, the sum of all the probabilities, $0.301 + 0.176 + 0.125 + \cdots + 0.046$, is 1. Because Rules 1 and 2 are satisfied, Table 2 represents a probability model.

Use Benford's Law to determine the probability that a randomly selected first digit is 1 or 2.

Approach

Use the [Addition Rule for Disjoint Events](#).

Solution

$$\begin{aligned}P(1 \text{ or } 2) &= P(1) + P(2) \\ &= 0.301 + 0.176 \\ &= 0.477\end{aligned}$$

If we looked at 100 numbers, we would expect about 48 of them to begin with 1 or 2.

Use Benford's Law to determine the probability that a randomly selected first digit is at least 6.

Approach

The phrase *at least* means "greater than or equal to". So *at least 6* means "6, 7, 8, or 9." Use the [Addition Rule for Disjoint Events](#).

Solution

$$\begin{aligned}P(\text{at least } 6) &= P(6 \text{ or } 7 \text{ or } 8 \text{ or } 9) \\ &= P(6) + P(7) + P(8) + P(9) \\ &= 0.067 + 0.058 + 0.051 + 0.046 \\ &= 0.222\end{aligned}$$

If we looked at 100 numbers, we would expect about 22 of them to begin with 6, 7, 8, or 9.

EXAMPLE 2 A Deck of Cards and the Addition Rule for Disjoint Events

Problem

Suppose that a single card is selected from a standard 52-card deck, such as the one shown in Figure 3.

Video Solution

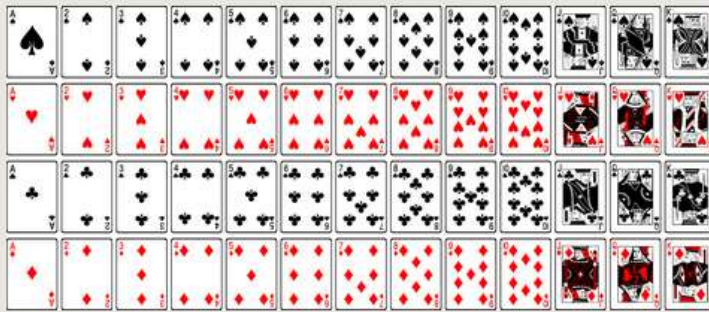


Figure 3

Part (A)

Part (B)

Compute the probability of the event $E =$ “drawing a king.”

Approach

Because the outcomes are **equally likely**, use the **classical method** for computing the probabilities.

Solution

The sample space consists of the 52 cards in the deck, so $N(S) = 52$. A standard deck of cards has four kings, so $N(E) = 4$. Therefore,

$$P(\text{king}) = P(E) = \frac{N(E)}{N(S)} = \frac{4}{52} = \frac{1}{13}$$

Compute the probability of the event

$E =$ “drawing a king” or $F =$ “drawing a queen” or $G =$ “drawing a jack”.

Approach

Notice that the events are **mutually exclusive** because you cannot simultaneously draw a king and a queen.

Therefore, use the **Addition Rule for Disjoint Events** to compute the probability.

Solution

A standard deck of cards has four kings, four queens, and four jacks. Because events E , F , and G are mutually exclusive, we use the Addition Rule for Disjoint Events extended to two or more disjoint events. So

$$\begin{aligned}P(\text{king or queen or jack}) &= P(E \text{ or } F \text{ or } G) \\&= P(E) + P(F) + P(G) \\&= \frac{4}{52} + \frac{4}{52} + \frac{4}{52} = \frac{12}{52} = \frac{3}{13}\end{aligned}$$

Classical Method

If an experiment has n equally likely outcomes and if the number of ways an event E can occur is m , then the probability of E , $P(E)$, is

$$P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{m}{n}$$

5.2 Objective 2 Use the General Addition Rule

October 22, 2016 09:10 AM

What happens when you need to compute the probability of two events that are not disjoint?

Suppose we are randomly selecting chips labeled 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9 from a bag. Let E represent the event "choose an odd number" and let F represent the event "choose a number less than or equal to 4." Because $E = \{1, 3, 5, 7, 9\}$ and $F = \{0, 1, 2, 3, 4\}$ have the outcomes 1 and 3 in common, the events are not disjoint. Figure 4 shows a Venn diagram of these events.

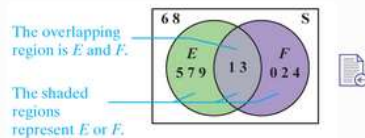


Figure 4

We can compute $P(E \text{ or } F)$ directly by counting because each outcome is equally likely. There are 8 outcomes in E or F and 10 outcomes in the sample space, so

$$P(E \text{ or } F) = \frac{N(E \text{ or } F)}{N(S)} = \frac{8}{10} = \frac{4}{5}$$

Notice that using the Addition Rule for Disjoint Events to find $P(E \text{ or } F)$ would be *incorrect*.

$$P(E \text{ or } F) \neq P(E) + P(F) = \frac{5}{10} + \frac{5}{10} = 1$$

This implies that the chips labeled 6 and 8 will never be selected, which contradicts our assumption that all the outcomes are equally likely. Our result is incorrect because we counted the outcomes 1 and 3 twice: once for event E and once for event F . To avoid this double counting, we must subtract the probability corresponding to the overlapping region, E and F . That is, we must subtract $P(E \text{ and } F) = \frac{2}{10}$ from the result and obtain:

$$\begin{aligned} P(E \text{ or } F) &= P(E) + P(F) - P(E \text{ and } F) \\ &= \frac{5}{10} + \frac{5}{10} - \frac{2}{10} \\ &= \frac{8}{10} = \frac{4}{5} \end{aligned}$$

This probability, $4/5$, agrees with the result we found by counting. The following rule generalizes these results.

The General Addition Rule

For any two events E and F ,

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

EXAMPLE 3 Computing Probabilities for Events That Are Not Disjoint

Problem

Suppose a single card is selected from a standard 52-card deck. Compute the probability of the event $E =$ “drawing a king” or $F =$ “drawing a diamond.”

Video Solution



Approach

The events are not disjoint because the outcome “king of diamonds” is in both events, so we use the [General Addition Rule](#).

Solution

$$\begin{aligned}P(E \text{ or } F) &= P(E) + P(F) - P(E \text{ and } F) \\P(\text{king or diamond}) &= P(\text{king}) + P(\text{diamond}) - P(\text{king of diamonds}) \\&= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} \\&= \frac{16}{52} \\&= \frac{4}{13}\end{aligned}$$

Contingency Tables

Consider the data shown in Table 3, which represents the marital status of males and females 15 years and older in the United States in 2013.

TABLE 3

Marital Status	Gender	
	Males (in millions)	Females (in millions)
Never married	41.6	36.9
Married	64.4	63.1
Widowed	3.1	11.2
Divorced	11.0	14.4
Separated	2.4	3.2

Data from U.S. Census Bureau, Current Population Reports

Table 3 is called a **contingency table** or **two-way table** because it relates two categories of data.

- The **row variable** is marital status because each row describes the marital status of an individual.
- The **column variable** is gender.
- Each box inside the table is called a **cell**.

For example, the cell corresponding to married individuals who are male is in the second row, first column. Each cell contains the frequency of the category: There were 64.4 million married males in the United States in 2013. Put another way, in the United States in 2013, there were 64.4 million individuals who were male *and* married.

Determine the probability that a randomly selected U.S. resident 15 years and older is male.

Approach

Add the entries in each row and column to get the total number of people in each category. Then determine the probability using the [Addition Rule for Disjoint Events](#) (because a male cannot simultaneously be married and divorced, for example).

Solution

We first add the entries in each column. For example, the “male” column shows there are $41.6 + 64.4 + 3.1 + 11.0 + 2.4 = 122.5$ million males 15 years old or older in the United States. Add the entries in each row. For example, in the “never married” row we find there are $41.6 + 36.9 = 78.5$ million U.S. residents 15 years old or older who have never married. Adding the row totals or column totals, we find there are $122.5 + 128.8 = 78.5 + 127.5 + 14.3 + 25.4 + 5.6 = 251.3$ million U.S. residents 15 years old or older.

There are 122.5 million males 15 years and older and 251.3 million U.S. residents 15 years and older. The probability that a randomly selected U.S. resident 15 years and older is male is $\frac{122.5}{251.3} = 0.487$.

Determine the probability that a randomly selected U.S. resident 15 years and older is widowed.

Approach

Determine the probability using the [Addition Rule for Disjoint Events](#) (since a widowed individual cannot be male and female simultaneously).

Solution

There are 14.3 million U.S. residents 15 years and older who are widowed, and there are 251.3 million U.S. residents 15 years and older. The probability that a randomly selected U.S. resident 15 years and older is widowed is

$$\frac{14.3}{251.3} = 0.057$$

Part (A)

Part (B)

Part (C)

Part (D)

Determine the probability that a randomly selected U.S. resident 15 years and older is widowed or divorced.

Approach

Determine the probability using the [Addition Rule for Disjoint Events](#) (since an individual cannot be widowed and divorced simultaneously).

Solution

$$\begin{aligned} P(\text{widowed or divorced}) &= P(\text{widowed}) + P(\text{divorced}) \\ &= \frac{14.3}{251.3} + \frac{25.4}{251.3} \\ &= \frac{39.7}{251.3} \\ &= 0.158 \end{aligned}$$

Part (A)

Part (B)

Part (C)

Part (D)

Determine the probability that a randomly selected U.S. resident 15 years and older is male or widowed.

Approach

Determine the probability using the [General Addition Rule](#) (because gender and marital status are not disjoint).

Solution

$$\begin{aligned} P(\text{male or widowed}) &= P(\text{male}) + P(\text{widowed}) - P(\text{male and widowed}) \\ &= \frac{122.5}{251.3} + \frac{14.3}{251.3} - \frac{3.1}{251.3} \\ &= \frac{133.7}{251.3} \\ &= 0.532 \end{aligned}$$

Addition Rule for Disjoint Events

If E and F are disjoint (or mutually exclusive) events, then

$$P(E \text{ or } F) = P(E) + P(F)$$

If E, F, G, \dots each have no outcomes in common (they are pairwise disjoint), then

$$P(E \text{ or } F \text{ or } G \text{ or } \dots) = P(E) + P(F) + P(G) + \dots$$

5.2 Objective 3 Compute the Probability of an Event Using the Complement Rule

October 22, 2016 09:40 AM

Suppose that the probability of an event E is known and we would like to determine the probability that E does not occur. This can be accomplished using the idea of *complements*.

DEFINITION

Let S denote the sample space of a probability experiment and let E denote an event. The **complement of E** , denoted E^C , is all outcomes in the sample space S that are not outcomes in the event E .

Because E and E^C are mutually exclusive,

$$P(E \text{ or } E^c) = P(E) + P(E^c) = P(S) = 1$$

Subtracting $P(E)$ from both sides, we obtain the following result.

Complement Rule

If E represents any event and E^C represents the complement of E , then

$$P(E^c) = 1 - P(E)$$

Figure 5 illustrates the Complement Rule using a Venn diagram.

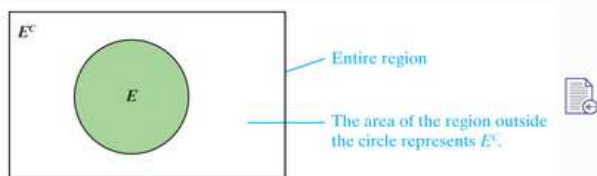


Figure 5

EXAMPLE *Illustrating the Complement Rule*

According to the American Veterinary Medical Association, 31.6% of American households own a dog. What is the probability that a randomly selected household does not own a dog?

$$\begin{aligned} P(\text{no dog}) &= 1 - P(\text{dog}) \\ &= 1 - 0.316 \\ &= 0.684 \end{aligned}$$

EXAMPLE 6 Computing Probabilities Using Complements

Problem

The data in Table 4 represent the travel time to work for residents of Hartford County, Connecticut.

Video Solution



TABLE 4

Travel Time	Frequency
Less than 5 minutes	24,358
5 to 9 minutes	39,112
10 to 14 minutes	62,124
15 to 19 minutes	72,854
20 to 24 minutes	74,386
25 to 29 minutes	30,099
30 to 34 minutes	45,043
35 to 39 minutes	11,169
40 to 44 minutes	8045
45 to 59 minutes	15,650
60 to 89 minutes	5451
90 or more minutes	4895

Data from United States Census Bureau

What is the probability that a randomly selected resident has a travel time of **90** or more minutes?

Approach

Determine the probability by finding the **relative frequency** of the event "**90** or more minutes."

Solution

There are a total of $24,358 + 39,112 + \dots + 4895 = 393,186$ residents in Hartford County, CT. Of these, **4895** had a travel time of **90** or more minutes; so

$$P(\text{90 or more minutes}) = \frac{4895}{393,186} = 0.012$$

If we randomly selected **1000** residents of Hartford County, CT, we would expect about **12** of the residents to have a travel time of **90** or more minutes. A travel time of **90** or more minutes is unusual.

What is the probability that a randomly selected resident of Hartford County, Connecticut will have a travel time less than **90** minutes?

Approach

Although this probability *could* be computed using the **Addition Rule for Disjoint Events**, it is more efficient to use the **Complement Rule** because the complement of "less than **90** minutes" is "**90** or more minutes."

Solution

$$\begin{aligned} P(\text{less than 90 minutes}) &= 1 - P(\text{90 minutes or more}) \\ &= 1 - 0.012 \\ &= 0.988 \end{aligned}$$

If we randomly selected **1000** residents of Hartford County, CT, we would expect about **988** of the residents to have a travel time less than **90** minutes.

5.3 Independence and the Multiplication Rule

October 23, 2016 10:56 AM

5.3 Objective 1 Identify Independent Events

October 23, 2016 10:59 AM

Disjoint (or Mutually Exclusive) Events versus Independent Events

Disjoint events and independent events are different concepts.

- Recall that two events are disjoint if they have no outcomes in common, that is, if knowing that one of the events occurs, we know that the other event did not occur.
- Independence means that one event occurring does not affect the probability of the other event occurring.

Therefore, knowing that two events are disjoint means that the events are not independent.

Consider the experiment of rolling a single die. Let E represent the event "roll an even number" and let F represent the event "roll an odd number." We can see that E and F are disjoint (mutually exclusive) because they have no outcomes in common. In addition, $P(E) = 1/2$ and $P(F) = 1/2$. However, if we are told that the roll of the die is going to be an even number, then what is the probability of event F ? Because the outcome will be even, the probability of event F is now 0 (and the probability of event E is now 1). So knowledge of event E changes the likelihood of observing event F .

5.3 Objective 2 Use the Multiplication Rule for Independent Events

October 23, 2016 11:00 AM

Suppose that you flip a fair coin twice. What is the probability that you obtain a head on both flips, that is, a head on the first flip *and* a head on the second flip? If \mathbf{H} represents the outcome “heads” and \mathbf{T} represents the outcome “tails,” then the sample space of this experiment is

$$S = \{\mathbf{HH}, \mathbf{HT}, \mathbf{TH}, \mathbf{TT}\}$$

There is one outcome with both heads. Because each outcome is equally likely, we have

$$\begin{aligned} P(\text{heads on Flip 1 and heads on Flip 2}) &= \frac{N(\text{heads on Flip 1 and heads on Flip 2})}{N(S)} \\ &= \frac{1}{4} \end{aligned}$$

We may have intuitively figured this out by recognizing $P(\text{head}) = \frac{1}{2}$ for each flip. So it seems reasonable that

$$\begin{aligned} P(\text{heads on Flip 1 and heads on Flip 2}) &= P(\text{heads on Flip 1}) \cdot P(\text{heads on Flip 2}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{4} \end{aligned}$$

Because both approaches result in the same answer, we conjecture that $P(E \text{ and } F) = P(E) \cdot P(F)$, which is true.

Multiplication Rule for Independent Events

If E and F are independent events, then

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

EXAMPLE 1 Computing Probabilities of Independent Events

Problem

In the game of roulette, the wheel has slots numbered 0, 00, and 1 through 36.

A metal ball rolls around a wheel until it falls into one of the numbered slots. What is the probability that the ball will land in the slot numbered 17 two times in a row?

Video Solution



Approach

Approach

The sample space of the experiment has 38 outcomes. Because the outcomes are equally likely, we use the classical method of computing probabilities. In addition, we use the Multiplication Rule for Independent Events. The events “17 on Spin 1” and “17 on Spin 2” are independent because the ball does not remember it landed on 17 on the first spin; so this cannot affect the probability of landing on 17 on the second spin.

Solution

There are 38 possible outcomes, so the probability of landing on 17 is $\frac{1}{38}$. Because the events “17 on Spin 1” and “17 on Spin 2” are independent, we have

$$\begin{aligned}P(17 \text{ on Spin 1 and } 17 \text{ on Spin 2}) &= P(17 \text{ on Spin 1}) \cdot P(17 \text{ on Spin 2}) \\ &= \frac{1}{38} \cdot \frac{1}{38} = \frac{1}{1444} \approx 0.0006925\end{aligned}$$

Round this result to 0.0007, which is read “seven ten-thousandths.” Therefore, we expect the ball to land on 17 twice in a row about 7 times in 10,000 trials.

We can extend the Multiplication Rule to three or more independent events.

Multiplication Rule for n Independent Events

If $E_1, E_2, E_3, \dots, E_n$ are independent events, then

$$P(E_1 \text{ and } E_2 \text{ and } E_3 \text{ and } \dots \text{ and } E_n) = P(E_1) \cdot P(E_2) \cdot \dots \cdot P(E_n)$$

EXAMPLE 2 Life Expectancy

Problem

The probability that a randomly selected 24-year-old male will survive the year is 0.9986 according to the *National Vital Statistics Report*, Vol. 56, No. 9.

Video Solution



Part (A)

Part (B)

What is the probability that three randomly selected 24-year-old males will survive the year?

Approach

It is safe to assume that the outcomes of the probability experiment are independent because the males were randomly selected. Therefore, there is no reason to believe that the survival of one male will affect the survival of the others. So, use the [Multiplication Rule for \$n\$ Independent Events](#).

Solution

$$\begin{aligned}P(\text{all three males survive}) &= P(\text{1st survives and 2nd survives and 3rd survives}) \\ &= P(\text{1st survives}) \cdot P(\text{2nd survives}) \cdot P(\text{3rd survives}) \\ &= (0.9986)(0.9986)(0.9986) \\ &= 0.9958\end{aligned}$$

If we randomly selected three 24-year-old males 1000 different times, we would expect all three to survive one year in 996 of the samples.

Part (A)

Part (B)

What is the probability that twenty randomly selected 24-year-old males will survive the year?

Approach

It is safe to assume that the outcomes of the probability experiment are independent because the males were randomly selected. Therefore, there is no reason to believe that the survival of one male affects the survival of the others. So, use the [Multiplication Rule for \$n\$ Independent Events](#).

Solution

$$\begin{aligned}P(\text{all 20 males survive}) &= P(\text{1st survives and 2nd survives and } \dots \text{ and 20th survives}) \\ &= P(\text{1st survives}) \cdot P(\text{2nd survives}) \cdot \dots \cdot P(\text{20th survives}) \\ &= \underbrace{(0.9986) \cdot (0.9986) \cdot \dots \cdot (0.9986)}_{\text{Multiply 0.9986 by itself 20 times}} \\ &= (0.9986)^{20} \\ &\approx 0.9724\end{aligned}$$

If we randomly selected twenty 24-year-old males 1000 different times, we would expect all twenty to survive one year in 972 of the samples.

5.3 Objective 3 Compute At-Least Probabilities

October 23, 2016 11:06 AM

Usually, when computing probabilities involving the phrase *at least*, use the [Complement Rule](#).

The phrase *at least* means “greater than or equal to.” For example, a person must be at least 17 years old to see an R-rated movie. This means that the person’s age must be greater than or equal to 17 to watch the movie.

Complement Rule:

If E represents any event and E^C represents the complement of E , then $P(E^C) = 1 - P(E)$

EXAMPLE 3 Computing At-least Probabilities

Problem

The probability that a randomly selected female aged 60 years will survive the year is 0.99186 according to the *National Vital Statistics Report*. What is the probability that at least one of 500 randomly selected 60-year-old females will die during the course of the year?

Video Solution



Approach

The phrase *at least* means “greater than or equal to,” so we want to know the probability that 1 or 2 or 3 or . . . or 500 60-year-old females will die during the year. These events are mutually exclusive, so

$$P(1 \text{ or } 2 \text{ or } 3 \text{ or } \dots \text{ or } 500 \text{ die}) = P(1 \text{ dies}) + P(2 \text{ die}) + P(3 \text{ die}) + \dots + P(500 \text{ die})$$

Computing these probabilities is very time-consuming. However, notice that the complement of “at least one dying” is “none die.” Use the [Complement Rule](#) to compute the probability.

Solution

$$\begin{aligned} P(\text{at least one dies}) &= 1 - P(\text{none die}) \\ &= 1 - P(1\text{st survives and } 2\text{nd survives and } \dots \text{ and } 500\text{th survives}) \\ &= 1 - P(1\text{st survives}) \cdot P(2\text{nd survives}) \cdot \dots \cdot P(500\text{th survives}) \\ &= 1 - (0.99186)^{500} \\ &\approx 1 - 0.0168 \\ &= 0.9832 \end{aligned}$$

If we randomly selected 500 females 60 years of age 1000 different times, we would expect at least one to die in 983 of the samples.

Summary: Rules of Probability

Rule 1 The probability of any event must be between 0 and 1, inclusive. If we let E denote any event, then $0 \leq P(E) \leq 1$.

Rule 2 The sum of the probabilities of all outcomes in the sample space must equal 1. That is, if the sample space $S = \{e_1, e_2, \dots, e_n\}$, then

$$P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

Rule 3 If E and F are disjoint events, then $P(E \text{ or } F) = P(E) + P(F)$. If E and F are not disjoint events, then $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$.

Rule 4 If E represents any event and E^c represents the complement of E , then $P(E^c) = 1 - P(E)$.

Rule 5 If E and F are independent events, then

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

Notice that *or* probabilities use the Addition Rule, whereas *and* probabilities use the Multiplication Rule. Accordingly, *or* probabilities imply addition, whereas *and* probabilities imply multiplication.

5.4 Conditional Probability and the General Multiplication Rule

October 23, 2016 11:55 AM

5.4 Objective 1 Computer Conditional Probabilities

October 23, 2016 11:55 AM

We know that when two events are independent, the occurrence of one event has no effect on the probability of the second event. However, we cannot always assume that two events will be independent. Will the probability of being in a car accident change depending on driving conditions? We would expect that the probability of an accident will be higher for driving on icy roads than for driving on dry roads.

According to data from the Centers for Disease Control and Prevention, 33.3% of adult men in the United States are obese. So the probability is 0.333 that a randomly selected U.S. adult male is obese. However, 28% of adult men aged 20 to 39 are obese compared with 40% of adult men aged 40 to 59. The probability is 0.28 that an adult male is obese, *given* that he is aged 20 to 39. The probability is 0.40 that an adult male is obese, *given* that he is aged 40 to 59. The probability that an adult male is obese changes depending on his age group. Therefore, obesity and age are not independent. This is called *conditional probability*.

DEFINITION

The notation $P(F|E)$ is read "the probability of event F given event E ." It is the **conditional probability** that event F occurs, given that event E has occurred.

For example, $P(\text{obese}|\text{20 to 39}) = 0.28$ and $P(\text{obese}|\text{40 to 59}) = 0.40$.

EXAMPLE 1 An Introduction to Conditional Probability

Problem

Suppose a single die is rolled. What is the probability that the die comes up three?
Now suppose that the die is rolled a second time, but we are told the outcome will be an odd number. What is the probability that the die comes up three?

Video Solution



Approach

Assume that the die is fair. This means that the **outcomes** are **equally likely**, so we use the **classical method** of computing probabilities.

Solution

In the first instance, there are six possibilities in the sample space, $S = \{1, 2, 3, 4, 5, 6\}$; so $P(3) = \frac{1}{6}$. In the second instance, there are three possibilities in the sample space because the only possible outcomes are odd; so $S = \{1, 3, 5\}$. This probability is expressed symbolically as $P(3|\text{roll is odd}) = \frac{1}{3}$, which is read as "the probability of rolling a three, given that the roll is odd, is one-third." Notice that the conditional probability reduces the size of the sample space under consideration (from six outcomes to three outcomes).

Classical Method:

If an experiment has n equally likely outcomes and if the number of ways an event E can occur is m , then the probability of E , $P(E)$, is

$$P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{m}{n}$$

The data in Table 5 represent the marital status of males and females 15 years and older in the United States in 2013. To find the probability that a randomly selected individual is widowed, divide the number of widowed individuals by the total number of individuals in the data set.

$$\begin{aligned} P(\text{widowed}) &= \frac{14.3}{251.3} \\ &= 0.057 \end{aligned}$$

TABLE 5

	Males (in millions)	Females (in millions)	Totals (in millions)
Never Married	41.6	36.9	78.5
Married	64.4	63.1	127.5
Widowed	3.1	11.2	14.3
Divorced	11.0	14.4	25.4
Separated	2.4	3.2	5.6
Totals (in millions)	122.5	128.8	251.3

Suppose we know that the individual is female. Does this change the probability that she is widowed? The sample space now consists only of females; so the probability that the individual is widowed, given that the individual is female, is

$$\begin{aligned} P(\text{widowed}|\text{female}) &= \frac{N(\text{widowed females})}{N(\text{females})} \\ &= \frac{11.2}{128.8} \\ &= 0.087 \end{aligned}$$

So knowing that the individual is female increases the likelihood that the individual is widowed (from 0.057 to 0.087). This leads to the following rule.

Conditional Probability Rule

If E and F are any two events, then

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

The probability of event F occurring, given the occurrence of event E , is found by dividing the probability of E and F by the probability of E , or by dividing the number of outcomes in E and F by the number of outcomes in E .

EXAMPLE 2 Conditional Probabilities on Marital Status and Gender

Problem

The data in Table 5 represent the marital status and gender of U.S. residents aged 15 years and older in 2013.

Video Solution



Part (A)

Part (B)

Compute the probability that a randomly selected individual never married, given that the individual is male.

Approach

Because the randomly selected person is male, concentrate on the “male” column. There are 122.5 million males and 41.6 million males who never married, so $N(\text{male}) = 122.5$ million and $N(\text{male and never married}) = 41.6$ million. Compute the probability using the [Conditional Probability Rule](#).

Solution

Substituting into the Conditional Probability Rule, we obtain

$$\begin{aligned} P(\text{never married}|\text{male}) &= \frac{N(\text{never married and male})}{N(\text{male})} \\ &= \frac{41.6}{122.5} \\ &= 0.340 \end{aligned}$$

The probability that the randomly selected individual never married, given that he is male, is 0.340.

Compute the probability that a randomly selected individual is male, given that the individual never married.

Approach

Because the randomly selected person never married, concentrate on the “never married” row. There are 78.5 million people who never married and 41.6 million males who never married, so

$$N(\text{never married}) = 78.5 \text{ million and } N(\text{male and never married}) = 41.6 \text{ million.}$$

Compute the probability using the [Conditional Probability Rule](#).

Solution

Substituting into the Conditional Probability Rule, we obtain

$$\begin{aligned} P(\text{male}|\text{never married}) &= \frac{N(\text{male and never married})}{N(\text{never married})} \\ &= \frac{41.6}{78.5} \\ &= 0.530 \end{aligned}$$

The probability that the randomly selected individual is male, given that the individual never married, is 0.530.

What is the difference between the results of parts (A) and (B)? In part (A), we found that 34.0% of males never married, whereas in part (B), we found that 53.0% of individuals who never married are male.

EXAMPLE 3 Birth Weights of Preterm Babies

Problem

Suppose that 12.2% of all births are preterm. (Preterm means that the gestation period of the pregnancy is less than 37 weeks.) Also, 0.2% of all births result in a preterm baby who weighs 8 pounds, 13 ounces or more. What is the probability that a randomly selected baby weighs 8 pounds, 13 ounces or more, given that the baby is preterm? Is this unusual? Data based on the Vital Statistics Reports.

Video Solution



Approach

We want to know the probability that the baby weighs 8 pounds, 13 ounces or more, given that the baby is preterm, $P(8 \text{ pounds, } 13 \text{ ounces or more}|\text{preterm})$. Because 0.2% of all babies weigh 8 pounds, 13 ounces or more and are preterm, $P(\text{weighs 8 lb, 13 oz or more and preterm}) = 0.002$. Because 12.2% of all births are preterm, $P(\text{preterm}) = 0.122$. The phrase “given that” suggests we use the [Conditional Probability Rule](#) to compute the probability.

Solution

$$\begin{aligned} P(\text{weighs 8 lb, 13 oz or more}|\text{preterm}) &= \frac{P(\text{weighs 8 lb, 13 oz or more and preterm})}{P(\text{preterm})} \\ &= \frac{0.002}{0.122} \approx 0.0164 \end{aligned}$$

If we randomly selected 100 preterm babies, we would expect about 2 of them to weigh 8 pounds, 13 ounces or more. This would be an unusual result (remember, we typically say that events with probability less than 0.05 are unusual).

5.4 Objective 2 Compute Probabilities Using the General Multiplication Rule

October 23, 2016 12:10 PM

Conditional Probability Rule

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

If we solve the [Conditional Probability Rule](#) for $P(E \text{ and } F)$, we obtain the General Multiplication Rule.

General Multiplication Rule

The probability that two events E and F both occur is

$$P(E \text{ and } F) = P(E) \cdot P(F|E)$$

EXAMPLE 4 Using the General Multiplication Rule

Problem

The probability that a driver who is speeding gets pulled over is 0.8. The probability that a driver gets a ticket, given that he or she is pulled over, is 0.9.

What is the probability that a randomly selected driver who is speeding gets pulled over and gets a ticket?

Video Solution



Approach

Let E represent the event "driver who is speeding gets pulled over" and let F represent the event "driver gets a ticket." We use the [General Multiplication Rule](#) to compute $P(E \text{ and } F)$.

Solution

$$\begin{aligned} P(\text{driver who is speeding gets pulled over and gets a ticket}) &= P(E \text{ and } F) \\ &= P(E) \cdot P(F|E) \\ &= 0.8(0.9) \\ &= 0.72 \end{aligned}$$

The probability that a driver who is speeding gets pulled over and gets a ticket is 0.72.

EXAMPLE 5 Acceptance Sampling

Problem

Suppose that of 100 circuits sent to a manufacturing plant, 5 are defective. The plant manager receiving the circuits randomly selects two and tests them. If both circuits work, she will accept the shipment. Otherwise, the shipment is rejected. What is the probability that the plant manager discovers at least one defective circuit and rejects the shipment?

Video Solution



Approach #1 Approach #2

To determine the probability that at least one of the tested circuits is defective, consider four possibilities. Neither of the circuits is defective, the first is defective while the second is not, the first is not defective while the second is defective, or both circuits are defective. Note that the outcomes are not equally likely. We might use two different approaches.

The first approach is to use a tree diagram to list all possible outcomes and the [General Multiplication Rule](#) to compute the probability of each outcome. Then determine the probability of at least one defective by adding the probability that only the first is defective, only the second is defective, or both are defective, using the [Addition Rule](#) (because the events are disjoint).

Solution

Of the 100 circuits, 5 are defective, so 95 are not defective. Construct a tree diagram to determine the possible outcomes for the experiment. See Figure 6, where D stands for defective and G stands for good (not defective). Because the outcomes are not equally likely, we include the probabilities in our diagram to show how the probability of each outcome is obtained. (Multiply the individual probabilities along the corresponding path in the diagram.)

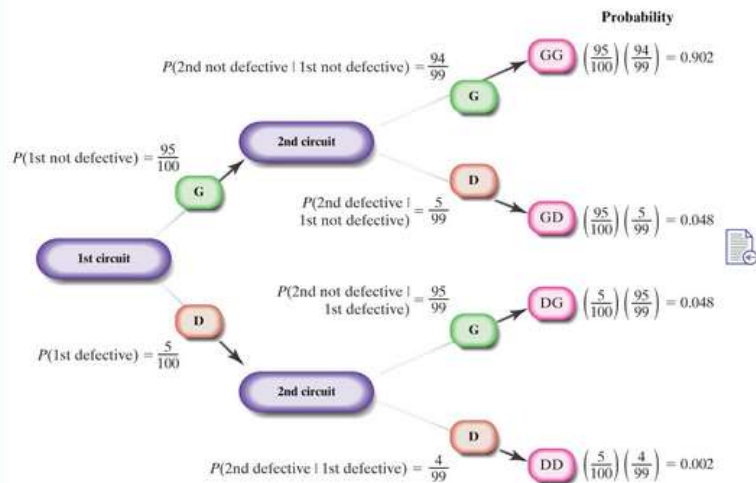


Figure 6

From our tree diagram, and using the [Addition Rule](#), we can write

$$\begin{aligned}
 P(\text{at least 1 defective}) &= P(\text{GD}) + P(\text{DG}) + P(\text{DD}) \\
 &= 0.048 + 0.048 + 0.002 \\
 &= 0.098
 \end{aligned}$$

The probability that the shipment will not be accepted is 0.098.

HIDE SOLUTION

The second approach is to compute the probability that both circuits are not defective and use the [Complement Rule](#) to determine the probability of at least one defective circuit.

Solution

$$\begin{aligned}
 P(\text{at least 1 defective}) &= 1 - P(\text{none defective}) \\
 &= 1 - P(\text{1st not defective}) \cdot P(\text{2nd not defective} \mid \text{1st not defective}) \\
 &= 1 - \frac{95}{100} \cdot \frac{94}{99} \\
 &= 1 - 0.902 \\
 &= 0.098
 \end{aligned}$$

The probability that the shipment will not be accepted is 0.098.

HIDE SOLUTION

EXAMPLE 6 Sickle-Cell Anemia

Problem

A survey of 10,000 African Americans found that 27 had sickle-cell anemia.

Video Solution



Part (A)

Part (B)

Part (C)

Suppose we randomly select 1 of the 10,000 African Americans surveyed. What is the probability that he or she has sickle-cell anemia?

Approach

Let the event E = "sickle-cell anemia"; so $P(E)$ = number of African Americans who have sickle-cell anemia divided by the number in the survey.

Solution

If one individual is selected,

$$P(E) = \frac{27}{10,000} = 0.0027$$

If two individuals from this group are randomly selected, what is the probability that both have sickle-cell anemia?

Approach

Let E_1 = "first person has sickle-cell anemia" and E_2 = "second person has sickle-cell anemia"; then compute $P(E_1 \text{ and } E_2) = P(E_1) P(E_2|E_1)$.

Solution

Using the [General Multiplication Rule](#), we have

$$\begin{aligned} P(E_1 \text{ and } E_2) &= P(E_1) \cdot P(E_2|E_1) \\ &= \frac{27}{10,000} \cdot \frac{26}{9999} \\ &\approx 0.00000702 \end{aligned}$$

Notice that $P(E_2|E_1) = \frac{26}{9999}$ because we are sampling without replacement, so after event E_1 occurs, there is one less person with sickle-cell anemia and one less person in the sample space.

Compute the probability of randomly selecting two individuals from this group who have sickle-cell anemia, assuming independence.

Approach

Use the [Multiplication Rule for Independent Events](#).

Solution

The assumption of independence means that the outcome of the first trial of the experiment does not affect the probability of the second trial. (It is like sampling with replacement.) Therefore, we assume that

$$P(E_1) = P(E_2) = \frac{27}{10,000}$$

Then

$$\begin{aligned} P(E_1 \text{ and } E_2) &= P(E_1) \cdot P(E_2) \\ &= \frac{27}{10,000} \cdot \frac{27}{10,000} \\ &\approx 0.00000729 \end{aligned}$$

Independence and Small Samples

Notice that the probabilities from Examples 6(B) and 6(C) are extremely close.

- In Example 6(B) the probability of randomly selecting two African Americans from a population of 10,000 who have sickle-cell anemia without replacement is 0.00000702.
- In Example 6(C), the probability of randomly selecting two African Americans from a population of 10,000 who have sickle-cell anemia with replacement—that is, assuming independence—is 0.00000729.

Based on this, we infer the following principle.

If small random samples are taken from large populations without replacement, it is reasonable to assume independence of the events. As a rule of thumb, if the sample size, n , is less than 5% of the population size, N , we treat the events as independent. That is, if $n < 0.05N$, treat the events as independent.

In Example 6, we can compute the probability of randomly selecting two African Americans who have sickle-cell anemia assuming independence because the sample size, 2, is only $\frac{2}{10,000} = 0.02\%$ of the population size, 10,000.

We now express independence using conditional probabilities.

DEFINITION

Two events E and F are independent if $P(E|F) = P(E)$ or, equivalently, if $P(F|E) = P(F)$.

Consider the data in Table 5. Because $P(\text{widowed}) = 0.057$ does not equal $P(\text{widowed}|\text{female}) = 0.087$, the events “widowed” and “female” are not independent. In fact, knowing an individual is female increases the likelihood that the individual is also widowed.

5.5 Computing Techniques

October 24, 2016 09:02 AM

5.5 Objective 1 Solve Counting Problems using the Multiplication Rule

October 24, 2016 09:03 AM

Multiplication Rule of Counting

If a task consists of a sequence of choices in which there are p selections for the first choice, q selections for the second choice, r selections for the third choice, and so on, then the task of making these selections can be done in

$$p \cdot q \cdot r \cdots$$

different ways.

EXAMPLE 2 Counting Airport Codes (Repetition Allowed)

Problem

The International Airline Transportation Association (IATA) assigns three-letter codes to represent airport locations. For example, the code for Fort Lauderdale International Airport is FLL. How many different airport codes are possible?

Video Solution



Approach

We are choosing 3 letters from 26 letters and arranging them in order. Notice that repetition of letters is allowed. Use the [Multiplication Rule of Counting](#), recognizing that we have 26 ways to choose the first letter, 26 ways to choose the second letter, and 26 ways to choose the third letter.

Solution

By the Multiplication Rule,

$$\begin{aligned} 26 \cdot 26 \cdot 26 &= 26^3 \\ &= 17,576 \end{aligned}$$

different airport codes are possible.

EXAMPLE 3 Counting (Without Repetition)

Problem

Three members from a 14-member committee are to be randomly selected to serve as chair, vice-chair, and secretary. The first person selected is the chair, the second is the vice-chair, and the third is the secretary. How many different committee structures are possible?

Video Solution



Approach

The task consists of making three selections. The first selection requires choosing from 14 members. Because a member cannot serve in more than one capacity, the second selection requires choosing from the 13 remaining members. The third selection requires choosing from the 12 remaining members. (Do you see why?) We use the [Multiplication Rule of Counting](#) to determine the number of possible committee structures.

Solution

By the Multiplication Rule,

$$14 \cdot 13 \cdot 12 = 2184$$

different committee structures are possible.

The Factorial Symbol

We now introduce a special symbol.

DEFINITION

If $n \geq 0$ is an integer, the **factorial symbol**, $n!$, is defined as follows:

$$n! = n(n-1) \cdot \dots \cdot 3 \cdot 2 \cdot 1$$

$$0! = 1 \quad 1! = 1$$

For example,

$$\begin{array}{l} 2! = 2 \cdot 1 \\ \quad = 2 \end{array} \quad \begin{array}{l} 3! = 3 \cdot 2 \cdot 1 \\ \quad = 6 \end{array} \quad \begin{array}{l} 4! = 4 \cdot 3 \cdot 2 \cdot 1 \\ \quad = 24 \end{array}$$

and so on. Table 6 lists values for $n!$ for $0 \leq n \leq 6$. Notice how quickly the values of the factorials increase.

TABLE 6

n	0	1	2	3	4	5	6
$n!$	1	1	2	6	24	120	720

5.5 Objective 2 Solve Counting Problem Using Permutations

October 24, 2016 09:06 AM

The solutions in [Example 3](#) and [Example 4](#) illustrate a type of counting problem referred to as a *permutation*.

DEFINITION

A **permutation** is an ordered arrangement in which r objects are chosen from n distinct (different) objects so that $r \leq n$ and repetition is not allowed. The symbol ${}_n P_r$ represents the number of permutations of r objects selected from n objects.

The solution in Example 3 could be represented as

$$\begin{aligned} {}_n P_r &= {}_{14} P_3 \\ &= 14 \cdot 13 \cdot 12 \\ &= 2184 \end{aligned}$$

and the solution to Example 4 as

$$\begin{aligned} {}_7 P_7 &= 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \\ &= 5040 \end{aligned}$$

Deriving a Formula for the Number of Permutations

To arrive at a formula for ${}_n P_r$, we note that there are n choices for the first selection, $n - 1$ choices for the second selection, $n - 2$ choices for the third selection, ..., and $n - (r - 1)$ choices for the r th selection. By the [Multiplication Rule](#),

$$\begin{aligned} {}_n P_r &= \overset{\text{1st}}{n} \cdot \overset{\text{2nd}}{(n-1)} \cdot \overset{\text{3rd}}{(n-2)} \cdot \dots \cdot \overset{\text{rth}}{[n-(r-1)]} \\ &= n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1) \end{aligned}$$

This formula for ${}_n P_r$ can be written in factorial notation:

$$\begin{aligned} {}_n P_r &= n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1) \\ &= n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1) \cdot \frac{(n-r) \cdot \dots \cdot 3 \cdot 2 \cdot 1}{(n-r) \cdot \dots \cdot 3 \cdot 2 \cdot 1} \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

Number of Permutations of n Distinct Objects Taken r at a Time

The formula below gives the number of arrangements of r objects chosen from n objects, in which

1. The n objects are distinct
2. Repetition of objects is not allowed, and
3. Order is important (so ABC is different from BCA)

$${}_n P_r = \frac{n!}{(n-r)!}$$

EXAMPLE 5 Computing Permutations**Problem**Evaluate: (a) ${}_7P_5$ (b) ${}_5P_5$

Video Solution



Technology Step-By-Step



Part (A)

Part (B)

ApproachUse the formula ${}_nP_r = \frac{n!}{(n-r)!}$ where $n = 7$ and $r = 5$.**Solution**

$$\begin{aligned}
 {}_7P_5 &= \frac{7!}{(7-5)!} \\
 &= \frac{7!}{2!} \\
 &= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2!}{2!} \\
 &= \underbrace{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3}_{5 \text{ factors}} \\
 &= 2520
 \end{aligned}$$

Part (A)

Part (B)

ApproachUse the formula ${}_nP_r = \frac{n!}{(n-r)!}$ where $n = 5$ and $r = 5$.**Solution**

$$\begin{aligned}
 {}_5P_5 &= \frac{5!}{(5-5)!} \\
 &= \frac{5!}{0!} \\
 &= \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} \\
 &= 120
 \end{aligned}$$

0! Always equals 1

EXAMPLE 6 Betting the Trifecta

Problem

In how many ways can horses in a ten-horse race finish first, second, and third?

Approach

The ten horses are distinct. Once a horse crosses the finish line, that horse will not cross the finish line again, and in a race, finishing order is important. We have a permutation of ten objects taken three at a time.

Solution

The top three horses can finish a ten-horse race in

$$\begin{aligned} {}_{10}P_3 &= \frac{10!}{(10-3)!} \\ &= \frac{10!}{7!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7!}{7!} \\ &= \underbrace{10 \cdot 9 \cdot 8}_{\text{3 factors}} \\ &= 720 \text{ ways} \end{aligned}$$

Video Solution



Technology Step-By-Step



5.5 Objective 3 Solve Counting Problems using Combinations

October 24, 2016 09:12 AM

In a permutation, order is important. For example, the arrangements ABC and BAC are considered different arrangements of the letters A , B , and C . If order is unimportant, we do not distinguish ABC from BAC . In poker, the order in which the cards are received does not matter. The *combination* of the cards is what matters.

DEFINITION

A **combination** is a collection, without regard to order, in which r objects are chosen from n distinct objects with $r \leq n$ without repetition. The symbol ${}_n C_r$ represents the number of combinations of n distinct objects taken r at a time.

EXAMPLE 7 Listing Combinations

Problem

Roger, Ken, Tom, and Jay are going to play golf. They will randomly select teams of two players each. List all possible team combinations. That is, list all the combinations of the four people Roger, Ken, Tom, and Jay taken two at a time. What is ${}_4 C_2$?

Video Solution



Approach

List the possible teams. Note that order is unimportant, so {Roger, Ken} is the same as {Ken, Roger}.

Solution

The list of all such teams (combinations) is

Roger, Ken Roger, Tom Roger, Jay Ken, Tom Ken, Jay Tom, Jay

So

$${}_4 C_2 = 6$$

There are six ways of forming teams of two from a group of four players.

Deriving a Formula for the Number of Combinations

We can find a formula for ${}_n C_r$ by noting that the only difference between a permutation and a combination is that we disregard order in combinations. To determine ${}_n C_r$, we eliminate from the formula for ${}_n P_r$ the number of permutations that were rearrangements of a given set of r objects. In Example 7, for example, selecting $\{\text{Roger, Ken}\}$ was the same as selecting $\{\text{Ken, Roger}\}$; so there were $2! = 2$ rearrangements of the two objects. This can be determined from the formula for ${}_n P_r$ by calculating ${}_r P_r = r!$. So, if we divide ${}_n P_r$ by $r!$, we will have the desired formula for ${}_n C_r$:

$${}_n C_r = \frac{{}_n P_r}{r!} = \frac{n!}{r!(n-r)!}$$

Number of Combinations of n Distinct Objects Taken r at a Time

The formula below gives the number of arrangements of r objects chosen from n objects, in which

1. The n objects are distinct
2. Repetition of objects is not allowed, and
3. Order is not important

$${}_n C_r = \frac{n!}{r!(n-r)!}$$

EXAMPLE 8 Computing Combinations

Problem

Evaluate: (a) ${}_4 C_1$ (b) ${}_6 C_4$ (c) ${}_6 C_2$

Video Solution



Technology Step-By-Step



Part (A)

Part (B)

Part (C)

Approach

Use the formula ${}_n C_r = \frac{n!}{r!(n-r)!}$ where $n = 4$ and $r = 1$.

Solution

$$\begin{aligned} {}_4 C_1 &= \frac{4!}{1!(4-1)!} \\ &= \frac{4!}{1! \cdot 3!} \\ &= \frac{4 \cdot 3!}{1 \cdot 3!} \\ &= 4 \end{aligned}$$

Approach

Use the formula ${}_n C_r = \frac{n!}{r!(n-r)!}$ where $n = 6$ and $r = 4$.

Solution

$$\begin{aligned} {}_6 C_4 &= \frac{6!}{4!(6-4)!} \\ &= \frac{6!}{4! \cdot 2!} \\ &= \frac{6 \cdot 5 \cdot 4!}{4! \cdot 2 \cdot 1} \\ &= \frac{30}{2} \\ &= 15 \end{aligned}$$

Approach

Use the formula ${}_n C_r = \frac{n!}{r!(n-r)!}$ where $n = 6$ and $r = 2$.

Solution

$$\begin{aligned} {}_6 C_2 &= \frac{6!}{2!(6-2)!} \\ &= \frac{6!}{2! \cdot 4!} \\ &= \frac{6 \cdot 5 \cdot 4!}{2 \cdot 1 \cdot 4!} \\ &= \frac{30}{2} \\ &= 15 \end{aligned}$$

EXAMPLE 9 Simple Random Samples

Problem

How many different **simple random samples** of size 4 can be obtained from a population whose size is 20?

Video Solution



Approach

The 20 individuals in the population are distinct. In addition, the order in which individuals are selected is unimportant. Thus, the number of simple random samples of size 4 from a population of size 20 is a combination of $n = 20$ objects taken $r = 4$ at a time. Use the formula ${}_n C_r = \frac{n!}{r!(n-r)!}$ where $n = 20$ and $r = 4$.

Technology Step-By-Step



Solution

$$\begin{aligned} {}_{20}C_4 &= \frac{20!}{4!(20-4)!} \\ &= \frac{20!}{4! \cdot 16!} \\ &= \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16!}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 16!} \\ &= \frac{116,280}{24} \\ &= 4845 \end{aligned}$$

There are 4845 different simple random samples of size 4 from a population whose size is 20.

5.5 Objective 4 Solve Counting Problems Involving Permutations with Non-distinct Items

October 24, 2016 09:15 AM

Example – DNA Sequence

A DNA sequence consists of a series of letters representing a DNA strand that spells out the genetic code. There are four possible letters (A, C, G, and T), each representing a specific nucleotide base in the DNA strand (adenine, cytosine, guanine, and thymine, respectively). How many distinguishable sequences can be formed using two As, two Cs, three Gs, and one T?

The process of forming a sequence consists of four tasks:

Task 1: Choose the positions for the two As. 8C_2

Task 2: Choose the positions for the two Cs. 6C_2

Task 3: Choose the positions for the three Gs.

Task 4: Choose the position for the one T.

The process of forming a sequence consists of four tasks:

Task 1: Choose the positions for the two As. 8C_2

Task 2: Choose the positions for the two Cs. 6C_2

Task 3: Choose the positions for the three Gs.

Task 4: Choose the position for the one T.

The process of forming a sequence consists of four tasks:

Task 1: Choose the positions for the two As. 8C_2

Task 2: Choose the positions for the two Cs. 6C_2

Task 3: Choose the positions for the three Gs. 4C_3

Task 4: Choose the position for the one T. 1C_1

$$\begin{aligned} & {}^8C_2 \cdot {}^6C_2 \cdot {}^4C_3 \cdot {}^1C_1 \\ &= \frac{8!}{2! \cdot 6!} \cdot \frac{6!}{2! \cdot 4!} \cdot \frac{4!}{3! \cdot 1!} \cdot \frac{1!}{1! \cdot 0!} \\ &= \frac{8!}{2! \cdot 2! \cdot 3! \cdot 1! \cdot 0!} \\ &= 1680 \end{aligned}$$

A Formula for Finding the Number of Permutations with Nondistinct Items

Example 10 suggests a general result. Had the letters in the sequence each been different, ${}_8P_8 = 8!$ possible sequences would have been formed. This is the numerator of the answer. The presence of two A's, two C's, and three G's reduces the number of different sequences, as the entries in the denominator illustrate. We are led to the following result:

Permutations with Nondistinct Items

The number of permutations of n objects of which n_1 are of one kind, n_2 are of a second kind, ..., and n_k are of a k th kind is given by

$$\frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_k!}$$

where $n = n_1 + n_2 + \cdots + n_k$.

EXAMPLE 11 Arranging Flags

Problem

How many different vertical arrangements are there of ten flags if five are white, three are blue, and two are red?

Video Solution



Approach

Because there are nondistinct items and order matters, use the formula for finding the number of permutations with nondistinct items. We seek the number of permutations of $n = 10$ objects, of which $n_1 = 5$ are of one kind (white), $n_2 = 3$ are of a second kind (blue), and $n_3 = 2$ are of a third kind (red).

Solution

$$\begin{aligned}\frac{n!}{n_1! \cdot n_2! \cdot n_3!} &= \frac{10!}{5! \cdot 3! \cdot 2!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{5! \cdot 3! \cdot 2!} \\ &= 2520 \text{ different vertical arrangements}\end{aligned}$$

One of the challenges in solving counting problems is selecting the appropriate formula for the given situation. The table below reviews the situations in which each counting problem applies.

	Description	Formula
Combination	The selection of r objects from a set of n different objects when the order in which the objects are selected does not matter (so AB is the same as BA) and an object cannot be selected more than once (repetition is not allowed)	${}_n C_r = \frac{n!}{r!(n-r)!}$
Permutation of Distinct Items with Replacement	The selection of r objects from a set of n different objects when the order in which the objects are selected matters (so AB is different from BA) and an object may be selected more than once (repetition is allowed)	n^r
Permutation of Distinct Items without Replacement	The selection of r objects from a set of n different objects when the order in which the objects are selected matters so (AB is different from BA) and an object cannot be selected more than once (repetition is not allowed)	${}_n P_r = \frac{n!}{(n-r)!}$
Permutation of Nondistinct Items without Replacement	The number of ways n objects can be arranged (order matters) in which there are n_1 of one kind, n_2 of a second kind, \dots , and n_k of a k th kind, where $n = n_1 + n_2 + \dots + n_k$	$\frac{n!}{n_1! n_2! \cdots n_k!}$

5.5 Objective 5 Compute Probabilities Involving Permutations and Combinations

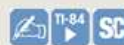
October 24, 2016 09:26 AM

EXAMPLE 12 Winning the Lottery

Problem

In the Illinois Lottery, an urn contains balls numbered 1 to 52. From this urn, six balls are randomly chosen without replacement. For a \$1 bet, a player chooses two sets of six numbers. To win, all six numbers must match those chosen from the urn. The order in which the balls are picked does not matter. What is the probability of winning the lottery?

Video Solution



Technology Step-By-Step



Approach

The probability of winning is given by the number of ways a ticket could win divided by the size of the sample space. Each ticket has two sets of six numbers and therefore two chances of winning. The size of the sample space S is the number of ways 6 objects can be selected from 52 objects without replacement and without regard to order, so $N(S) = {}_{52}C_6$.

Solution

The size of the sample space is

$$N(S) = {}_{52}C_6 = \frac{52!}{6! \cdot (52 - 6)!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46!}{6! \cdot 46!} = 20,358,520$$

Each ticket has two chances of winning. If E is the event "winning ticket," then $N(E) = 2$ and

$$P(E) = \frac{2}{20,358,520} = 0.000000098$$

There is about a 1 in 10,000,000 chance of winning the Illinois Lottery!

EXAMPLE 13 Acceptance Sampling

Problem

A shipment of 120 fasteners that contains 4 defective fasteners was sent to a manufacturing plant. The plant's quality control manager randomly selects and inspects 5 fasteners. What is the probability that exactly 1 of the inspected fasteners is defective?

Video Solution



Technology Step-By-Step



Approach

Find the probability that exactly 1 fastener is defective by calculating the number of ways of selecting exactly 1 defective fastener in 5 fasteners and dividing this result by the number of ways of selecting 5 fasteners from 120 fasteners. To choose exactly 1 defective in the 5 requires choosing 1 defective from the 4 defectives and 4 nondefectives from the 116 nondefectives. The order in which the fasteners are selected does not matter, so we use combinations.

Solution

The number of ways of choosing 1 defective fastener from 4 defective fasteners is ${}_4C_1$. The number of ways of choosing 4 nondefective fasteners from 116 nondefectives is ${}_{116}C_4$. Using the **Multiplication Rule**, we find that the number of ways of choosing 1 defective and 4 nondefective fasteners is

$$\begin{aligned}({}_4C_1) \cdot ({}_{116}C_4) &= 4 \cdot 7,160,245 \\ &= 28,640,980\end{aligned}$$

The number of ways of selecting 5 fasteners from 120 fasteners is ${}_{120}C_5 = 190,578,024$. The probability of selecting exactly 1 defective fastener is

$$P(1 \text{ defective fastener}) = \frac{({}_4C_1) \cdot ({}_{116}C_4)}{{}_{120}C_5} = \frac{28,640,980}{190,578,024} = 0.1503$$

The probability of randomly selecting exactly one defective fastener is 0.1503. If we selected 5 fasteners, 100 different times, we would expect about 15 of the samples to have exactly one defective fastener.

5.5.55

October 24, 2016 09:54 AM

5.5 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 10/24/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 5.5 Interactive Assignment

How many different 10-letter words (real or imaginary) can be formed from the following letters?

Y, Y, T, A, J, Z, Y, S, I, X

Before you begin calculating, consider some questions about the situation.

Does order of the letters in a word matter?

- Yes
 No

Are all of the letters distinct?

- Yes
 No

Can the any of the ten given letters be used more than once?

- Yes
 No

To create a word from a given set of letters, we are arranging the letters in a certain order. However, in this case we have the added complication that we have been given some letters which are duplicates of each other. Since we can use each of the ten letters we have been given only once, replacements are not allowed. In this situation we count permutations of nondistinct items without replacement.

A permutation of nondistinct items without replacement is the number of ways n objects can be arranged (order matters) in which there are n_1 of one kind, n_2 of a second kind, and n_k of a k th kind, where $n = n_1 + n_2 + \dots + n_k$. The number of such permutations is given by the following formula.

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

In the expression below, the appropriate numbers have been substituted into the formula. Evaluate this expression to find the number of ten-letter words (real or imaginary) that can be formed.

$$\frac{10!}{3! \cdot 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1!} = 604800 \quad (\text{Type a whole number.})$$

Therefore, 604,800 ten-letter words (real or imaginary) that can be formed with the letters above.

YOU ANSWERED: No

Yes

5.5.61

October 24, 2016 09:54 AM

5.5 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 10/24/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 5.5 Interactive Assignment

The grade appeal process at a university requires that a jury be structured by selecting five individuals randomly from a pool of five students and seven faculty. (a) What is the probability of selecting a jury of all students? (b) What is the probability of selecting a jury of all faculty? (c) What is the probability of selecting a jury of three students and two faculty?

(a) Recall that according to the classical method, if an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m , then the probability of E , $P(E)$, is given by the following formula.

$$P(E) = \frac{\text{Number of ways that } E \text{ can occur}}{\text{Number of possible outcomes}} = \frac{m}{n}$$

The first step in solving this problem is to find the number of possible juries. To do this disregard whether an individual is a student or faculty member and treat the jury pool as a whole.

Are the individuals in the combined jury pool distinct?

- No
 Yes

Can an individual be chosen more than once per jury?

- Yes
 No

Is the order in which individuals are selected for the jury important?

- No
 Yes

In this pool, the individuals are all distinct. Each person will be chosen only once per jury, and the order in which they are chosen does not matter. Given these facts, the situation lends itself to counting by combinations.

A combination is the selection of r objects from a set of n different objects when the order in which the objects is selected does not matter (so AB is the same as BA) and an object cannot be selected more than once (repetition is not allowed). The number of such combinations involving n objects and r choices is given by the following formula.

$${}_n C_r = \frac{n!}{r!(n-r)!}$$

How large is the combined student-faculty jury pool?

(Type a whole number.)

How many people are chosen for a jury?

(Type a whole number.)

Use the combination formula and the numbers above to find how many different juries are possible.

$${}_{12} C_5 = \frac{12!}{5!(12-5)!} = \text{ } \text{ (Type a whole number.)}$$

There are 792 possible juries. Now find the number of possible all-student juries. The number of ways an all-student jury can be chosen is given by letting only the students be the current jury pool and finding how many juries can be made from that pool.

For a student only jury, the jurors can only be chosen from the students. How many students are available?

(Type a whole number.)

Given that the jury needs five members, use the combination formula to calculate the number of juries that can be made from only the students.

$${}_5C_5 = \frac{5!}{5!(5-5)!} = \text{ } \text{ (Type a whole number.)}$$

Now use the classical method to find the probability of selecting an all-student jury.

$$\begin{aligned} P(\text{jury is all students}) &= \frac{\text{Number of juries with only students}}{\text{Number of possible juries}} \\ &= \text{ } \text{ (Round to five decimal places as needed.)} \end{aligned}$$

Thus 0.00126 is the probability that an all-student jury is chosen.

(b) To find the probability that an all-faculty jury is chosen, follow a similar procedure. How many faculty members are available?

(Type a whole number.)

Given that the jury needs five members, use the combination formula to calculate the number of juries that can be made from only the faculty.

$${}_7C_5 = \frac{7!}{5!(7-5)!} = \text{ } \text{ (Type a whole number.)}$$

Now use the classical method to find the probability of selecting an all-faculty jury

$$\begin{aligned} P(\text{jury is all faculty}) &= \frac{\text{Number of juries with only faculty}}{\text{Number of possible juries}} \\ &= \text{ } \text{ (Round to five decimal places as needed.)} \end{aligned}$$

The probability of an all-faculty jury is 0.02652.

(c) Now we would like to find the probability of selecting a jury of three students and two faculty.

To find the number of possible juries under these conditions, you can use both the multiplicative rule of counting and combinations. First find the number of ways to choose three students, and then find the number of ways to choose two faculty.

Use the combination formula to find the number of ways that three jurors can be chosen from the five students.

$${}_5C_3 = \frac{5!}{3!(5-3)!} = \text{ } \text{ (Type a whole number.)}$$

Again use the combination formula to find the number of ways that two jurors can be chosen from the seven faculty.

$${}_7C_2 = \frac{7!}{2!(7-2)!} = \text{ } \text{ (Type a whole number.)}$$

The multiplication rule of counting states that if a task consists of a sequence of choices in which there are p selections for the first choice, q selections for the second choice, r selections for the third choice, and so on, then the task of making these selections can be done in $p \cdot q \cdot r \cdot \dots$ different ways.

Multiply the number of ways to choose three students by the number of ways to choose two faculty.

$$10 \cdot 21 = \text{ } \text{ (Type a whole number.)}$$

Finally, to find the answer to the last part of the problem, use the classical method to compute the probability that a jury of three students and two faculty will be selected.

$$\begin{aligned} P(\text{mixed jury}) &= \frac{\text{Number of ways to chose two faculty and three students}}{\text{Number of possible juries}} \\ &= \text{0.26515} \quad (\text{Round to five decimal places as needed.}) \end{aligned}$$

Therefore, A jury of three students and two faculty is selected with probability 0.26515.

YOU ANSWERED: No

.29167

5.6 Putting it Together: Which Method Do I Use?

October 25, 2016 08:17 AM

5.6 Objective 1 Determine the Appropriate Rule to Use

October 25, 2016 08:19 AM

Working with probabilities can be challenging because of the number of different probability rules. This section will help you learn when to use a particular rule. To aid you, consider the flowchart in Figure 7. While not all situations can be handled directly with the formulas provided, they can be combined and expanded to many more situations. The [video](#) explains how to use the flowchart.

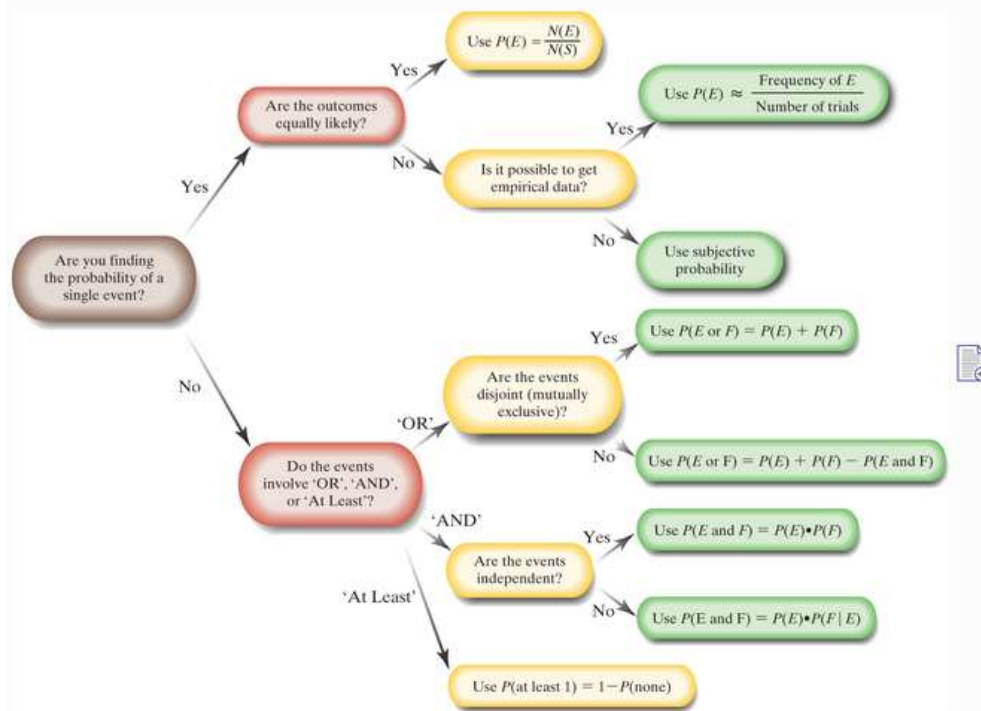


Figure 7

The first step is to determine whether we are finding the probability of a single event. If we are dealing with a single event, we must decide whether to use the classical method (equally likely outcomes), the empirical method (relative frequencies), or subjective probability. For experiments involving more than one event, we first decide which type of statement we have. For events involving 'AND', we must know if the events are **independent**. For events involving 'OR', we need to know if the events are **disjoint** (mutually exclusive).

EXAMPLE 1 Probability: Which Rule Do I Use?

Problem

In the game show *Deal or No Deal?*, a contestant is presented with 26 suitcases that contain amounts ranging from \$0.01 to \$1,000,000. The contestant must pick an initial case that is set aside as the game progresses. The amounts are randomly distributed among the suitcases prior to the game as shown in Table 7. What is the probability that the contestant picks a case worth at least \$100,000?

Video Solution



TABLE 7

Prize	Number of Suitcases
\$0.01–\$100	8
\$200–\$1000	6
\$5000–\$50,000	5
\$100,000–\$1,000,000	7

Approach

Follow the flowchart in [Figure 7](#) to determine which formula to use.

Solution

There is a single event, so we must decide among the empirical, classical, or subjective approaches to determine the probability. The probability experiment is selecting a suitcase. Each prize amount is randomly assigned to one of the 26 suitcases, so the outcomes are equally likely. Table 7 shows that seven cases contain at least \$100,000. Letting $E =$ “worth at least \$100,000,” we compute $P(E)$ using the classical approach.

$$P(E) = \frac{N(E)}{N(S)} = \frac{7}{26} \approx 0.269$$

The probability that the contestant selects a suitcase worth at least \$100,000 is 0.269. In 100 different games, we would expect about 27 games to result in a contestant choosing a suitcase worth at least \$100,000.

EXAMPLE 2 Probability: Which Rule Do I Use?

Problem

According to a Harris poll, 14% of adult Americans have one or more tattoos, 50% have pierced ears, and 65% of those with one or more tattoos also have pierced ears. What is the probability that a randomly selected adult American has one or more tattoos and pierced ears?

Video Solution



Approach

Follow the flowchart in [Figure 7](#) to determine which formula to use.

Solution

We are finding the probability of an event involving 'AND'. Letting T = "one or more tattoos" and E = "ears pierced," we must find $P(T \text{ and } E)$. We need to determine if the two events, T and E , are independent. The problem statement tells us that $P(T) = 0.14$, $P(E) = 0.50$, and $P(E|T) = 0.65$. Because $P(E) \neq P(E|T)$, the two events are not independent. We find $P(T \text{ and } E)$ using the [General Multiplication Rule](#).

$$\begin{aligned}P(T \text{ and } E) &= P(T) \cdot P(E|T) \\ &= (0.14)(0.65) \\ &= 0.091\end{aligned}$$

So the probability of selecting an adult American at random who has one or more tattoos and pierced ears is **0.091**.

5.6 Objective 2 Determine the Appropriate Counting Technique to Use

October 25, 2016 08:36 AM

To determine the appropriate counting technique to use, we need to distinguish between a sequence of choices and an arrangement of items. We also need to determine whether order matters in the arrangements. See Figure 8. Keep in mind that one problem may require several counting rules. The [video](#) explains how to use the flowchart.

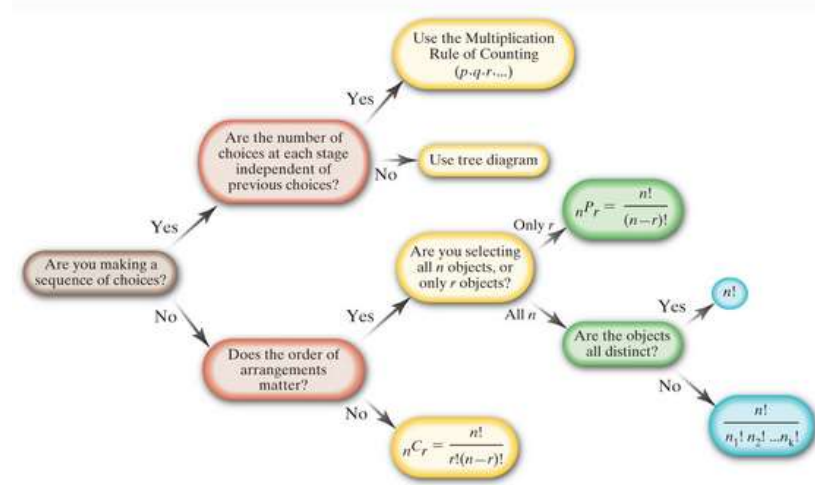


Figure 8

We first must decide whether we have a sequence of choices or an arrangement of items. For a sequence of choices, we use the Multiplication Rule of Counting if the number of choices at each stage is independent of previous choices. If the number of choices at each stage is not independent of previous choices, use a tree diagram. When determining the number of arrangements of items, we want to know whether the order of selection matters. If order matters, we also want to know whether we are arranging all the items available or only a subset of the items.

EXAMPLE 3 Counting: Which Technique Do I Use?

Problem

The Hazelwood city council consists of 5 men and 4 women. How many different subcommittees can be formed that consist of 3 men and 2 women?

Video Solution



Approach

Follow the flowchart in Figure 8 to determine which technique to use.

Solution

We need to find the number of subcommittees having 3 men and 2 women. So we consider a sequence of events: select the men and then select the women. Because the number of choices at each stage is independent of previous choices (the men chosen will not impact which women are chosen), we use the [Multiplication Rule of Counting](#) to obtain

$$N(\text{subcommittees}) = N(\text{ways to pick 3 men}) \cdot N(\text{ways to pick 2 women})$$

To select 3 men, we must consider the number of arrangements of 5 men taken 3 at a time. Because the order of selection does not matter, we use the [combination formula](#).

$$N(\text{ways to pick 3 men}) = {}_5C_3 = \frac{5!}{3! \cdot 2!} = 10$$

To select 2 women, we must consider the number of arrangements of 4 women taken 2 at a time. Because the order of selection does not matter, we use the [combination formula](#) again.

$$N(\text{ways to pick 2 women}) = {}_4C_2 = \frac{4!}{2! \cdot 2!} = 6$$

Combining our results, we obtain $N(\text{subcommittees}) = 10 \cdot 6 = 60$. There are 60 possible subcommittees that contain 3 men and 2 women.

EXAMPLE 4 Counting: Which Technique Do I Use?

Problem

The Daytona 500, the season opening NASCAR event, has 43 drivers in the race. In how many different ways could the top four finishers (first, second, third, and fourth place) occur?

Video Solution



Approach

Follow the flowchart in Figure 8 to determine which technique to use.

Solution

We need to find the number of ways to select the top four finishers. There are two different ways to solve this problem.

1. View this as a sequence of choices, where the first choice is the first-place driver, the second choice is the second-place driver, and so on. There are 43 ways to pick the first driver, 42 ways to pick the second, 41 ways to

1. View this as a sequence of choices, where the first choice is the first-place driver, the second choice is the second-place driver, and so on. There are 43 ways to pick the first driver, 42 ways to pick the second, 41 ways to pick the third, and 40 ways to pick the fourth. The number of choices at each stage is independent of previous choices, so we can use the [Multiplication Rule of Counting](#). The number of ways the top four finishers could occur is

$$N(\text{top four}) = 43 \cdot 42 \cdot 41 \cdot 40 = 2,961,840$$

2. We could also approach this problem as an arrangement of units. Because each race position is distinguishable, order matters. We are arranging the 43 drivers taken 4 at a time. Using our [permutation formula](#), we get

$$N(\text{top four}) = {}_{43}P_4 = \frac{43!}{(43 - 4)!} = \frac{43!}{39!} = 43 \cdot 42 \cdot 41 \cdot 40 = 2,961,840$$

Again, there are 2,961,840 different ways that the top four finishers could occur.

6.1 Discrete Random Variables

October 30, 2016 11:04 AM

6.1 Objective 1 Distinguish between Discrete and Continuous Random Variables

October 30, 2016 11:41 AM

Consider a **probability experiment** in which we flip a coin two times. The possible outcomes of the experiment are {HH, HT, TH, TT}. Rather than being interested in a particular outcome, we might be interested in the number of heads. If the outcome of a probability experiment is a numerical result, we say that the outcome is a *random variable*.

DEFINITION

A **random variable** is a numerical measure of the outcome of a probability experiment; so its value is determined by chance.

Random variables are typically denoted using capital letters such as X .

A probability experiment is any process with uncertain results that can be repeated.

Distinguishing between Discrete and Continuous Random Variables

So in our coin-flipping example, if the random variable X represents the number of heads in two flips of a coin, the possible values of X are $x = 0, 1, \text{ or } 2$. Notice that we follow the practice of using a capital letter, such as X , to identify the random variable and a lowercase letter, x , to list the possible values of the random variable, or the sample space of the experiment.

As another example, consider an experiment that measures the time between arrivals of cars at a drive-through. The random variable T describes the time between arrivals; so the sample space of the experiment is $t > 0$.

There are two types of random variables, *discrete* and *continuous*.

DEFINITION

A **discrete random variable** has either a finite or countable number of values. The values of a discrete random variable can be plotted on a number line with space between each point. See Figure 1(a).

A **continuous random variable** has infinitely many values. The values of a continuous random variable can be plotted on a line in an uninterrupted fashion. See Figure 1(b).

IN
OTHER
WORDS



Figure 1

The coin-flipping experiment involves discrete random variables, whereas the drive-through example involves continuous random variables.

Discrete Random Variable = Coin Flip, Fair Die Roll

EXAMPLE 1 Distinguishing between Discrete and Continuous Random Variables

(a) The number of A's earned in a section of statistics with 15 students enrolled is a discrete random variable because its value results from counting. If the random variable X represents the number of A's, then the possible values of X are $x = 0, 1, 2, \dots, 15$.

Video Solution



(b) The number of cars that travel through a McDonald's drive-through in the next hour is a discrete random variable because its value results from counting. If the random variable X represents the number of cars, the possible values of X are $x = 0, 1, 2, \dots$. That is, the number of cars can be any whole number and we do not impose an upper limit on the number of cars.

(c) The speed of the next car that passes a state trooper is a continuous random variable because speed is measured. If the random variable S represents the speed, the possible values of S are all positive real numbers; that is, $s > 0$. Even though a radar gun may report the speed of a car as 37 miles per hour, it is actually any number greater than or equal to 36.5 mph and less than 37.5 mph. That is, $36.5 \leq s < 37.5$.

6.1 Objective 2 Identify Discrete Probability Distributions

October 30, 2016 11:46 AM

Because the value of a random variable is determined by chance, we may assign probabilities to the possible values of the random variable.

DEFINITION

The **probability distribution** of a discrete random variable X provides the possible values of the random variable and their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

Remember, probabilities must follow certain rules.

Below are the rules for a **discrete probability distribution** using the notation just introduced.

Rules for a Discrete Probability Distribution

Let $P(x)$ denote the probability that the random variable X equals x ; then

$$1. \sum P(x) = 1$$

$$2. 0 \leq P(x) \leq 1$$



Table 1 is a discrete probability distribution because the sum of the probabilities equals **1** and each probability is between **0** and **1**, inclusive.

EXAMPLE 3 Identifying Discrete Probability Distributions

Problem

Which of the following is a **discrete probability distribution**?

Video Solution



The **probability distribution** of a discrete random variable X provides the possible values of the random variable and their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

Part A

Part B

Part C

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	0.01

Approach

In a discrete probability distribution, the sum of the probabilities must equal 1 and all probabilities must be between 0 and 1, inclusive.

Solution

This is not a discrete probability distribution because

$$\sum P(x) = 0.16 + 0.18 + 0.22 + 0.10 + 0.30 + 0.01 = 0.97 \neq 1$$

Part A

Part B

Part C

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	-0.01

Approach

In a discrete probability distribution, the sum of the probabilities must equal 1 and all probabilities must be between 0 and 1, inclusive.

Approach

In a discrete probability distribution, the sum of the probabilities must equal **1** and all probabilities must be between **0** and **1**, inclusive.

Solution

This is not a discrete probability distribution because $P(5) = -0.01$, which is less than **0**.

Part A

Part B

Part C

x	$P(x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.30
5	0.04

Approach

In a discrete probability distribution, the sum of the probabilities must equal **1**, and all probabilities must be between **0** and **1**, inclusive.

Solution

This is a discrete probability distribution because the sum of the probabilities equals **1**, and each probability is between **0** and **1**, inclusive.

6.1 Objective 3 Graph Discrete Probability Distributions

October 30, 2016 11:54 AM

In the graph of a discrete probability distribution, the horizontal axis is the value of the discrete random variable and the vertical axis is the corresponding probability of the discrete random variable. When graphing a discrete probability distribution, we want to emphasize that the data are discrete. Therefore, draw the graph of discrete probability distributions using vertical lines above each value of the random variable to a height that is the probability of the random variable.

TABLE 1

x	$P(x)$
0	0.01
1	0.10
2	0.38
3	0.51

EXAMPLE 4 Graph a Discrete Probability Distribution

Problem

Graph the discrete probability distribution given in [Table 1](#).

[Video Solution](#)



Approach

In the graph of a discrete probability distribution, the horizontal axis is the value of the discrete random variable and the vertical axis is the corresponding probability of the discrete random variable. Draw the graph using vertical lines above each value of the random variable to a height that is the probability of the random variable.

Solution

Figure 2 shows the graph of the distribution in [Table 1](#).



Figure 2

Distribution Shape of Discrete Probability Distributions

Graphs of discrete probability distributions help determine the shape of the distribution.

Recall that we describe distributions as *skewed left*, *skewed right*, or *symmetric*. The graph in [Figure 2](#) is skewed left.

6.1 Objective 4 Compute and Interpret the Mean of a Discrete Random Variable

October 30, 2016 11:56 AM

Mean of a Discrete Random Variable

The mean of a discrete random variable is given by the formula

$$\mu_X = \sum[x \cdot P(x)]$$

where x is the value of the random variable and $P(x)$ is the probability of observing the value x .

EXAMPLE 5 Computing the Mean of a Discrete Random Variable

Problem

Compute the mean of the discrete random variable given in Table 1.

Video Solution



Approach

Find the mean of a discrete random variable by multiplying each value of the random variable by its probability and adding these products.

Technology Step-By-Step



Solution

Refer to Table 2. The first two columns represent the discrete probability distribution. The third column represents $x \cdot P(x)$.

TABLE 2

x	$P(x)$	$x \cdot P(x)$
0	0.01	$0(0.01) = 0$
1	0.10	$1(0.10) = 0.1$
2	0.38	$2(0.38) = 0.76$
3	0.51	$3(0.51) = 1.53$

Now substitute into the formula for the mean of a discrete random variable.

$$\begin{aligned}\mu_X &= \sum[x \cdot P(x)] \\ &= 0(0.01) + 1(0.10) + 2(0.38) + 3(0.51) \\ &= 0 + 0.10 + 0.76 + 1.53 \\ &= 2.39 \\ &\approx 2.4 \quad \text{Round the mean to one more decimal place than the value of the random variable.}\end{aligned}$$

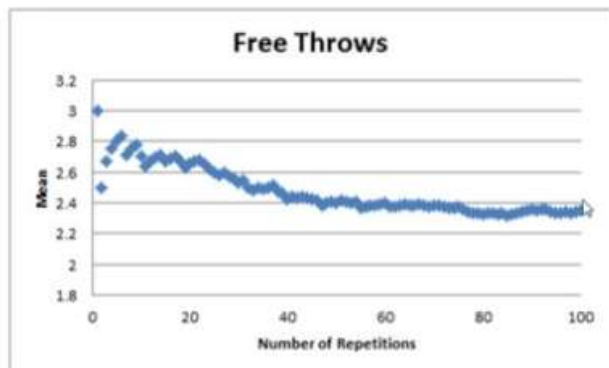
Example – Interpretation of the Mean of a Discrete Random Variable

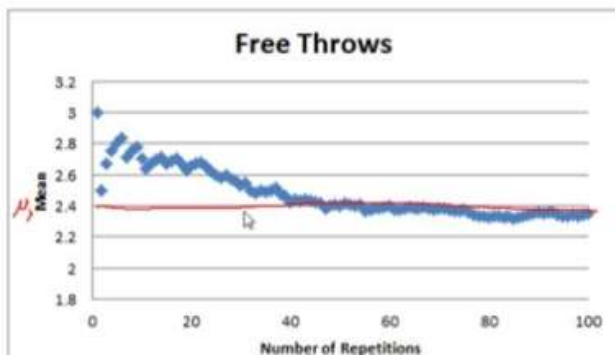
The basketball player from Example 2 is asked to shoot three free throws 100 times. Compute the mean number of free throws made.

3	2	3	3	3	3	1	2	3	2
2	3	3	1	2	2	2	2	2	3
3	3	2	2	3	2	3	2	2	2
3	3	2	3	2	3	3	2	3	1
3	2	2	2	2	0	2	3	1	2
3	3	2	3	2	3	2	1	3	2
2	3	3	3	1	3	3	1	3	3
3	2	2	1	3	2	2	2	3	2
3	2	2	2	3	3	2	2	3	3
2	3	2	1	2	3	3	2	3	3

Solution

$$\bar{x} = \frac{3 + 2 + 3 + \cdots + 3}{100} = 2.35$$





6.1 Objective 5 Interpret the Mean of a Discrete Random Variable as an Expected Value

October 30, 2016 12:04 PM

Because the mean of a random variable represents what we would expect to happen in the long run, it is also called the **expected value**, $E(X)$. The interpretation of the expected value is the same as the interpretation of the mean of a discrete random variable.

EXAMPLE 7 Computing the Expected Value of a Discrete Random Variable

Problem

A term life insurance policy will pay a beneficiary a certain sum of money upon the death of the policy holder. These policies have premiums that must be paid annually. Suppose a life insurance company sells a \$250,000 one-year term life insurance policy to a 49-year-old female for \$530. According to the *National Vital Statistics Report*, Vol. 47, No. 28, the probability that the female will survive the year is 0.99791. Compute the expected value of this policy to the insurance company.

Video Solution



Approach

The experiment has two possible outcomes: survival or death. Let the random variable X represent the *payout* (money lost or gained), depending on survival or death of the insured. Assign probabilities to each payout and substitute these values into $\mu_X = \sum [x \cdot P(x)]$.

Solution

Step 1 Because $P(\text{survives}) = 0.99791$, $P(\text{dies}) = 0.00209$. If the client survives the year, the insurance company makes \$530, or $x = +530$. If the client dies during the year, the insurance company must pay \$250,000 to the client's beneficiary, but still keeps the \$530 premium; so $x = \$530 - \$250,000 = -\$249,470$. The value is negative because it is money paid by the insurance company. The probability distribution is listed in Table 3.

TABLE 3

x	$P(x)$
\$530 (survives)	0.99791
-\$249,470 (dies)	0.00209

Step 2 The expected value (from the point of view of the insurance company) of the policy is

$$\begin{aligned} E(X) &= \mu_X \\ &= \sum x \cdot P(x) \\ &= \$530(0.99791) + (-\$249,470)(0.00209) \\ &= \$7.50 \end{aligned}$$

Interpretation

The company expects to make \$7.50 for each 49-year-old female client it insures. The \$7.50 profit of the insurance company is a long-term result. It does not make \$7.50 on each 49-year-old female it insures; rather, the average profit per 49-year-old female insured is \$7.50. Because this is a long-term result, the insurance "idea" will not work with only a few insured.

6.1 Objective 6 Compute the Standard Deviation of a Discrete Random Variable

October 30, 2016 12:18 PM

Below is the formula for computing the standard deviation of a discrete random variable.

Standard Deviation of a Discrete Random Variable

The standard deviation of a discrete random variable X is given by

$$\sigma_X = \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]}$$

where x is the value of the random variable, μ_X is the mean of the random variable, and $P(x)$ is the probability of observing x .

EXAMPLE 8 Computing the Standard Deviation of a Discrete Random Variable

Problem

Compute the standard deviation of the discrete random variable given in [Table 1](#).

Video Solution



Approach

Use the formula $\sigma_X = \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]}$ where $\mu_X = 2.39$

Technology Step-By-Step



NOTE

Use the unrounded mean to avoid round-off error.

Solution

Refer to Table 4. Columns 1 and 2 represent the discrete probability distribution. Column 3 represents $(x - \mu_X)^2 \cdot P(x)$. Find the sum of the entries in Column 3.

TABLE 4

x	$P(x)$	$(x - \mu_X)^2 \cdot P(x)$
0	0.01	$(0 - 2.39)^2 \cdot 0.01 = 0.057121$
1	0.10	$(1 - 2.39)^2 \cdot 0.10 = 0.19321$
2	0.38	$(2 - 2.39)^2 \cdot 0.38 = 0.057798$
3	0.51	$(3 - 2.39)^2 \cdot 0.51 = 0.189771$
		$\sum (x - \mu_X)^2 \cdot P(x) = 0.4979$

The standard deviation of the discrete random variable X is

$$\begin{aligned}\sigma_X &= \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]} \\ &= \sqrt{0.4979} \\ &\approx 0.7\end{aligned}$$

Finding the Variance of a Discrete Random Variable

The variance of the discrete random variable, σ_X^2 , is the value under the square root in the computation of the standard deviation. The variance of the discrete random variable in Table 1 is

$$\begin{aligned}\sigma_X^2 &= 0.4979 \\ &\approx 0.5\end{aligned}$$

6.2 The Binomial Probability Distribution

October 31, 2016 08:35 AM

6.2 Objective 1 Determine Whether a Probability Experiment is a Binomial Experiment

October 31, 2016 08:42 AM

We stated that [probability distributions](#) could be presented using tables, graphs, or mathematical formulas. We now introduce a specific type of discrete probability distribution that can be presented using a formula, the *binomial probability distribution*.

DEFINITION

The **binomial probability distribution** is a discrete probability distribution that describes probabilities for experiments in which there are two mutually exclusive (disjoint) outcomes. These two outcomes are generally referred to as *success* (such as making a free throw) and *failure* (such as missing a free throw). Experiments in which only two outcomes are possible are referred to as *binomial experiments*, provided that certain criteria are met.



The **probability distribution** of a discrete random variable X provides the possible values of the random variable and their corresponding probabilities. These distributions must satisfy the rules of a probability model:

1. The probability of any event (and random variable) must be greater than or equal to 0 and less than or equal to 1 .
2. The sum of the probabilities of all outcomes (all possible values of the random variable) must equal 1 .

EXAMPLE 1 Identifying Binomial Experiments

Problem

Determine which of the following probability experiments qualify as binomial experiments. For those that are binomial experiments, identify the number of trials, probability of success, probability of failure, and possible values of the random variable X .

Video Solution



An experiment in which a basketball player who historically makes 80% of his free throws is asked to shoot three free throws and the number of free throws made is recorded.

Approach

Determine whether the four conditions for a binomial experiment are satisfied.

1. The experiment is performed a fixed number of times.
2. The trials are independent.
3. There are only two possible outcomes of the experiment.
4. The probability of success for each trial is constant.

Solution

This is a binomial experiment because

1. There are $n = 3$ trials.
2. The trials are independent.
3. There are two possible outcomes: make or miss.
4. The probability of success (make) is $p = 0.8$, and the probability of failure (miss) is $1 - 0.8 = 0.2$. The probabilities are the same for each trial.

The random variable X is the number of free throws made with $x = 0, 1, 2, \text{ or } 3$.

According to a recent Harris Poll, 28% of Americans state that chocolate is their favorite flavor of ice cream. Suppose a simple random sample of size 10 is obtained and the number of Americans who choose chocolate as their favorite ice cream flavor is recorded.

Approach

Determine whether the four conditions for a binomial experiment are satisfied.

1. The experiment is performed a fixed number of times.
2. The trials are independent.
3. There are only two possible outcomes of the experiment.
4. The probability of success for each trial is constant.

Solution

This is a binomial experiment because

1. There are $n = 10$ trials (the ten randomly selected people).
2. The trials are **independent**.
3. There are two possible outcomes: finding an American who chooses chocolate as his or her favorite ice cream or not.
4. The probability of success is $p = 0.28$, and the probability of failure is $1 - 0.28 = 0.72$.

cream or not.

4. The probability of success is $p = 0.28$, and the probability of failure is $1 - 0.28 = 0.72$.

The random variable X is the number of people who choose chocolate as their favorite ice cream with $x = 0, 1, 2, 3, \dots, 10$.

Part A

Part B

Part C

A probability experiment in which three cards are drawn from a deck without replacement and the number of aces is recorded.

Approach

Determine whether the four conditions for a binomial experiment are satisfied.

1. The experiment is performed a fixed number of times.
2. The trials are independent.
3. There are only two possible outcomes of the experiment.
4. The probability of success for each trial is constant.

Solution

This is not a binomial experiment because the trials are not independent. The probability of an ace on the first trial is $\frac{4}{52}$. Because we are sampling without replacement, if an ace is selected on the first trial, the probability of an ace on the second trial is $\frac{3}{51}$. If an ace is not selected on the first trial, the probability of an ace on the second trial is $\frac{4}{51}$.

Conditions for a Binomial Experiment

1. The experiment is performed a fixed number of times.
2. The trials are independent.
3. There are only two possible outcomes of the experiment.
4. The probability of success for each trial is constant.

6.2 Objective 2 Compute Probabilities of Binomial Events

October 31, 2016 08:46 AM

In the previous video, there were $n = 4$ trials of a binomial experiment with a probability of success $p = 0.79$. The probability of obtaining $x = 3$ successes is found as follows:

$$P(3) = 4(0.79)^3(0.21)^1$$

4 is the number of ways to obtain three successes in four trials of the experiment. Here, it is ${}_4C_3$.

0.79 is the probability of success. The exponent 3 is the number of successes.

0.21 is the probability of failure. The exponent 1 is the number of failures.

Note* The video did not work

We summarize the *binomial probability distribution function (pdf)* below:

Binomial Probability Distribution Function

The probability of obtaining x successes in n independent trials of a **binomial experiment** is given by

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

where p is the probability of success.

When reading probability problems, pay special attention to key phrases that translate into mathematical symbols. Table 5 lists various phrases and their corresponding mathematical equivalent.

TABLE 5

Phrase	Math Symbol
<i>at least or no less than or greater than or equal to</i>	\geq
<i>more than or greater than</i>	$>$
<i>fewer than or less than</i>	$<$
<i>no more than or at most or less than or equal to</i>	\leq
<i>exactly or equals or is</i>	$=$

EXAMPLE 2 Using the Binomial Probability Distribution Function

Problem

According to CTIA, 25% of all U.S. households are wireless-only households (no landline).

Video Solution

Formula Tables



Technology Step-By-Step



Part A

Part B

Part C

Part D

What is the probability of obtaining exactly five wireless-only households based on a random sample of twenty households?

Approach

This is a binomial experiment with $n = 20$ independent trials. We define a success as selecting a household that is wireless-only. The probability of success, p , is 0.25. The possible values of the random variable X are $x = 0, 1, 2, \dots, 20$. Now decide on a method for obtaining the probabilities (formula, tables, or technology). Watch the video solution for the method you choose.

Result and Interpretation

The probability of getting exactly 5 households out of 20 that are wireless-only is 0.2023. In 100 trials of this experiment (that is, if we surveyed 20 households 100 different times), we would expect about 20 trials to result in 5 households that are wireless-only.

What is the probability of obtaining fewer than three wireless-only households based on a random sample of twenty households?

Approach

This is a binomial experiment with $n = 20$ independent trials. We define a success as selecting a household that is wireless-only. The probability of success, p , is 0.25 . The possible values of the random variable X are $x = 0, 1, 2, \dots, 20$. The phrase *fewer than* means "less than." The values of the random variable X less than 3 are $x = 0, 1$, or 2 . Now decide on a method for obtaining the probabilities (formula, tables, or technology). Watch the video solution for the method you choose.

Result and Interpretation

There is a 0.0912 probability that in a random sample of 20 households, fewer than 3 will be a wireless-only household. In 100 trials of this experiment, we would expect about 9 trials to result in fewer than 3 wireless-only households.

What is the probability of obtaining at least three wireless-only households based on a random sample of twenty households?

Approach

This is a binomial experiment with $n = 20$ independent trials. We define a success as selecting a household that is wireless-only. The probability of success, p , is 0.25 . The possible values of the random variable X are $x = 0, 1, 2, \dots, 20$. The complement of "at least 3 " is "less than 3 ". Because $P(X < 3)$ was found in part (B), we can use the Complement Rule to find $P(X \geq 3)$. Now, decide on a method for obtaining the probabilities (formula, tables, or technology). Watch the video solution for the method you choose.

Result and Interpretation

There is a 0.9088 probability that in a random sample of 20 households, at least 3 will be a wireless-only household. In 100 trials of this experiment, we would expect about 91 trials to result in at least 3 being wireless-only households.

What is the probability of obtaining between five and seven, inclusive, wireless-only households based on a random sample of twenty households?

Approach

This is a binomial experiment with $n = 20$ independent trials. We define a success as selecting a household that is wireless-only. The probability of success, p , is 0.25 . The possible values of the random variable X are $x = 0, 1, 2, \dots, 20$. The word *inclusive* means "including," so we want to determine the probability that five, six, or seven households are wireless-only. Now, decide on a method for obtaining the probabilities (formula, tables, or technology). Watch the video solution for the method you choose.

Result and Interpretation

The probability that the number of wireless-only households is between five and seven, inclusive, is 0.4833 . In 100 trials of this experiment, we would expect about 48 trials to result in 5 to 7 households that are wireless-only.

6.2 Objective 3 Compute the Mean and Standard Deviation of a Binomial Random Value

October 31, 2016 09:39 AM

We discussed finding the **mean** (or **expected value**) and **standard deviation** of a discrete random variable. These formulas can be used to find the mean and standard deviation of a binomial random variable, but a simpler method exists.

Mean (or Expected Value) and Standard Deviation of a Binomial Random Variable

A binomial experiment with n independent trials and probability of success p has mean, μ_X , and standard deviation, σ_X , given by the formulas

$$\mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)}$$

IN
OTHER
WORDS



EXAMPLE 3 Finding the Mean and Standard Deviation of a Binomial Random Variable

Problem

According to CTIA, 25% of all U.S. households are wireless-only households. In a simple random sample of 300 households, determine the mean and standard deviation number of wireless-only households.

Video Solution



Approach

This is a binomial experiment with $n = 300$ and $p = 0.25$. Use $\mu_X = np$ to find the mean and $\sigma_X = \sqrt{np(1-p)}$ to find the standard deviation.

Solution

$$\mu_X = np = 300(0.25) = 75$$

and

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{300(0.25)(1-0.25)} = \sqrt{56.25} = 7.5$$

Interpretation

We expect that in a random sample of 300 households, 75 will be wireless-only.

Shape of the Graph of a Binomial Probability Distribution for Various Values of n

For a fixed p , as the number of trials n in a binomial experiment increases, the probability distribution of the random variable X becomes bell-shaped.

As a rule of thumb, if $np(1 - p) \geq 10$, the probability distribution will be approximately bell-shaped.

This result allows us to use the **Empirical Rule** to identify unusual observations in a binomial experiment. Recall that the Empirical Rule states that in a bell-shaped distribution, about 95% of all observations lie within 2 standard deviations of the mean. That is, about 95% of the observations lie between $\mu - 2\sigma$ and $\mu + 2\sigma$. Any observation that lies outside this interval may be considered unusual because the observation occurs less than 5% of the time.

EXAMPLE 5 Using the Mean, Standard Deviation, and Empirical Rule to Check for Unusual Results in a Binomial Experiment

Problem

According to CTIA, 25% of all U.S. households are wireless-only. In a simple random sample of 300 households, 92 were wireless-only. Is this result unusual?

Video Solution



Approach

Because $np(1 - p) = 300(0.25)(1 - 0.25) = 56.25 \geq 10$, the binomial probability distribution is approximately bell-shaped. Therefore, we can use the Empirical Rule: If the observation is less than $\mu - 2\sigma$ or greater than $\mu + 2\sigma$, it is unusual.

Solution

We have $\mu_X = 300(0.25) = 75$ and $\sigma_X = \sqrt{np(1 - p)} = \sqrt{300(0.25)(1 - 0.25)} = 7.5$. Now

$$\begin{aligned}\mu_X - 2\sigma_X &= 75 - 2(7.5) & \mu_X + 2\sigma_X &= 75 + 2(7.5) \\ &= 75 - 15 & &= 75 + 15 \\ &= 60 & \text{and} &= 90\end{aligned}$$

Any value less than 60 or greater than 90 is unusual; therefore, 92 is an unusual result. We should try to identify the reason for its value. Perhaps the percentage of households that are wireless-only has increased.

6.2 Objective 4 Graph a Binomial Probability Distribution

October 31, 2016 09:45 AM

To graph a binomial probability distribution, first find the probabilities for each possible value of the random variable. Then follow the same approach as was used to graph [discrete probability distributions](#).

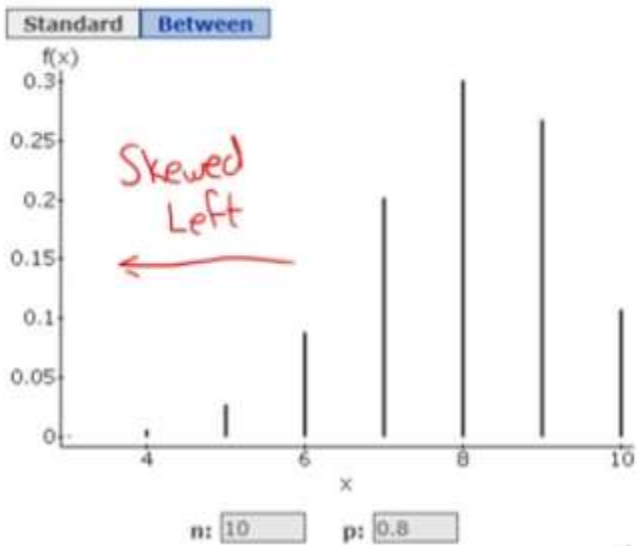
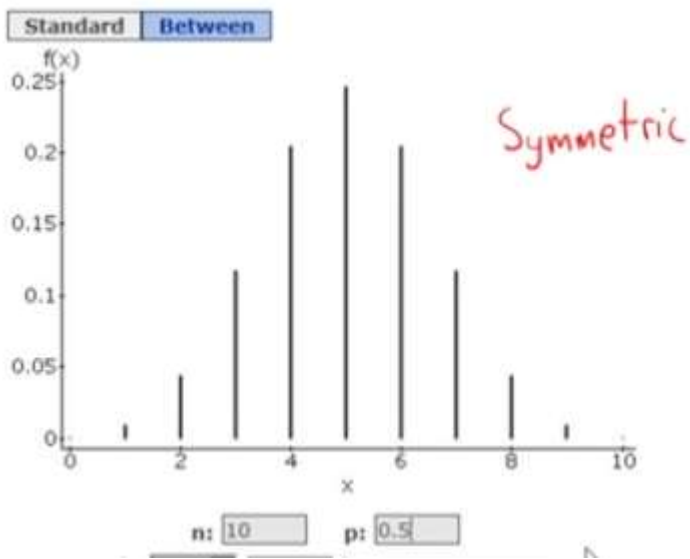
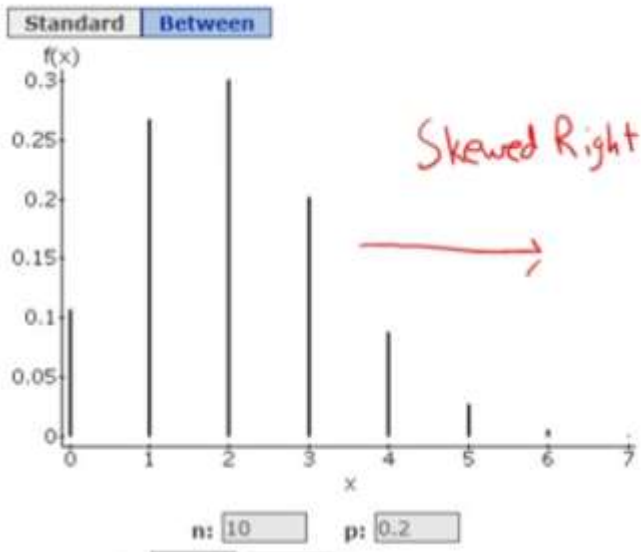
In the graph of a discrete probability distribution, the horizontal axis corresponds to the value of the random variable, and the vertical axis represents the probability of each value of the random variable.



Example – Graph a Binomial Probability Distribution

- (a)** Graph a binomial probability distribution with $n = 10$ and $p = 0.2$. Comment on the shape of the distribution.
- (b)** Graph a binomial probability distribution with $n = 10$ and $p = 0.5$. Comment on the shape of the distribution.
- (c)** Graph a binomial probability distribution with $n = 10$ and $p = 0.8$. Comment on the shape of the distribution.

Statcrunch: Stat>Calculators>Binomial



Based on the graphs drawn in the previous example, we have the following result.

Shape of Binomial Probability Distribution for Various Values of p

The binomial probability distribution is skewed right if $p < 0.5$, is symmetric and approximately bell-shaped if $p = 0.5$, and is skewed left if $p > 0.5$.

6.3 The Poisson Probability Distribution

November 3, 2016 08:45 AM

6.3 Objective 1 Determine Whether a Probability Experiment Follows as a Poisson Process

November 3, 2016 08:46 AM

Another discrete probability model is the *Poisson probability distribution*, named after Siméon Denis Poisson. This probability distribution can be used to compute probabilities of experiments in which the random variable X counts the number of occurrences (successes) of a particular event within a specified interval (usually time or space).

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 1 Illustrating a Poisson Process

A McDonald's® manager knows from experience that cars arrive at the drive-through at an average rate of two cars per minute between the hours of 12:00 noon and 1:00 PM. The random variable X , the number of cars that arrive between 12:20 and 12:40, follows a Poisson process.

DEFINITION

A random variable X , the number of successes in a fixed interval, follows a **Poisson process** provided the following conditions are met.

1. The probability of two or more successes in any sufficiently small subinterval is 0. For example, the fixed interval might be any time between 0 and 5 minutes. A subinterval could be any time between 1 and 2 seconds.
2. The probability of success is the same for any two intervals of equal length.
3. The number of successes in any interval is independent of the number of successes in any other interval provided the intervals are not overlapping.

Verifying That the Conditions for a Poisson Process Are Met

In the McDonald's example, if we divide the time interval into a sufficiently small length (for example, 1 second), it is impossible for more than one car to arrive. This satisfies part 1 of the definition.

Part 2 is satisfied because the cars arrive at an average rate of two cars per minute over the one-hour interval.

Part 3 is satisfied because the number of cars that arrive in any one-minute interval (for example, between 12:23 PM and 12:24 PM) is independent of the number of cars that arrive in any other one-minute interval (for example, between 12:35 PM and 12:36 PM).

6.3 Objective 2 Compute Probabilities of a Poisson Random Variable

November 3, 2016 08:49 AM

If the random variable X follows a Poisson process, we can use the following probability rule to compute Poisson probabilities.

Poisson Probability Distribution Function

If X is the number of successes in an interval of fixed length t , then the probability of obtaining x successes in the interval is

$$P(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad x = 0, 1, 2, 3, \dots$$

where λ (the Greek letter lambda) represents the average number of occurrences of the event in some interval of length 1 and $e \approx 2.71828$.

To clarify the roles of λ and t , revisit the [McDonald's](#) example. Here, $\lambda = 2$ cars per minute, while $t = 20$ minutes (the length of time between 12:20 P.M. and 12:40 P.M.).

A McDonald's® manager knows from prior experience that cars arrive at the drive-through at an average rate of two cars per minute between the hours of 12:00 noon and 1:00 P.M. The random variable X , the number of cars that arrive between 12:20 and 12:40, follows a Poisson process.

EXAMPLE 2 Computing Probabilities of a Poisson Process

Problem

A McDonald's manager knows that cars arrive at the drive-through at the average rate of two cars per minute between the hours of 12 noon and 1:00 P.M. Find the following probabilities.

Video Solution



Technology Step-By-Step



Find the probability that exactly six cars arrive between 12 noon and 12:05 PM.

Approach

The manager needs a method to determine the probabilities. The cars arrive at a rate of two per minute over the time interval between 12 noon and 1:00 P.M. The random variable X follows a Poisson process, where $x = 0, 1, 2, \dots$. The **Poisson probability distribution function** requires a value for λ and t . Because the cars arrive at a rate of two per minute, $\lambda = 2$. The interval of time we are interested in is five minutes, so $t = 5$. Compute the probability either by hand or using technology. Watch the video solution.

Solution

The probability that exactly six cars arrive between 12 noon and 12:05 P.M. is

$$P(6) = \frac{[2(5)]^6}{6!} e^{-2(5)} = \frac{1,000,000}{720} e^{-10} = 0.0631$$

Interpretation

On about 6 of every 100 days, exactly 6 cars will arrive between 12:00 noon and 12:05 P.M.

Find the probability that fewer than six cars arrive between 12 noon and 12:05 PM.

Approach

The cars arrive at a rate of two per minute over the time interval between 12 noon and 1:00 P.M. The random variable X follows a Poisson process, where $x = 0, 1, 2, \dots$. The **Poisson probability distribution function** requires a value for λ and t . Because the cars arrive at a rate of two per minute, $\lambda = 2$. The interval of time we are interested in is five minutes, so $t = 5$.

Solution

The probability that fewer than six cars arrive between 12:00 noon and 12:05 P.M. is

$$\begin{aligned}
 P(X < 6) &= P(X \leq 5) \\
 &= P(0) + P(1) + P(2) + P(3) + P(4) + P(5) \\
 &= \frac{[2(5)]^0}{0!} e^{-2(5)} + \frac{[2(5)]^1}{1!} e^{-2(5)} + \frac{[2(5)]^2}{2!} e^{-2(5)} \\
 &\quad + \frac{[2(5)]^3}{3!} e^{-2(5)} + \frac{[2(5)]^4}{4!} e^{-2(5)} + \frac{[2(5)]^5}{5!} e^{-2(5)} \\
 &= \frac{1}{1} e^{-10} + \frac{10}{1} e^{-10} + \frac{100}{2} e^{-10} + \frac{1000}{6} e^{-10} + \frac{10,000}{24} e^{-10} + \frac{100,000}{120} e^{-10} \\
 &= 0.0671
 \end{aligned}$$

Interpretation

On about 7 of every 100 days, fewer than six cars will arrive between 12:00 noon and 12:05 P.M.

Find the probability that at least six cars arrive between 12 noon and 12:05 PM.

Approach

The cars arrive at a rate of two per minute over the time interval between 12 noon and 1:00 P.M. The random variable X follows a Poisson process, where $x = 0, 1, 2, \dots$. The [Poisson probability distribution function](#) requires a value for λ and t . Because the cars arrive at a rate of two per minute, $\lambda = 2$. The interval of time we are interested in is five minutes, so $t = 5$.

At-least and *more-than* probabilities for a Poisson process must be found using the [Complement Rule](#) because the random variable X can be any integer greater than or equal to 0.

Solution

The probability that at least six cars arrive between 12 noon and 12:05 P.M. is the complement of the probability that fewer than six cars arrive during that time. That is,

$$\begin{aligned}P(X \geq 6) &= 1 - P(X < 6) \\ &= 1 - 0.0671 \\ &= 0.9329\end{aligned}$$

Interpretation

On about **93** of every **100** days, at least six cars will arrive between 12:00 noon and 12:05 p.m.

6.3 Objective 3 Find the Mean and Standard Deviation of a Poisson Random Variable

November 3, 2016 09:23 AM

If two cars per minute arrive at McDonald's between 12 noon and 1:00 P.M., how many cars would you expect to arrive between 12 noon and 12:05 P.M.? Considering that two cars arrive every minute (on average) and we are observing the arrival of cars for five minutes, it seems reasonable to expect $2(5) = 10$ cars to arrive. Because the expected value of a random variable is the mean of the random variable, it is reasonable that $\mu_X = \lambda t$ for an interval of length t .

Mean and Standard Deviation of a Poisson Random Variable

A random variable X that follows a Poisson process with parameter λ has mean (or expected value) and standard deviation given by the formulas

$$\mu_X = \lambda t \quad \text{and} \quad \sigma_X = \sqrt{\lambda t} = \sqrt{\mu_X}$$

where t is the length of the interval.

Because $\mu_X = \lambda t$, we restate the [Poisson probability distribution function](#) in terms of its mean.

Poisson Probability Distribution Function

If X is the number of successes in an interval of fixed length and X follows a Poisson process with mean $\mu_X = \lambda t$, then the probability distribution function for X is

$$P(x) = \frac{\mu^x}{x!} e^{-\mu} \quad x = 0, 1, 2, 3, \dots$$

where $e \approx 2.71828$.

Formula for the Poisson Probability Distribution Function

$$P(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad x = 0, 1, 2, 3, \dots$$

EXAMPLE 3 Beetles and the Poisson Distribution

Problem

A biologist performs an experiment in which 2000 Asian beetles are allowed to roam in an enclosed area of 1000 square feet. The area is divided into 200 subsections of 5 square feet each.

Video Solution



Technology Step-By-Step



Part A

Part B

Part C

Part D

If the beetles spread themselves evenly throughout the enclosed area, how many beetles would you expect in each subsection?

Approach

If the beetles spread themselves evenly throughout the enclosed region, we can model the distribution of the beetles using Poisson probabilities.

Solution

If the beetles spread themselves evenly throughout the enclosed area, we expect

$$\begin{aligned}\mu_X &= \frac{2000 \text{ beetles}}{200 \text{ subsections}} \\ &= 10 \text{ beetles per subsection}\end{aligned}$$

EXAMPLE 3 Beetles and the Poisson Distribution

Problem

A biologist performs an experiment in which 2000 Asian beetles are allowed to roam in an enclosed area of 1000 square feet. The area is divided into 200 subsections of 5 square feet each.

Video Solution



Technology Step-By-Step



Part A

Part B

Part C

Part D

What is the standard deviation of X , the number of beetles in a particular subsection?

Approach

$$\sigma_X = \sqrt{\lambda t} = \sqrt{\mu_X}$$

Solution

$$\begin{aligned}\sigma_X &= \sqrt{\mu_X} \\ &= \sqrt{10} \\ &\approx 3.2\end{aligned}$$

Part A

Part B

Part C

Part D

What is the probability of finding exactly eight beetles in a particular subsection?

Approach

Use the expected value $\mu_X = 10$ in the [Poisson probability distribution function](#) to compute the probability of finding exactly eight beetles in a subsection.

Solution

$$\begin{aligned}P(8) &= \frac{10^8}{8!} e^{-10} & P(x) &= \frac{\mu^x}{x!} e^{-\mu}, \mu = 10, x = 8 \\ &= 0.1126\end{aligned}$$

In 100 trials of this experiment, we expect to find eight beetles in a particular subsection about 11 times.

Would it be unusual to find more than 16 beetles in a particular subsection?

Approach

Compute $P(X > 16)$. An event is unusual if the probability of the event is less than 0.05.

Solution

$$\begin{aligned}P(X > 16) &= 1 - P(X \leq 16) \\ &= 1 - 0.9730 \quad \text{Use Technology} \\ &= 0.0270\end{aligned}$$

According to the Poisson probability model, there will be more than 16 beetles in a subsection about 3 times in 100. To observe more than 16 beetles in a subsection is rather unusual.

6.3.11

November 3, 2016 09:17 AM

The number of hits to a Web site follows a Poisson process. Hits occur at the rate of 3.6 per minute between 7:00 P.M. and 9:00 P.M. Given below are three scenarios for the number of hits to the Web site. Compute the probability of each scenario between 8:26 P.M. and 8:31 P.M.

- (a) exactly eight.
- (b) fewer than eight.
- (c) at least eight.

First, identify the values for λ and t .

$$\lambda = 3.6$$

$$t = 5$$

Substitute the values for λ , t and X into the equation.

$$P(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad P(8) = \frac{(3.6 \cdot 5)^8}{8!} e^{-3.6 \cdot 5}$$

7.1 Properties of the Normal Distribution

November 7, 2016 08:01 AM

7.1 Objective 1 Use the Uniform Probability Distribution

November 7, 2016 08:09 AM

EXAMPLE 1 The Uniform Distribution

Assume that United Parcel Service is supposed to deliver a package to your front door and the arrival time is somewhere between 10 AM and 11 AM. Let the random variable X represent the time from 10 AM when the delivery is supposed to take place.

Video Solution



The delivery could be at 10 AM ($x = 0$) or at 11 AM ($x = 60$), with all one-minute intervals of time between $x = 0$ and $x = 60$ equally likely. That is to say, your package is just as likely to arrive between 10:15 and 10:16 as it is to arrive between 10:40 and 10:41.

The random variable X can be any value in the interval from 0 to 60, that is, $0 \leq X \leq 60$. Because any two intervals of equal length between 0 and 60, inclusive, are equally likely, the random variable X is said to follow a **uniform probability distribution**.

When we compute probabilities for [discrete random variables](#), we usually substitute the value of the random variable into a formula.

Things are not as easy for [continuous random variables](#). Because an infinite number of outcomes are possible for continuous random variables, the probability of observing one *particular* value is zero. In the UPS example, the probability that the package arrives exactly 12.9438823 minutes after 10 AM is zero. This result is based on [classical probability](#): there is one way to observe 12.9438823, and there are an infinite number of possible values between 0 and 60. To resolve this problem, we compute probabilities of continuous random variables over an *interval* of values. For example, we might compute the probability that your package arrives between $x=10$ minutes and $x=15$ minutes after 10 AM. To find probabilities for continuous random variables, we use *probability density functions*.

A **discrete random variable** has either a finite or countable number of values. The values of a discrete random variable can be plotted on a number line with space between each point.

A **continuous random variable** has infinitely many values. The values of a continuous random variable can be plotted on a line in an uninterrupted fashion.

Classical Probability

If an experiment has n equally likely outcomes and if the number of ways an event E can occur is m , then the probability of E , $P(E)$, is

$$\begin{aligned} P(E) &= \frac{\text{number of ways } E \text{ can occur}}{\text{number of possible outcomes}} \\ &= \frac{m}{n} \end{aligned}$$

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

7.1 Objective 2 Graph a Normal Curve

November 7, 2016 08:20 AM

When describing a uniform random variable using a probability distribution, a rectangle is used to find probabilities of observing an interval of numbers (such as 10 to 20 minutes after 10 AM). However, not all continuous random variables follow a uniform distribution. For example, continuous random variables such as IQ scores and birth weights of babies have distributions that are symmetric and bell-shaped. Consider the histograms in Figure 1, which represent the IQ scores of 10,000 randomly selected adults. Notice that as the class width of the histogram decreases, the histogram becomes closely approximated by the smooth red curve. For this reason, we can use the curve to *model* the probability distribution of this continuous random variable.

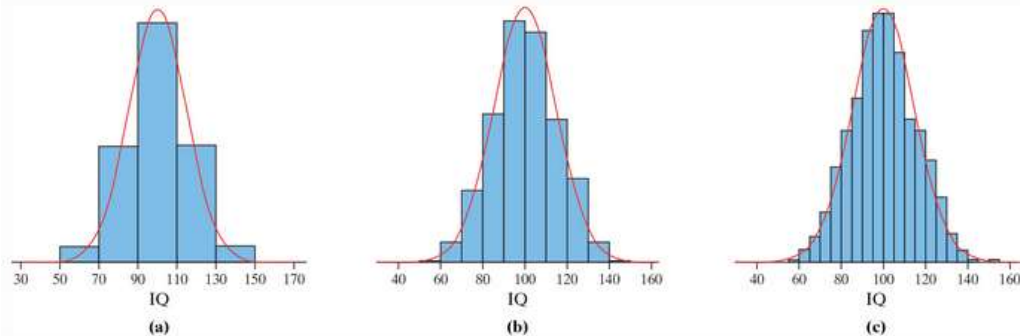


Figure 1

In mathematics, a **model** is an equation, a table, or a graph used to describe reality. The red curve in [Figure 1](#) is a model called the **normal curve**, which is used to describe continuous random variables that are *normally distributed*.

Definition

A continuous random variable is **normally distributed**, or has a **normal probability distribution**, if its relative frequency histogram has the shape of a normal curve.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

The Normal Curve

Figure 2 shows a normal curve, demonstrating the roles that the mean μ and standard deviation σ play in the curve being drawn. The vertical scale on the graph, which indicates **density**, is purposely omitted and will not play a role in any of the computations using this curve.

The mode represents the "high point" of the graph of any distribution. The median represents the point where 50% of the area under the distribution is to the left and 50% is to the right. The mean represents the balancing point of the graph of the distribution.

For symmetric distributions with a single peak, such as the normal distribution, the **mean = median = mode**. Because of this, the mean, μ , is the high point of the graph of the distribution.

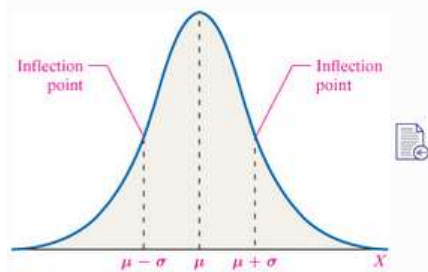


Figure 2

The points at $x = \mu - \sigma$ and $x = \mu + \sigma$ are the **inflection points** on the normal curve, the points on the curve where the curvature of the graph changes. To the left of $x = \mu - \sigma$ and to the right of $x = \mu + \sigma$, the curve is drawn upward:



Between $x = \mu - \sigma$ and $x = \mu + \sigma$, the curve is drawn downward:

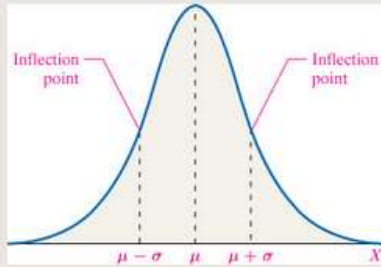


7.1 Objective 3 State the Properties of the Normal Curve

November 7, 2016 08:25 AM

Properties of the Normal Curve

1. The normal curve is symmetric about its mean, μ .



2. Because mean = median = mode, the normal curve has a single peak and the highest point occurs at $x = \mu$.

3. The normal curve has inflection points at $\mu - \sigma$ and $\mu + \sigma$.

4. The area under the normal curve is 1.

5. The area under the normal curve to the right of μ equals the area under the normal curve to the left of μ , which equals $\frac{1}{2}$.

6. As x increases without bound (gets larger and larger), the graph approaches, but never reaches, the horizontal axis. As x decreases without bound (gets more and more negative), the graph approaches, but never reaches, the horizontal axis.

7. The Empirical Rule: Approximately 68% of the area under the normal curve is between $x = \mu - \sigma$ and $x = \mu + \sigma$,

approximately 95% of the area is between $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$, and approximately 99.7% of the area is between $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$.

See Figure 3.

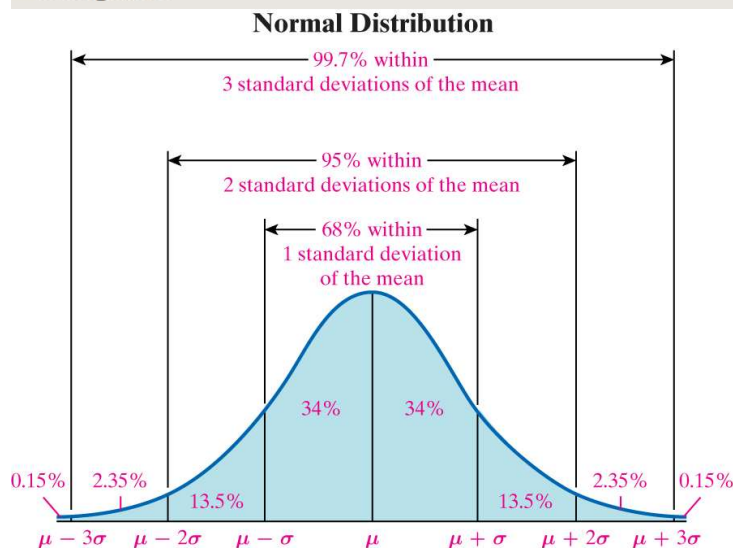


Figure 3: Normal Distribution, a.k.a Empirical Rule

7.1 Objective 4 Explain the Role of Area in the Normal Density Function

November 7, 2016 08:26 AM

Area under a Normal Curve

Suppose that a random variable X is normally distributed with mean μ and standard deviation σ . The area under the normal curve for any interval of values of the random variable X represents either

- the *proportion* of the population with the characteristic described by the interval of values
- or
- the *probability* that a randomly selected individual from the population will have the characteristic described by the interval of values.

EXAMPLE 2 Interpreting the Area Under a Normal Curve

Problem

The serum total cholesterol for males 20 to 29 years old is approximately normally distributed with mean $\mu = 180$ and standard deviation $\sigma = 36.2$, based on data obtained from the National Health and Nutrition Examination Survey.

[Video Solution](#)



Part A

Part B

Part C

Draw a normal curve with the parameters labeled.

Approach

Draw the normal curve with the mean $\mu = 180$ labeled at the high point and the inflection points at $\mu - \sigma = 180 - 36.2 = 143.8$ and $\mu + \sigma = 180 + 36.2 = 216.2$.

Solution

Figure 4(a) shows the graph of the normal curve.

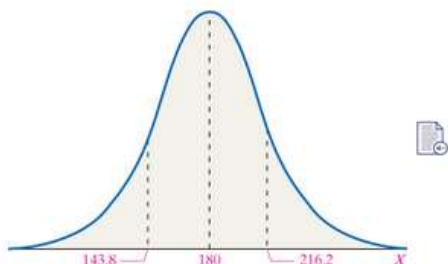


Figure 4(a)

An individual with total cholesterol greater than 200 is considered to have high cholesterol. Shade the region under the normal curve to the right of $x = 200$.

Approach

Shade the region under the normal curve drawn in Figure 4(a) to the right of $x = 200$.

Solution

Figure 4(b) shows the region under the normal curve to the right of $x = 200$ shaded.

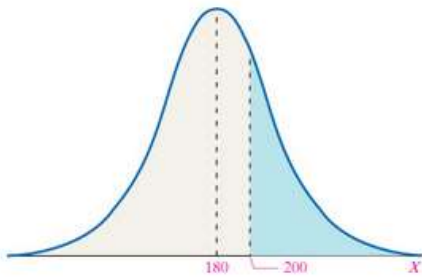


Figure 4(b)

Suppose that the area under the normal curve to the right of $x = 200$ is 0.2903. (You will learn how to find this area in the next section.) Provide two interpretations of this result.

Approach

The two interpretations of the area under a normal curve are (1) a proportion and (2) a probability.

Solution

The two interpretations for the area of this shaded region are as follows:

- (1) The proportion of 20- to 29-year-old males that have high cholesterol is 0.2903.
- (2) The probability that a randomly selected 20- to 29-year-old male has high cholesterol is 0.2903.

7.2 Applications of Normal Distribution

November 8, 2016 08:37 AM

If X is a normally distributed random variable, then the area under the normal curve represents the proportion of the population with a certain characteristic or the probability that a randomly selected individual from the population has the characteristic.

The question then is, "How do I find the area under the normal curve?"

We have two options:

1. Performing calculations by hand with the aid of a table, or
2. Using technology

7.2 Objective 1 Find and Interpret the Area under a Normal Curve

November 8, 2016 08:38 AM

We use z -scores to find the area under a normal curve by hand. Recall that the z -score allows us to transform a random variable X with mean μ and standard deviation σ into a random variable Z with mean 0 and standard deviation 1.

Standardizing a Normal Random Variable

Suppose that the random variable X is normally distributed with mean μ and standard deviation σ .

Then the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$.

The random variable Z is said to have the **standard normal distribution**.

The result from the previous slide is powerful.

If a normal random variable X has a mean different from 0 or a standard deviation different from 1, we can transform X into a **standard normal random variable** Z whose mean is 0 and standard deviation is 1. Then we can use [Table V](#) to find the area to the left of a specified z -score, z , as shown in Figure 5, which is also the area to the left of the value of x in the distribution of X . The graph in Figure 5 is called the **standard normal curve**.

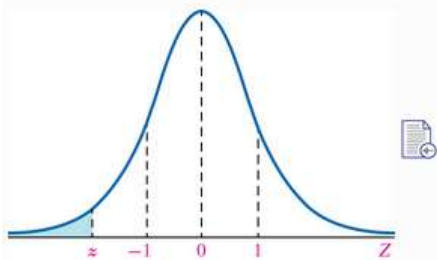


Figure 5

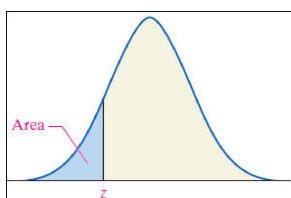


TABLE V

Standard Normal Distribution										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823

Standardizing a Normal Random Variable

Suppose that the random variable X is normally distributed with mean μ and standard deviation σ .

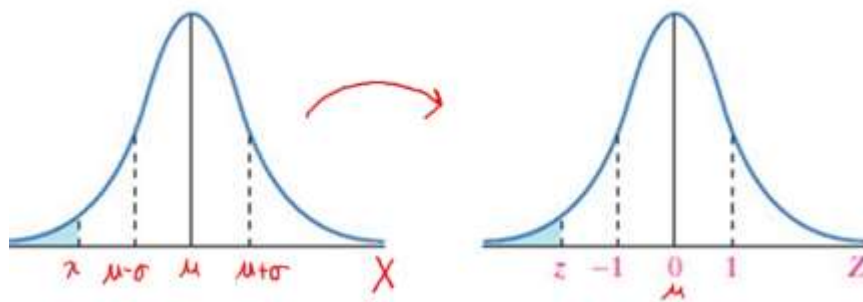
Then the random variable $Z = \frac{X - \mu}{\sigma}$ is

normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$.

The random variable Z is said to have the **standard normal distribution**.

If a normal random variable X has a mean different from 0 or a standard deviation different from 1, we can ...

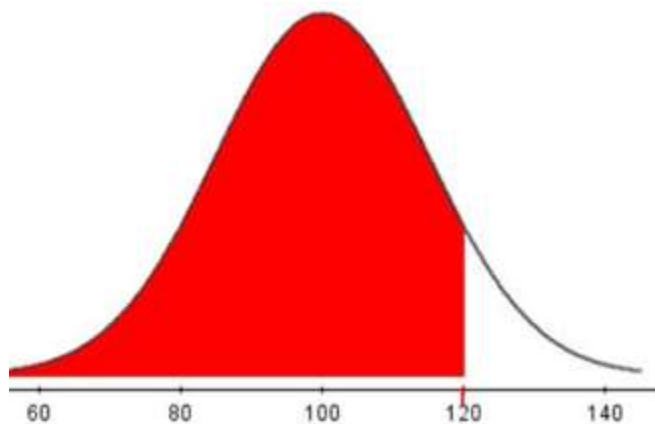
Transform X into a **standard normal random variable** Z whose mean is 0 and standard deviation is 1.



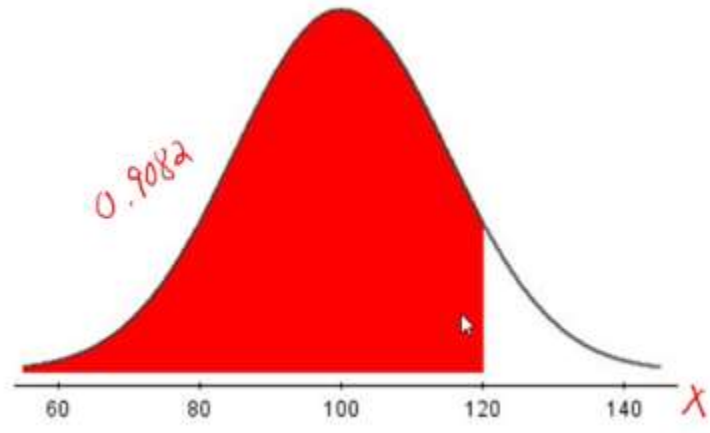
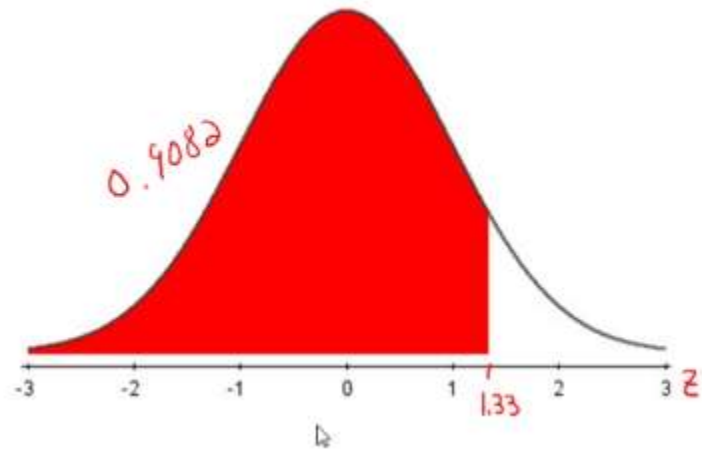
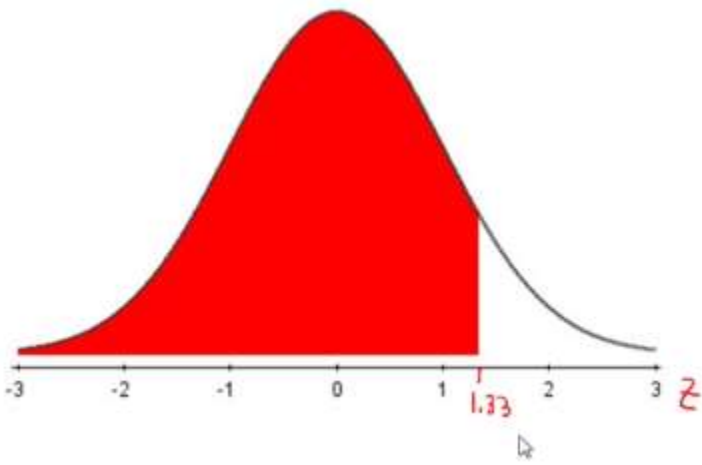
Use Table V (found in Appendix A) to find the area to the left of a specified z -score, z , which is also the area to the left of the value of x in the distribution of X .

IQ scores can be modeled by a normal distribution with $\mu = 100$ and $\sigma = 15$.

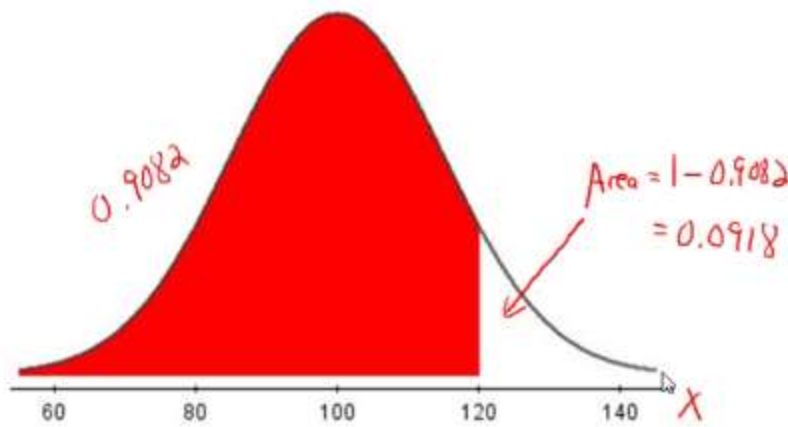
Suppose a person has an IQ of 120.



$$Z = \frac{X - \mu}{\sigma} = \frac{120 - 100}{15} = 1.33$$



Find area of the right by using compliment rule



A Better Table V

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99997	.99997	.99997

IQ scores can be modeled by a normal distribution with $\mu = 100$ and $\sigma = 15$. An individual whose IQ is 120 is

$$z = \frac{x - \mu}{\sigma} = \frac{120 - 100}{15} = 1.33$$

standard deviations above the mean (recall that we round z -scores to two decimal places.) We look in Table V and find that the area under the standard normal curve to the left of $z = 1.33$ is 0.9082. See Figure 6. Therefore, the area under the normal curve to the left of $x = 120$ is 0.9082. Figure 7 illustrates the area to the left of 120 using a normal model.

z	Standard Normal Distribution						
	.00	.01	.02	.03	.04	.05	.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5479	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6025
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406

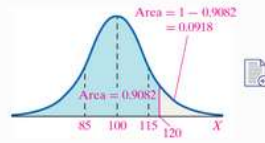


Figure 7

Figure 6

To find the area to the right of the value of a random variable, we use the **Complement Rule** and determine 1 minus the area to the left.

To find the area under the normal curve with mean $\mu = 100$ and standard deviation $\sigma = 15$ to the right of $x = 120$, we compute

$$\text{Area} = 1 - 0.9082 = 0.0918$$

as shown in Figure 7.

EXAMPLE 1 Finding and Interpreting Area Under a Normal Curve

Problem

A pediatrician obtains the heights of her 200 three-year-old female patients as shown in Table 1. The heights are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. See Figure 8. Use the normal model to determine the proportion of the 3-year-old females who have a height less than 35 inches.

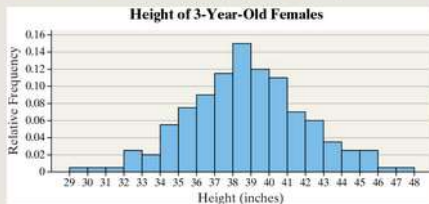
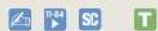


Figure 8

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

Results and Interpretation

By hand, the normal model indicates that the proportion of the pediatrician's 3-year-old females who are less than 35 inches tall is 0.1210. Using technology, the normal model indicates that the proportion of the pediatrician's 3-year-old females who are less than 35 inches tall is 0.1203.

If we randomly selected 100 three-year old females, we would expect about 12 of them to be less than 35 inches tall.

HIDE RESULTS AND INTERPRETATION

TABLE 1

Height (Inches)	Relative Frequency
29.0–29.9	0.005
30.0–30.9	0.005
31.0–31.9	0.005
32.0–32.9	0.025
33.0–33.9	0.02
34.0–34.9	0.055
35.0–35.9	0.075
36.0–36.9	0.09
37.0–37.9	0.115
38.0–38.9	0.15
39.0–39.9	0.12
40.0–40.9	0.11
41.0–41.9	0.07
42.0–42.9	0.06
43.0–43.9	0.035
44.0–44.9	0.025
45.0–45.9	0.025
46.0–46.9	0.005
47.0–47.9	0.005

Finding Area Under a Normal Curve

1. From the HOME screen, press 2^{nd} VARS to access the DISTRIBution menu.
2. Select 2:normalcdf{.
3. Enter the lower bound, upper bound, μ , and σ . For example, to find the area to the left of $x = 35$ under the normal curve with $\mu = 40$ and $\sigma = 10$, enter $-1E99$ for the lower bound, 35 for the upper bound, 40 for μ , and 10 for σ . Highlight Paste and hit ENTER. Hit ENTER again with the formula on the HOME screen.

NOTE

When there is no lower bound, enter $-1E99$. When there is no upper bound, enter $1E99$. The E shown is scientific notation; it is selected by pressing 2^{nd} then the comma key.

Finding Area Under a Normal Curve

1. Select Stat, highlight Calculators, and select Normal.
2. Enter the mean and the standard deviation.
 - If you want to compute $P(X \leq x)$ or $P(X \geq x)$, select Standard. Then select the \leq or \geq from the pull-down menu and enter the value of x . Click Compute.
 - If you want to compute $P(a \leq X \leq b)$, select Between. Then enter the values of a and b . Click Compute.

The Normal Curve as a Model

According to the Example 1 results, the proportion of 3-year-old females who are shorter than 35 inches is approximately 0.12. If the normal curve is a good model for determining proportions (or probabilities), then about 12% of the 200 three-year-olds in Table 1 should be shorter than 35 inches.

The relative frequency distribution in Table 1 shows that

$0.005 + 0.005 + 0.005 + 0.025 + 0.02 + 0.055 = 0.115 = 11.5\%$ of the 3-year-old females are less than 35 inches tall. The results based on the normal curve are close to the actual results. The normal curve accurately models the heights.

If we wanted to know the proportion of 3-year-old females whose height is greater than 35 inches, we would use the [Complement Rule](#) to find that the proportion is $1 - 0.1210 = 0.879$ (using the "by-hand" computation).

Percentiles

Because the area under the normal curve represents a proportion, we can also use the area to find [percentile](#) ranks of scores. In Example 1, 12% of the females have a height less than 35 inches and 88% of the females have a height greater than 35 inches; so a child whose height is 35 inches is at the 12th percentile.

EXAMPLE 2 Finding and Interpreting Area Under a Normal Curve

Problem

A pediatrician obtains the heights of her 200 three-year-old female patients as shown in Table 1. The heights are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. See Figure 8. Use the normal model to determine the probability that a randomly selected 3-year-old female is between 35 and 40 inches tall, inclusive. That is, find $P(35 \leq X \leq 40)$.

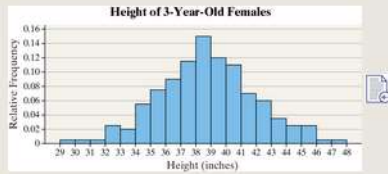


Figure 8

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

Results and Interpretation

By hand, the probability that a randomly selected 3-year-old female is between 35 and 40 inches tall is 0.5344. That is, $P(35 \leq X \leq 40) = 0.5344$. Using technology, the probability that a randomly selected 3-year-old female is between 35 and 40 inches tall is 0.5365. That is, $P(35 \leq X \leq 40) = 0.5365$.

If we randomly selected 100 three-year-old females, we would expect about 53 or 54 of them to be between 35 and 40 inches tall.

HIDE RESULTS AND INTERPRETATION

TABLE 1

Height (Inches)	Relative Frequency
29.0–29.9	0.005
30.0–30.9	0.005
31.0–31.9	0.005
32.0–32.9	0.025
33.0–33.9	0.02
34.0–34.9	0.055
35.0–35.9	0.075
36.0–36.9	0.09
37.0–37.9	0.115
38.0–38.9	0.15
39.0–39.9	0.12
40.0–40.9	0.11
41.0–41.9	0.07
42.0–42.9	0.06
43.0–43.9	0.035
44.0–44.9	0.025
45.0–45.9	0.025
46.0–46.9	0.005
47.0–47.9	0.005

According to the relative frequency distribution in Table 1, the proportion of the 200 three-year-old females with heights between 35 and 40 inches is $0.075 + 0.09 + 0.115 + 0.15 + 0.12 = 0.55 = 55\%$. This is very close to the probability of 0.5365 that a randomly selected 3-year-old female is between 35 and 40 inches tall found from the normal model.

7.2 Objective 2 Find the Value of a Normal Random Variable

November 8, 2016 08:54 AM

Often, we do not want to find the proportion, probability, or percentile given a value of a normal random variable. Rather, we want to find the value of a normal random variable that corresponds to a certain proportion, probability, or percentile. For example, we might want to know the height of a 3-year-old girl who is at the 20th percentile. Or we might want to know the scores on a standardized exam that separate the middle 90% of scores from the bottom and top 5%.

EXAMPLE 3 Finding the Value of a Normal Random Variable

Problem

The heights of a pediatrician's 3-year-old females are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. Find the height of a 3-year-old female at the 20th percentile.

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

Result

The height of a 3-year-old female at the 20th percentile is 36.1 inches.

HIDE RESULT

TI-83/84 Plus

StatCrunch

Finding the Value of a Normal Random Variable

1. From the HOME screen, press 2nd VARS to access the DISTRibution menu.
2. Select 3: invNorm(.
3. Enter the area to the left, μ , and σ . For example, to find the normal value such that the area under the normal curve to the left of the value is 0.68, with $\mu = 40$ and $\sigma = 10$, enter 0.68 for the area, 40 for μ , and 10 for σ . Highlight Paste and hit ENTER. Hit ENTER again with the formula on the HOME screen.


TI-83/84 Plus

StatCrunch

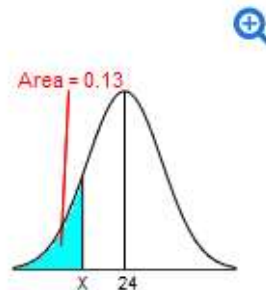
Finding the Value of a Normal Random Variable

1. Select Stat, highlight Calculators, and select Normal.
2. Enter the mean and the standard deviation. Select the Standard option. In the pull-down menu, decide if you are given the area to the left of the unknown score or the area to the right. If given the area to the left, in the pull-down menu, choose the \leq option; if given the area to the right, choose the \geq option. Finally, enter the area in the rightmost cell. Click Compute.

The mean incubation time of fertilized eggs is 24 days. Suppose the incubation times are approximately normally distributed with a standard deviation of 1 day. Determine the 13th percentile for incubation times.

 Click the icon to view a table of areas under the normal curve.

The first step is to draw a normal curve and shade the area corresponding to the 13th percentile. The 13th percentile means that 13% of the data is less than the random variable X . That is, the area under the normal curve to the left of X is equal to 0.13.



While technology or a standard normal table can be used to find the Z -score, in this problem, technology will be used. Use technology to find the Z -score that corresponds to an area of 0.13.

The Z -score associated with an area of 0.13 is -1.13 , rounded to the nearest hundredth.

Obtain the normal value from the fact that $X = \mu + Z\sigma$, rounding to the nearest integer.

$$\begin{aligned} X &= 24 + (-1.13)(1) \\ &= 23 \end{aligned}$$

Thus, the 13th percentile for incubation times is 23 days.

EXAMPLE 4 Finding the Value of a Normal Random Variable

Problem

The scores earned on the mathematics portion of the SAT, a college entrance exam, are approximately normally distributed with mean 516 and standard deviation 116. What scores separate the middle 90% of test takers from the bottom and top 5%? In other words, find the 5th and 95th percentiles. Data from The College Board

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

Result

SAT mathematics scores that separate the middle 90% of the scores from the bottom and top 5% are 325 and 707, respectively. Put another way, a student who scores 325 on the SAT math exam is at the 5th percentile. A student who scores 707 on the SAT math exam is at the 95th percentile. We might use these results to identify those scores that are unusual.

HIDE RESULT

Important Notation for the Future

In upcoming chapters, we must find the z -score that has a specified area to the right. We have special notation to represent this situation.

The notation z_α (pronounced "z sub alpha") is the z -score such that the area under the standard normal curve to the right of z_α is α . Figure 9 illustrates the notation.

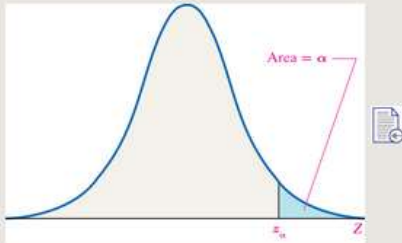


Figure 9

EXAMPLE 5 Finding the Value of z_α

Problem

Find the value of $z_{0.10}$.

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

Result

The value of $z_{0.10} = 1.28$.

HIDE RESULT

TI-83/84 Plus

StatCrunch

Finding the Value of z_α

1. From the HOME screen, press 2nd VARS to access the DISTRibution menu.
2. Select 3: invNorm(.
3. Enter the area to the left, $1 - \alpha$. Enter 0 for μ and 1 for σ . For example, to find $z_{0.20}$, enter 0.80 for the area, 0 for μ , and 1 for σ . Highlight Paste and hit ENTER. Hit ENTER again with the formula on the HOME screen.

Finding the Value of z_{α}

1. Select Stat, highlight Calculators, and select Normal.
2. Enter 0 for the mean and 1 for the standard deviation. Select the Standard option. In the pull-down menu, choose the \geq option. Finally, enter the value of α in the right-most cell. Click Compute.

Find the value of z_{α} .

$z_{0.46}$



Click the icon to view a table of areas under the normal curve.

$z_{0.46} = .10$ (Round to two decimal places as needed.)

```
invNorm(1-.46,0,1)
.1004337118
```

For any continuous random variable, the probability of observing a specific value of the random variable is 0. For example, for a normal random variable, $P(a) = 0$ for any value of a because there is no area under the normal curve associated with a single value. Therefore, the following probabilities are equivalent:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

7.3 Assessing Normality

November 10, 2016 09:24 AM

Up to this point, we have said that a random variable X is normally distributed, or at least approximately normal, provided the histogram of the data is symmetric and bell-shaped. This works well for large data sets, but the shape of a histogram drawn from a small sample of observations does not always accurately represent the shape of the population. For this reason, we need additional methods for assessing the normality of a random variable X when we are looking at a small set of sample data.

7.3 Objective 1 Use normal Probability Plots to Assess Normality

November 10, 2016 09:39 AM

A **normal probability plot** is a graph that plots observed data versus *normal scores*. A **normal score** is the expected z -score of the data value, assuming that the distribution of the random variable is normal. The expected z -score of an observed value depends on the number of observations in the data set.

Although it is recommended that you use technology to draw normal probability plots, we present below the steps for drawing a normal probability plot by hand. These steps help to explain the logic behind reading a normal probability plot.

Drawing a Normal Probability Plot by Hand

Step 1 Arrange the data in ascending order.

Step 2 Compute

$$f_i = \frac{i - 0.375}{n + 0.25}$$

where i is the index (the position of the data value in the ordered list) and n is the number of observations. The expected proportion of observations less than or equal to the i th data value is f_i .

Step 3 Find the z -score corresponding to f_i from [Table V](#).

Step 4 Plot the observed values on the horizontal axis and the corresponding expected z -scores on the vertical axis.

The idea behind finding the expected z -score is, if the data come from a normally distributed population, we could predict the area to the left of each data value. The value of f_i represents the expected area to the left of the i th observation when the data come from a population that is normally distributed. For example, f_1 is the expected area to the left of the smallest data value, f_2 is the expected area to the left of the second-smallest data value, and so on. See Figure 10.

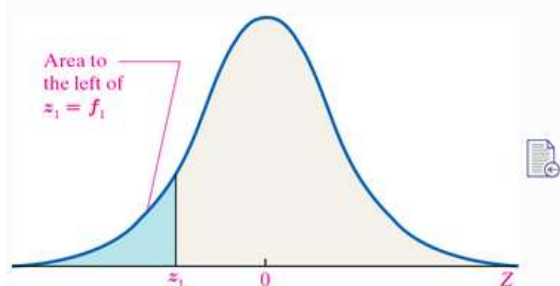


Figure 10

Once we determine each f_i , we find the z -scores corresponding to f_1 , f_2 , and so on. The smallest observation in the data set will be the smallest expected z -score, and the largest observation in the data set will be the largest expected z -score. Also, because of the symmetry of the normal curve, the expected z -scores are always paired as positive and negative values.

Values of normal random variables and their z -scores are linearly related ($x = \mu + z\sigma$), so a plot of observations of normal variables against their expected z -scores will be linear. We conclude the following:

If sample data are taken from a population that is normally distributed, a normal probability plot of the observed values versus the expected z -scores will be approximately linear.

It is difficult to determine whether a normal probability plot is “linear enough.” However, we can use a procedure based on the research of S. W. Looney and T. R. Gullledge in their paper “Use of the Correlation Coefficient with Normal Probability Plots,” published in the *American Statistician*. Basically, if the linearly correlation coefficient between the observed values and expected z -scores is greater than the critical value found in Table VI, then it is reasonable to conclude that the data could come from a population that is normally distributed.

Normal probability plots are typically drawn using technology (graphing calculators or statistical software). However, it is worthwhile to go through an example that demonstrates how to draw a normal probability plot by hand to better understand the results supplied by technology.

Sample Size, n	Critical Value
5	0.880
6	0.888
7	0.898
8	0.906
9	0.912
10	0.918
11	0.923
12	0.928
13	0.932
14	0.935
15	0.939
16	0.941
17	0.944
18	0.946
19	0.949
20	0.951
21	0.952
22	0.954
23	0.956
24	0.957
25	0.959
30	0.960

S. W. Looney and T. R. Gullledge, Jr. “Use of the Correlation Coefficient with Normal Probability Plots,” *American Statistician* 39(Feb. 1985): 75-79.

EXAMPLE 1 Drawing a Normal Probability Plot by Hand

Problem

The data in Table 2 represent the finishing time (in seconds) for six randomly selected races of a greyhound named Barbies Bomber in the 5/16-mile race at

Greyhound Park in Dubuque, Iowa. Is there evidence to support the belief that the variable “finishing time” is normally distributed?

Video Solution



TABLE 2

31.35	32.52
32.06	31.26
31.91	32.37

Data from Greyhound Park, Dubuque, IA

Approach

Approach

Follow [Steps 1 through 4](#) to draw the normal probability plot.

Solution

Step 1 Column 1 in Table 3 represents the index i . Column 2 represents the observed values in the data set, written in ascending order.

TABLE 3

Index, i	Observed Value
1	31.26
2	31.35
3	31.91
4	32.06
5	32.37
6	32.52

Step 2 Column 3 in Table 3 represents $f_i = \frac{i-0.375}{n+0.25}$ for each observation. This value is the expected area under the normal curve to the left of the i th observation, assuming normality. For example, $i = 1$ corresponds to the finishing time of 31.26, and

$$f_1 = \frac{1 - 0.375}{6 + 0.25} = 0.10$$

So we expect the area under the normal curve to the left of 31.26 to be 0.10 if the sample data come from a population that is normally distributed.

TABLE 3

Index, i	Observed Value	f_i
1	31.26	$\frac{1 - 0.375}{6 + 0.25} = 0.10$
2	31.35	$\frac{2 - 0.375}{6 + 0.25} = 0.26$
3	31.91	0.42
4	32.06	0.58
5	32.37	0.74
6	32.52	0.90

Step 3 Use Table V to find the z -scores that correspond to each f_i , then list them in Column 4 of Table 3. Look in Table V for the area closest to $f_1 = 0.1$. The expected z -score is -1.28 . Notice that for each negative expected z -score there is a corresponding positive expected z -score as a result of the symmetry of the normal curve.

TABLE 3

Index, i	Observed Value	f_i	Expected z -score
1	31.26	$\frac{1 - 0.375}{6 + 0.25} = 0.10$	-1.28
2	31.35	$\frac{2 - 0.375}{6 + 0.25} = 0.26$	-0.64
3	31.91	0.42	-0.20
4	32.06	0.58	0.20
5	32.37	0.74	0.64
6	32.52	0.90	1.28

Step 4 Plot the actual observations on the horizontal axis and the expected z -scores on the vertical axis. See Figure 11.

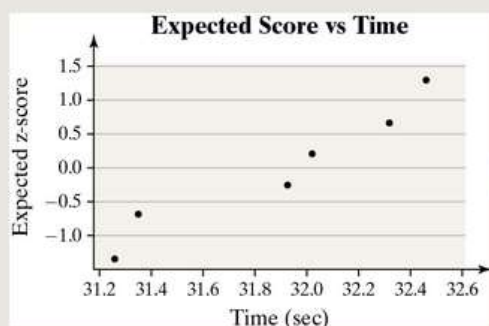


Figure 11

Interpretation The linear correlation between the observed values and expected z -scores from the data in Table 3 is 0.967 . The critical value in Table VI for $n = 6$ observations is 0.888 . Because the correlation coefficient is greater than the critical value ($0.967 > 0.888$), it is reasonable to conclude that the finishing times of Barbies Bomber in the 5/16-mile race are approximately normally distributed.

Typically, normal probability plots are drawn using either a graphing calculator with advanced statistical features or statistical software such as StatCrunch.

This will be explained in the next example.

TABLE 2

31.35	32.52
32.06	31.26
31.91	32.37

Data from Greyhound Park, Dubuque, IA

EXAMPLE 2 Drawing a Normal Probability Plot Using Technology**Problem**

Draw a normal probability plot of the Barbies Bomber data in Table 2 using technology. Is there evidence to support the belief that the variable “finishing time” is normally distributed?

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

TI-83/84 Plus

StatCrunch

Drawing Normal Probability Plots

1. Enter the raw data into L1.
2. Press 2^{nd} Y = to access STAT PLOTS.
3. Select 1: Plot1.
4. Turn Plot1 on by highlighting On and pressing ENTER. Press the down-arrow key and highlight the normal probability plot icon. Press ENTER to select this plot type. The Data List should be set at L1. The Data Axis should be the x -axis.
5. Press ZOOM and select 9: ZoomStat.

Once you have the graph, TRACE to find the values of the observations and the corresponding normal scores. Enter these observations into L1 and L2. Find the correlation coefficient for this data.

TI-83/84 Plus

StatCrunch

Drawing Normal Probability Plots

1. If necessary, enter the raw data into column var1. Name the column.
2. Select **Graph** and highlight **QQ Plot**.
3. Select the variable. Check the box to add the correlation statistic. Click **Compute!**

EXAMPLE 3 Assessing Normality

Problem

The data in Table 4 represent the time 100 randomly selected riders spent waiting in line (in minutes) for the Demon Roller Coaster. Is the random variable “wait time” normally distributed?

TABLE 4

7	3	5	107	8	37	16	41	7	25	22	19	1	40	1	29	93
33	76	14	8	9	45	15	81	94	10	115	18	0	18	11	60	34
30	6	21	0	86	6	11	1	1	3	9	79	41	2	9	6	19
4	3	2	7	18	0	93	68	6	94	16	13	24	6	12	121	30
35	39	9	15	53	9	47	5	55	64	51	80	26	24	12	0	
94	18	4	61	38	38	21	61	9	80	18	21	8	14	47	56	

Solution

Video Solution Technology Step-By-Step



HIDE SOLUTION

7.4 The Normal Approximation to the Binomial Probability Distribution

November 10, 2016 09:51 AM

7.4 Objective 1 Approximate Binomial Probabilities Using the Normal Distribution

November 10, 2016 10:15 AM

Criteria for a Binomial Probability Experiment

A probability experiment is a binomial experiment if all the following are true:

1. The experiment is performed n independent times. Each repetition of the experiment is called a **trial**.
2. For each trial, there are two mutually exclusive outcomes—success or failure.
3. The probability of success, p , is the same for each trial of the experiment.

Binomial Probability Formula

$$P(x) = {}_n C_x \cdot p^x (1-p)^{n-x}$$

Large Number of Trials

Suppose, given 500 trials of a binomial experiment, you need to compute the probability of 400 or more successes.

$$\begin{aligned} P(X \geq 400) \\ &= P(400) + P(401) + \dots + P(500) \\ &= 1 - [P(0) + P(1) + \dots + P(399)] \end{aligned}$$

Recall From Section 6.2 ...

For a fixed p , as the number of trials n in a binomial experiment increases, the probability distribution of the random variable X becomes more nearly symmetric and bell shaped.

As a rule of thumb, if $np(1 - p) \geq 10$, the probability distribution will be approximately symmetric and bell shaped.

From the video, we conclude that binomial probabilities can be approximated by the area under a normal curve, provided that $np(1 - p) \geq 10$.

The Normal Approximation to the Binomial Probability Distribution

If $np(1 - p) \geq 10$, the binomial random variable X is approximately normally distributed, with mean $\mu_X = np$ and standard deviation $\sigma_X = \sqrt{np(1 - p)}$.

Note: In a binomial experiment, n is the number of trials and p is the probability of success.

Figure 12 shows a graph for the binomial random variable X , with $n = 40$ and $p = 0.5$, drawn in StatCrunch.

Because $np(1 - p) = 40(0.5)(1 - 0.5) = 10$, we can use a normal model with mean $\mu_X = np = 40(0.5) = 20$ and standard deviation $\sigma_X = \sqrt{np(1 - p)} = \sqrt{40(0.5)(1 - 0.5)} = \sqrt{10}$ to describe the random variable X .

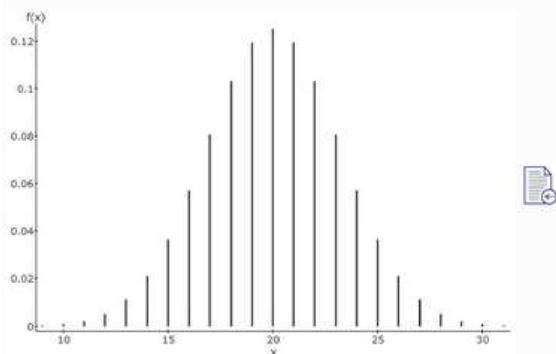
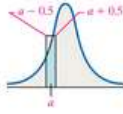
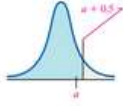
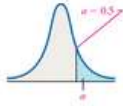
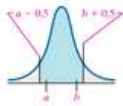


Figure 12

To approximate the probability of a specific value of the binomial random variable, such as $P(18)$, find the area under the normal curve from $x = 17.5$ to $x = 18.5$. We add and subtract 0.5 from $x = 18$ as a **correction for continuity** because we are using a continuous density function to approximate a discrete probability.

To approximate $P(X \leq 18)$, compute the area under the normal curve for $X \leq 18.5$. To approximate $P(X \geq 18)$, compute $P(X \geq 17.5)$. Do you see why?

TABLE 5

Exact Probability Using Binomial	Approximate Probability Using Normal Model	Graphical Depiction
$P(a)$	$P(a - 0.5 \leq X \leq a + 0.5)$	
$P(X \leq a)$	$P(X \leq a + 0.5)$	
$P(X \geq a)$	$P(X \geq a - 0.5)$	
$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5)$	

A question remains, however. What do we do if the probability is of the form $P(X > a)$, $P(X < a)$, or $P(a < X < b)$? The solution is to rewrite the inequality in a form with \leq or \geq . For example, $P(X > 4) = P(X \geq 5)$ and $P(X < 4) = P(X \leq 3)$ for binomial random variables because the values of the random variables must be whole numbers.

EXAMPLE 1 The Normal Approximation to a Binomial Random Variable

Problem

According to the American Red Cross, 7% of people in the United States have blood type O-negative. What is the probability that in a simple random sample of 500 people in the United States fewer than 30 have blood type O-negative?

Video Solution



Approach

Step 1 Verify that this is a binomial experiment.

Step 2 Computing the probability by hand would be very tedious. Verify $np(1 - p) \geq 10$. Then use the normal distribution to approximate the binomial probability.

Step 3 Approximate $P(X < 30) = P(X \leq 29)$ using the normal approximation to the binomial distribution.

Solution

Step 1 Each of the 500 independent trials has a probability of success equal to 0.07. This is a **binomial experiment**.

Step 2 Verify that $np(1 - p) \geq 10$.

$$\begin{aligned} np(1 - p) &= 500(0.07)(1 - 0.07) \\ &= 32.55 \end{aligned}$$

Since $np(1 - p) \geq 10$ we can use the normal distribution to approximate the binomial distribution.

Step 3 The probability that fewer than 30 people in the sample have blood type O-negative is

$P(X < 30) = P(X \leq 29)$. This is approximately equal to the area under the normal curve to the left of $x = 29.5$, with

$$\mu_X = np = 500(0.07) = 35 \text{ and } \sigma_X = \sqrt{np(1 - p)} = \sqrt{500(0.07)(1 - 0.07)} = \sqrt{32.55}$$

See Figure 13. The area to the left of $x = 29.5$ is 0.1685. Therefore, the approximate probability that fewer than 30 people in a simple random sample of 500 people will have blood type O-negative is 0.1685.

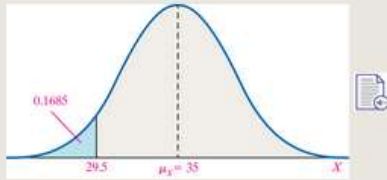


Figure 13

Using the binomial probability distribution and StatCrunch or a TI-84 Plus graphing calculator, we find that the exact probability is 0.1678. See Figure 14. The approximate result using the normal model is 0.1685, which is close to the exact probability. Also, notice the shape of the distribution in the StatCrunch output.

```
binomcdf(500,.07,29)
.1677676733
```

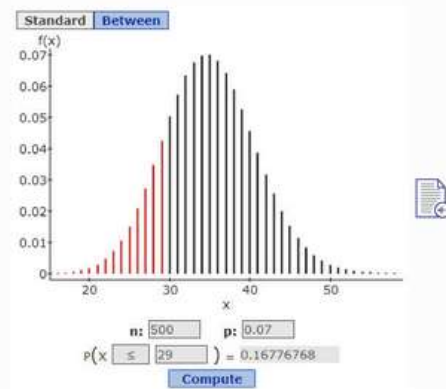


Figure 14

EXAMPLE 2 A Normal Approximation to the Binomial

Problem

According to the Gallup Organization, 65% of adult Americans are in favor of the death penalty for individuals convicted of murder. Erica selects a random sample of 1000 adult Americans in Will County, Illinois, and finds that 630 of them are in favor of the death penalty for individuals convicted of murder.

Video Solution



Part (A)

Part (B)

Assuming that 65% of adult Americans in Will County are in favor of the death penalty, what is the probability of obtaining a random sample of no more than 630 adult Americans in favor of the death penalty from a sample of size 1000?

Approach

This is a binomial experiment with $n = 1000$ and $p = 0.65$. Erica needs to determine the probability of obtaining a random sample of no more than 630 adult Americans who favor the death penalty, assuming that 65% of adult Americans favor the death penalty. Computing this probability using the binomial probability formula would be difficult, so Erica will use the normal approximation to the binomial because $np(1 - p) = 1000(0.65)(1 - 0.65) = 227.5 \geq 10$. Approximate $P(X \leq 630)$ by determining the area under the normal curve to the left of $x = 630.5$ with $\mu_X = np = 650$ and $\sigma_X = \sqrt{np(1 - p)} = \sqrt{1000(0.65)(1 - 0.65)} \approx 15.083$.

Solution

The area under the normal curve to the left of $X = 630.5$ is 0.0985. See Figure 16. The probability of obtaining 630 or fewer adult Americans who favor the death penalty from a sample of 1000 adult Americans, assuming the proportion of adult Americans who favor the death penalty is 0.65, is 0.0985.

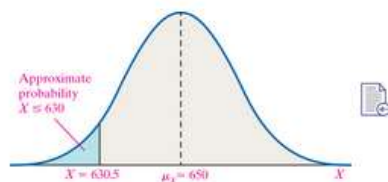


Figure 16

Does the result from part (A) contradict the Gallup Organization's findings? Explain.

Approach

Remember, results are unusual if the probability of the event is less than 0.05 . Determine whether obtaining 630 or fewer individuals from a sample of 1000 individuals is unusual if the probability of success (finding an individual in favor of the death penalty) is 0.65 .

Solution

From part (A): The probability of obtaining 630 or fewer adult Americans who favor the death penalty from a sample of 1000 adult Americans, assuming that the proportion of adult Americans who favor the death penalty is 0.65 , is 0.0985 .

Had we obtained 100 different simple random samples of size 1000 we would expect about 10 to result in 630 or fewer adult Americans favoring the death penalty if the true proportion is 0.65 . Because the results obtained are not unusual under the assumption that $p = 0.65$, Erica finds that the results of her survey do not contradict those of Gallup.

7.4.28

November 10, 2016 11:36 AM

7.4 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/10/16

Instructor: Matthew Naby
Course: MTH 243: Introduction to
Probability and Statistics

Assignment: 7.4 Interactive Assignment

A certain flight arrives on time 80 percent of the time. Suppose 100 flights are randomly selected. Use the normal approximation to the binomial to approximate the probability that

- (a) exactly 85 flights are on time.
- (b) at least 85 flights are on time.
- (c) fewer than 75 flights are on time.
- (d) between 75 and 90 are on time.

Because $np(1-p) \geq 10$, we can use the normal distribution with $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$ to approximate the binomial probability. Find μ_X and σ_X .

$$\begin{aligned}\mu_X &= np & \sigma_X &= \sqrt{np(1-p)} \\ &= 100 \cdot 0.80 & &= \sqrt{100 \cdot 0.80(1-0.80)} \\ &= 80 & &= 4\end{aligned}$$

(a) Approximate $P(85)$ using the normal distribution. First adjust for continuity by determining an interval of width 1 centered about X and find the probability that X is within that interval.

$$P(85) \approx P(84.5 < X < 85.5)$$

Convert $X = 84.5$ and $X = 85.5$ to normalized values Z_0 and Z_1 respectively.

$$\begin{aligned}Z_0 &= \frac{84.5 - 80}{4} & Z_1 &= \frac{85.5 - 80}{4} \\ &\approx 1.13 & &\approx 1.38\end{aligned}$$

$P(85)$ is approximated by finding the area under the curve between Z_0 and Z_1 .

$$\begin{aligned}P(85) &= P(1.38 < Z < 1.13) \\ &= P(Z < 1.38) - P(Z < 1.13) \\ &\approx 0.9162 - 0.8708 \\ &= 0.0454\end{aligned}$$

Thus, the probability that exactly 85 of the 100 flights are on time is approximately 0.0454.

(b) Approximate $P(X \geq 85)$. Because normal probabilities in the z-table are given for X less than or equal to a certain value, approximate $P(X \geq 85)$ as $1 - P(X \leq 84)$. Given that the inequality sign is less than or equal to, adjust for continuity by adding 0.5 to X .

$$P(X \geq 85) \approx 1 - P(X \leq 84) \approx 1 - P(X < 84.5)$$

Using $\mu_X = 80$ and $\sigma_X = 4$, find the appropriate value of Z .

$$Z = \frac{84.5 - 80}{4} \approx 1.13$$

Because the probability can be described as the area under the curve at or to the right of 85, approximate $P(X \geq 85)$ as follows.

$$\begin{aligned}P(X \geq 85) &\approx 1 - P(Z < 1.13) \\ &\approx 1 - 0.8708 \\ &= 0.1292\end{aligned}$$

Thus, the probability that at least 85 of the 100 flights are on time is approximately 0.1292.

(c) Approximate $P(X < 75)$. Write the the probability as $P(X \leq 74)$. Adjust for continuity by adding 0.5 to X .

$$P(X < 75) = P(X < 74.5)$$

Using $\mu_X = 80$ and $\sigma_X = 4$, find the appropriate value of Z .

$$Z = \frac{74.5 - 80}{4} \approx -1.38$$

Calculate $P(X < 75)$.

$$P(X < 75) \approx P(Z \leq -1.38) \approx 0.0838$$

Thus, the probability that fewer than 75 of the 100 flights are on time is approximately 0.0838.

(d) Approximate $P(75 \leq X \leq 90)$. Given that the interval for X includes the endpoints, adjust the interval by subtracting 0.5 from the lower limiting value and add 0.5 to the upper limiting value.

$$P(75 \leq X \leq 90) \approx P(74.5 < X < 90.5)$$

Using $\mu_X = 80$ and $\sigma_X = 4$, find the normalized values Z_0 and Z_1 corresponding to X -values $X_0 = 74.5$ and $X_1 = 90.5$.

$$\begin{aligned} Z_0 &= \frac{74.5 - 80}{4} & Z_1 &= \frac{90.5 - 80}{4} \\ &\approx -1.38 & &\approx 2.63 \end{aligned}$$

The probability can be described as the area under the curve between 75 and 90. Find it by calculating $P(-1.38 \leq Z \leq 2.63)$.

$$\begin{aligned} P(-1.38 \leq Z \leq 2.63) &= P(Z \leq 2.63) - P(Z \leq -1.38) \\ &\approx 0.9957 - 0.0838 \\ &= 0.9119 \end{aligned}$$

Thus, the probability that 75 to 90 of the 100 flights are on time is approximately 0.9119.

8.1 Distribution of the Sample Mean

November 17, 2016 08:52 AM

Suppose the government wanted to determine the mean income of all U.S. households. One approach the government could take is literally to survey each U.S. household to determine the population mean. This would be a very expensive and time-consuming survey.

A second approach the government could (and does) take is to survey a random sample of U.S. households and use the results to estimate the mean household income. The American Community Survey is administered to approximately 250,000 randomly selected households each month. Among the many questions on the survey, respondents are asked to report the income of each individual in the household. From this information, the federal government obtains a sample mean U.S. household income. For example, in 2012, the mean annual U.S. household income was approximately $\bar{x} = \$71,274$. The government might infer from this survey that the mean annual household income of all U.S. households in 2012 was $\mu = \$71,274$.

The households in the American Community Survey were determined by chance (random sampling). A second random sample of households would likely lead to a different sample mean, such as $\bar{x} = \$71,518$, and a third random sample of households would likely lead to a third sample mean, such as $\bar{x} = \$71,678$. Because the households selected will vary from sample to sample, the sample mean of household income will also vary from sample to sample. For this reason, the sample mean is a random variable; so it has a probability distribution. Our goal in this section is to describe the distribution of the sample mean. Remember, when we describe a distribution, we do so in terms of its shape, center, and spread.

DEFINITION

The **sampling distribution** of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n .

The **sampling distribution of the sample mean** \bar{x} is the probability distribution of all possible values of the random variable \bar{x} computed from a sample of size n from a population with mean μ and standard deviation σ .

The idea behind finding the sampling distribution of the mean is as follows:

Steps for Determining the Sampling Distribution

Step 1 Obtain a simple random sample of size n .

Step 2 Compute the sample mean.

Step 3 Assuming that we are sampling from a **finite** population, repeat Steps 1 and 2 until all distinct simple random samples of size n have been obtained.

NOTE

Once a particular sample is obtained, it cannot be obtained a second time.

Finite:

If the number of individuals in a population is a positive integer, we say that the population is **finite**. Otherwise, the population is infinite.

8.1 Objective 1 Describe the Distribution of the Sample Mean: Normal Population

November 17, 2016 08:53 AM

From the results of this video, we conclude that the mean of the distribution of the sample mean, \bar{x} , equals the mean of the parent population and as the sample size n increases, the standard deviation of the distribution of \bar{x} decreases.

The Mean and Standard Deviation of the Sampling Distribution of \bar{x}

Suppose that a simple random sample of size n is drawn from a population with mean μ and standard deviation σ . The sampling distribution of \bar{x} has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation of the sampling distribution of \bar{x} , $\sigma_{\bar{x}}$, is called the **standard error of the mean**.

NOTE

Technically, we assume that we are drawing a simple random sample from an infinite population. For populations of finite size N , $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$. However, if the sample size is less than 5% of the population size ($n < 0.05N$), the effect of $\sqrt{\frac{N-n}{N-1}}$ (the **finite population correction factor**) can be ignored without significantly affecting the result.

For the population presented in the video, the mean was $\mu = 100$ and the standard deviation was $\sigma = 15$. In the first simulation from the video, we drew a random sample of size $n = 9$ so the sampling distribution of \bar{x} will have mean $\mu_{\bar{x}} = 100$ and standard deviation

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{15}{\sqrt{9}} \\ &= 5\end{aligned}$$

This standard error of the mean is close to the approximate standard error of 4.86 found in the first simulation.

In the second simulation from the video, where $n = 20$, the sampling distribution of \bar{x} will have mean $\mu_{\bar{x}} = 100$ and standard deviation

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{15}{\sqrt{20}} \\ &\approx 3.35\end{aligned}$$

This standard error of the mean is close to the approximate standard error of 3.37 found in the second simulation.

Now that we can find the mean and standard deviation for any sampling distribution of \bar{x} , we can concentrate on the shape of the distribution. In the simulations from the video, both histograms of the 1000 different sample means appear to be normal. Recall that the population from which the sample was drawn was also normal. This result leads us to believe that if the population is normal, then the distribution of the sample mean is also normal.

The Shape of the Sampling Distribution of \bar{x} If X Is Normal

If a random variable X is normally distributed, the distribution of the sample mean, \bar{x} , is normally distributed.

For example, the IQ scores of individuals are modeled by a normal random variable with mean $\mu = 100$ and standard deviation $\sigma = 15$. The distribution of the sample mean, \bar{x} , the mean IQ of a simple random sample of $n = 9$ individuals, is normal, with mean $\mu_{\bar{x}} = 100$ and standard deviation $\sigma_{\bar{x}} = \frac{15}{\sqrt{9}} = 5$. See Figure 1.

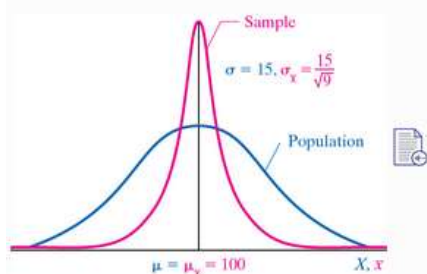


Figure 1

EXAMPLE 1 Finding Probabilities of a Sample Mean

Problem

The IQ, X , of humans is approximately normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. Compute the probability that a simple random sample of size $n = 10$ results in a sample mean greater than 110. That is, compute $P(\bar{x} > 110)$.

Approach

The random variable X is normally distributed, so the sampling distribution of \bar{x} will also be normally distributed. Verify the independence requirement. The mean of the sampling distribution is $\mu_{\bar{x}} = \mu$, and its standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. To find the probability by hand, convert the sample mean $\bar{x} = 110$ to a z -score and then find the area under the standard normal curve to the right of this z -score, or use technology to find the area.

Solution

Video Solution Technology Step-By-Step



Results and Interpretation

The probability of obtaining a sample mean IQ greater than 110 from a population whose mean is 100 is approximately 0.02. That is, $P(\bar{x} > 110) = 0.0174$ (or 0.0175 using technology). If we take 100 simple random samples of $n = 10$ individuals from this population and if the population mean is 100, about 2 of the samples will result in a mean IQ that is greater than 110.

8.1 Objective 2 Describe the Distribution of the Sample Mean: Non-Normal Population

November 17, 2016 09:19 AM

There are two key concepts to understand from the previous video.

1. The mean of the sampling distribution of the sample mean is equal to the mean of the underlying population. That is, $\mu_{\bar{x}} = \mu$. The standard deviation of the sampling distribution of the sample mean is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, regardless of the size of the sample.
2. The shape of the distribution of the sample mean becomes approximately normal as the sample size n increases, regardless of the shape of the underlying population.

We formally state point 2 as the *Central Limit Theorem*.

The Central Limit Theorem

Regardless of the shape of the underlying population, the sampling distribution of \bar{x} becomes approximately normal as the sample size, n , increases.



How large does the sample size need to be before we can say that the sampling distribution of \bar{x} is approximately normal? The answer depends on the shape of the distribution of the underlying population. Distributions that are highly **skewed** will require a larger sample size for the distribution of \bar{x} to become approximately normal.

For example, the right-skewed distribution in the video required a sample size of about 30 before the distribution of the sample mean became approximately normal. However, Figure 2(a) shows a uniform distribution for $0 \leq X \leq 10$. Figure 2(b) shows the distribution of the sample mean obtained via simulation using StatCrunch for $n = 4$. So, for samples as small as $n = 4$, the distribution of the sample mean is approximately normal.

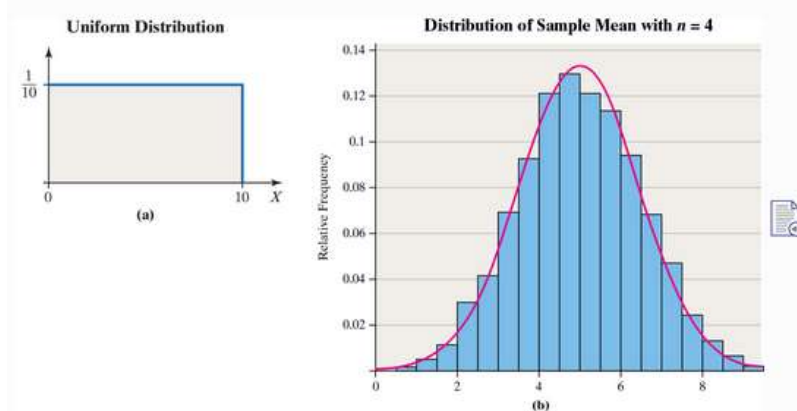


Figure 2

Table 1 shows the distribution of the cumulative number of children for 50- to 54-year-old mothers who had a live birth.

x (number of children)	$P(x)$
1	0.329
2	0.252
3	0.161
4	0.103
5	0.058
6	0.034
7	0.016
8	0.047

Data from National Vital Statistics Report

Figure 3(a) shows the graph of the probability distribution for this distribution. Figure 3(b) shows the distribution of the sample mean for a random sample of $n = 3$ mothers. Figure 3(c) shows the distribution of the sample mean for a random sample of $n = 12$ mothers, and Figure 3(d) shows the distribution of the sample mean for a random sample of $n = 20$ mothers. In this instance, the distribution of the sample mean is close to normal for $n = 20$.

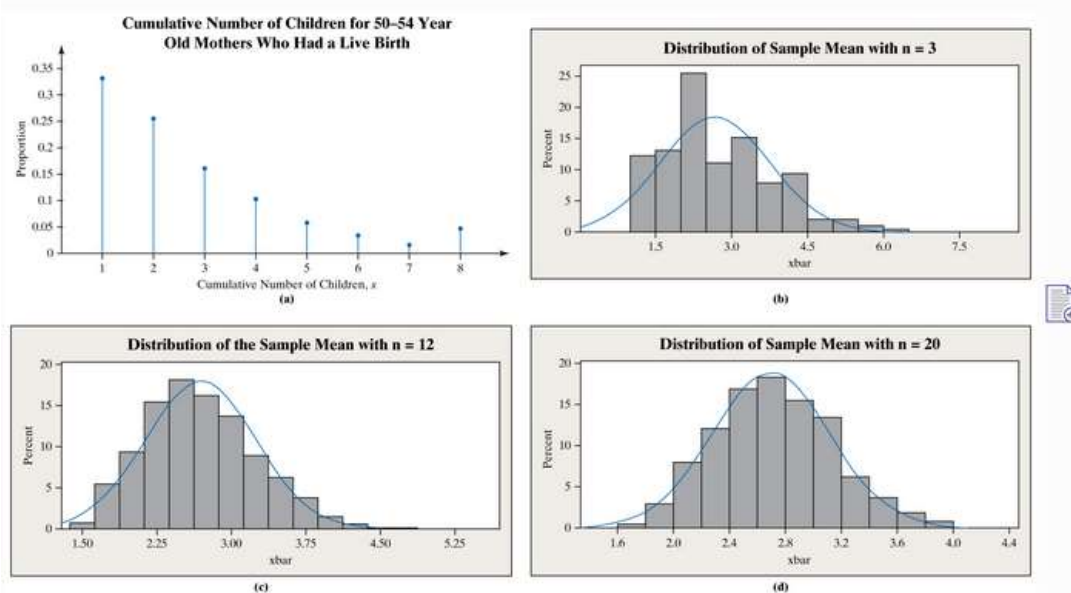


Figure 3

A Rule of Thumb for Invoking the Central Limit Theorem

Based on the previous discussion and on the graphs in [Figure 2](#) and [Figure 3](#), we draw the following important conclusions.

1. The shape of the distribution of the population from which the sample is drawn dictates the size of the sample required for the distribution of the sample mean to be normal.
2. The more skewed the distribution of the population is, the larger the sample size needed to invoke the Central Limit Theorem.

We will err on the side of caution and use the following rule of thumb.

If the distribution of the population is unknown or not normal, then the distribution of the sample mean is approximately normal provided that the sample size is greater than or equal to 30.

EXAMPLE 2 Weight Gain during Pregnancy

Problem

The mean weight gain during pregnancy is 30 pounds, with a standard deviation of 12.9 pounds. Weight gain during pregnancy is skewed right. An obstetrician obtains a random sample of 35 low-income patients and determines that their mean weight gain during pregnancy was 36.2 pounds. Does this result suggest anything unusual?

Approach

We want to know whether the sample mean obtained is unusual. Therefore, determine the likelihood of obtaining a sample mean of 36.2 pounds or higher. (If a 36.2 pound weight gain is unusual, certainly any weight gain above 36.2 pounds is also unusual.) Assume that the patients come from the population whose mean weight gain is 30 pounds. Verify the independence assumption. Use the normal model to obtain the probability because the sample size is large enough to use the Central Limit Theorem. Determine the area under the normal curve to the right of 36.2 pounds with

$$\mu_{\bar{x}} = 30 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.9}{\sqrt{35}}.$$

Solution

[Video Solution](#) [Technology Step-By-Step](#)



Results and Interpretation

If the population from which this sample is drawn has a mean weight gain of 30 pounds, the probability that a random sample of 35 women has a sample mean weight gain of 36.2 pounds (or more) is approximately 0.002 (by hand: 0.0023; technology: 0.0022). This means that about 2 samples in 1000 will result in a sample mean of 36.2 pounds or higher if the population mean is 30 pounds. We can conclude one of two things based on this result:

1. The mean weight gain for low-income patients is 30 pounds, and we happened to select women who, on average, gained more weight.
2. The mean weight gain for low-income patients is more than 30 pounds.

We are inclined to accept the second explanation over the first because our sample was obtained randomly. Therefore, the obstetrician should be concerned. Perhaps she should look at the diets and/or lifestyles of low-income patients while they are pregnant.

Finding Area Under the Normal Curve

1. From the HOME screen, press 2^{nd} VARS to access the DISTRibution menu.
2. Select 2:normalcdf{.
3. Enter the lower bound, upper bound, μ , and σ . For example, to find the area to the left of $x = 35$ under the normal curve with $\mu = 40$ and $\sigma = 10$, enter $-1E99$ for lower, 35 for upper, 40 for μ , and 10 for σ . Highlight Paste and hit ENTER. Hit ENTER a second time.

NOTE

When there is no lowerbound, enter $-1E99$. When there is no upperbound, enter $1E99$. The E shown is scientific notation; it is selected by pressing 2^{nd} then the comma button.

Finding Area Under the Normal Curve

1. Select Stat, highlight Calculators, and select Normal.
2. Enter the mean and the standard deviation.
 - If you want to compute $P(X \leq x)$ or $P(X \geq x)$, select Standard. Then select the \leq or \geq from the pull-down menu and enter the value of x . Click Compute.
 - If you want to compute $P(a \leq X \leq b)$, select Between. Then enter the values of a and b . Click Compute.

8.1.19-T

November 17, 2016 09:16 AM

8.1 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell Date: 11/17/16	Instructor: Matthew Nably Course: MTH 243: Introduction to Probability and Statistics	Assignment: 8.1 Interactive Assignment
---	--	---

Suppose the lengths of the pregnancies of a certain animal are approximately normally distributed with mean $\mu = 180$ days and standard deviation $\sigma = 17$ days. Complete parts (a) through (c).

(a) What is the probability that a randomly selected pregnancy lasts less than 174 days?

Notice the probability in this case is for a single observation. Therefore, use the population distribution to determine the probability.

What area should be determined to find $P(x < 174)$?

- A. The area under the normal curve between $x = 174$ and $x = 180$
- B. The area under the normal curve to the right of $x = 174$
- C. The area under the normal curve to the right of $x = 180$
- D. The area under the normal curve to the left of $x = 174$

The probability is the area to the left of $x = 174$ under the normal curve. Use technology to construct a normal curve with $\mu = 180$ and $\sigma = 17$, and find the area under the curve to the left of $x = 174$.

$P(x < 174) =$
(Round to four decimal places as needed.)



Therefore, the probability that a randomly selected pregnancy lasts less than 174 days is approximately 0.3621.

(b) What is the probability that a random sample of 13 pregnancies has a mean gestation period of less than 174 days?

Notice that the probability in this case is for the mean of a sample of the population. This means that the sampling distribution of the mean should be used and not the population distribution.

When a simple random sample of size n is drawn from a large population with mean μ and standard deviation σ , the sampling distribution of \bar{x} will have mean $\mu_{\bar{x}} = \mu$ and standard deviation given by the formula below.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

What is the mean of the sampling distribution of \bar{x} ?

$\mu_{\bar{x}} =$

Apply the definition for the standard deviation of the sampling distribution of \bar{x} for a sample size of 13.

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{17}{\sqrt{13}} \end{aligned}$$

Simplify to find the standard deviation of the sampling distribution of \bar{x} .

1 of 2

17-Nov-16 09:16 AM

8.1 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{17}{\sqrt{13}} \\ &= 4.714952 \end{aligned}$$

(Round to six decimal places as needed.)

The probability is the area to the left of $\bar{x} = 174$ under the normal curve. Use technology to construct a normal curve with $\mu_{\bar{x}} = 180$ and $\sigma_{\bar{x}} = 4.714952$, and find the area under the curve to the left of $\bar{x} = 174$.

$P(\bar{x} < 174) =$
(Round to four decimal places as needed.)



Therefore, the probability that the mean of a sample of 13 randomly selected pregnancies is less than 174 days is approximately 0.1016.

(c) What is the probability that a random sample of 34 pregnancies has a mean gestation period of less than 174 days?

Repeat the same calculations for a sample size of 34. The mean is again 180. Calculate the standard deviation of the sampling distribution.

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{17}{\sqrt{34}} \\ &= 2.915476 \end{aligned}$$

(Round to six decimal places as needed.)

The probability is the area to the left of $\bar{x} = 174$ under the normal curve. Use technology to construct a normal curve with $\mu_{\bar{x}} = 180$ and $\sigma_{\bar{x}} = 2.915476$, and find the area under the curve to the left of $\bar{x} = 174$.

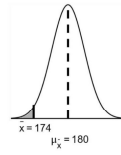
$P(\bar{x} < 174) =$
(Round to four decimal places as needed.)



the left of $\bar{x} = 174$.

$$P(\bar{x} < 174) = .0198$$

(Round to four decimal places as needed.)



Therefore, the probability that the mean of a sample of 34 randomly selected pregnancies is less than 174 days is approximately 0.0198.

8.1.31

November 17, 2016 10:23 AM

Section 8.1 Homework-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell Date: 11/17/16	Instructor: Matthew Nabity Course: MTH 243: Introduction to Probability and Statistics	Assignment: Section 8.1 Homework
---	---	---

The data in the table represent the ages of the winners of an award for the past five years. Use the data to answer questions (a) through (e).

50	33
44	34
33	

(a) Compute the population mean, μ .

$$\mu = \frac{50 + 33 + 44 + 34 + 33}{5} = 38.8$$

(b) Compute the mean for all ${}_5C_2 = 10$ samples with size $n = 2$.

Find the means by adding each pair and dividing by two.

Sample	Sample Mean	Sample	Sample Mean
50,33	41.5	33,34	33.5
50,44	47	33,33	33
50,34	42	44,34	39
50,33	41.5	44,33	38.5
33,44	38.5	34,33	33.5

(c) Construct a sampling distribution for the mean by listing the sample means and their corresponding probabilities.

The probability of each mean is the number of times it occurs divided by the total number of samples. Sort the means and identify all the unique values. The sorted sample means are 33, 33.5, 33.5, 38.5, 38.5, 39, 41.5, 41.5, 42, and 47. What are the unique values?

33,33.5,38.5,39,41.5,42,47

(Use a comma to separate answers as needed.)

The sorted sample means are 33, 33.5, 33.5, 38.5, 38.5, 39, 41.5, 41.5, 42, and 47. Count the number of times each mean occurs.

Sample Mean	Frequency	Probability	Sample Mean	Frequency	Probability
33	1		41.5	2	
33.5	2		42	1	
38.5	2		47	1	
39	1				

Then compute the probability for each mean. (Type an integer or decimal.)

Sample Mean	Frequency	Probability	Sample Mean	Frequency	Probability
33	1	.1	41.5	2	.2
33.5	2	.2	42	1	.1
38.5	2	.2	47	1	.1
39	1	.1			

(d) Compute the mean of the sampling distribution.

The sorted sample means are 33, 33.5, 33.5, 38.5, 38.5, 39, 41.5, 41.5, 42, and 47. Find the sum of all the sample means and divide by the number of samples, 10.

$$\mu_{\bar{x}} = \text{38.8}$$

(e) Compute the probability that the sample mean is within 2 years of the population mean age.

The sample means are 33, 33.5, 33.5, 38.5, 38.5, 39, 41.5, 41.5, 42, and 47. What unique values are within 2 years of the population mean age of 38.8?

(Use a comma to separate answers as needed.)

Add the probabilities for each of the means.

$$0.2 + 0.1 = \text{.3}$$

Therefore, the probability that a sample mean is within 2 years of the population mean age is 0.3.

YOU ANSWERED: 33,33.5,33.5,38.5,38.5,39,41.5,41.5,42,47

39, 41.5

8.1.33

November 17, 2016 10:19 AM

Section 8.1 Homework-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/17/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: Section 8.1 Homework

In the game of roulette, a wheel consists of 38 slots numbered 0, 00, 1, 2, ..., 36. To play the game, a metal ball is spun around the wheel and is allowed to fall into one of the numbered slots. If the number of the slot the ball falls into matches the number you selected, you win \$35; otherwise you lose \$1. Complete parts (a) through (g) below.

(a) Construct a probability distribution for the random variable X , the winnings of each spin.

There is an equal chance of the ball falling into any slot. Determine the number of slots that correspond to each winning amount and divide by the total number of slots. Do not overlook slots 0 and 00.

There are two possible amounts to win. Find the probabilities of winning either amount.

(Type an integer or decimal rounded to four decimal places as needed.)

x	$P(x)$
35	0.0263
-1	.9737

(b) Determine the mean and standard deviation of the random variable X . Round your results to the nearest penny.

There are 38 possible outcomes; 37 result in $x = -1$ and 1 results in $x = 35$. Use technology to find the mean and the standard deviation.

$\mu =$	-0.05
$\sigma =$	5.76

(c) Suppose that you play the game 80 times so that $n = 80$. Describe the sampling distribution of \bar{x} , the mean amount won per game.

Recall that the Central Limit Theorem states that regardless of the underlying population, the sampling distribution of \bar{x} becomes approximately normal as the sample size, n , increases.

The sampling distribution of \bar{x} is approximately normal.

Suppose that a simple random sample of size n is drawn from a population with mean μ and standard deviation σ . The

sampling distribution of \bar{x} has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

What are the mean and standard deviation of the sampling distribution of \bar{x} ?

$\mu_{\bar{x}} =$	-0.05
$\sigma_{\bar{x}} =$.64

(Type integers or decimals rounded to two decimal places as needed.)

(d) What is the probability of being ahead after playing the game 80 times? That is, what is the probability that the sample mean is greater than 0 for $n = 80$?

To calculate the probability, generate a normal curve with $\mu_{\bar{x}} = -0.05$ for the mean and $\sigma_{\bar{x}} = \frac{5.76}{\sqrt{80}}$ for the standard deviation.

Use technology to find the area to the right of 0.

$$P(\bar{x} > 0) = .4687 \quad (\text{Round to four decimal places as needed.})$$

(e) What is the probability of being ahead after playing the game 240 times?

To calculate the probability, generate a normal curve with $\mu_{\bar{x}} = -0.05$ for the mean and $\sigma_{\bar{x}} = \frac{5.76}{\sqrt{240}}$ for the standard deviation.

Use technology to find the area to the right of 0.

$$P(\bar{x} > 0) = \text{.4465} \quad (\text{Round to four decimal places as needed.})$$

(f) What is the probability of being ahead after playing the game 800 times?

To calculate the probability, generate a normal curve with $\mu_{\bar{x}} = -0.05$ for the mean and $\sigma_{\bar{x}} = \frac{5.76}{\sqrt{800}}$ for the standard deviation.

Use technology to find the area to the right of 0.

$$P(\bar{x} > 0) = \text{.4030} \quad (\text{Round to four decimal places as needed.})$$

(g) Compare the results of parts (d) through (f). What lesson does this teach you?

Examine the effect of n on the different probabilities computed to determine what lesson can be learned. Recall that x represents the amount won per game.

YOU ANSWERED: .02633

19.5

11.1131

8.2 Distribution of the Sample Mean

November 18, 2016 05:30 PM

8.2 Objective 1 Describe the Sampling Distribution of a Sample Proportion

November 18, 2016 05:35 PM

The Sample Proportion

Suppose we want to determine the proportion of households in a 100-house homeowners association that favor an increase in the annual assessments to pay for neighborhood improvements. We could survey all households to learn which are in favor of higher assessments. If 65 of the 100 households favor the higher assessment, the population proportion, p , of households in favor of a higher assessment is

$$p = \frac{65}{100} = 0.65$$

Of course, gaining access to all the individuals in a population is rare, so we usually obtain estimates of population parameters such as p .

DEFINITION

Suppose that a random sample of size n is obtained from a population in which each individual either does or does not have a certain characteristic. The **sample proportion**, denoted \hat{p} (read “ p -hat”), is given by

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the specified characteristic. The sample proportion, \hat{p} , is a statistic that estimates the population proportion, p .

NOTE

For those who studied binomial probabilities, x can be thought of as the number of successes in n trials of a binomial experiment.

EXAMPLE 1 Computing a Sample Proportion

Problem

The Harris Poll conducted a survey of 1200 adult Americans who vacation during the summer and asked whether the individuals planned to work while on summer vacation. Of those surveyed, 552 indicated that they planned to work while on vacation. Find the sample proportion of individuals surveyed who planned to work while on summer vacation.

Approach

Use the formula $\hat{p} = \frac{x}{n}$, where x is the number of individuals who plan to work while on summer vacation and n is the sample size.

Solution

Substituting $x = 552$ and $n = 1200$ into $\hat{p} = \frac{x}{n}$, we find that $\hat{p} = \frac{552}{1200} = 0.46$.

The Harris Poll estimates that 0.46, or 46%, of adult Americans plan to work while on summer vacation.

The Sample Proportion is a Random Variable

A second survey of 1200 adult Americans would likely have a different estimate of the proportion of adult Americans who plan to work on summer vacation because different individuals would be in the sample. Because the value of \hat{p} varies from sample to sample, it is a random variable and has a probability distribution.

To get a sense of the shape, center, and spread of the sampling distribution of \hat{p} , we could repeat the exercise of obtaining simple random samples of 1200 adult Americans over and over. This would lead to a list of sample proportions. A histogram of the sample proportions will give us a feel for the shape of the distribution of the sample proportion. The mean of the sample proportions will give us an idea of the center of the distribution, and the standard deviation of the sample proportions will give us an idea of the spread of the distribution.

The display below summarizes the sampling distribution of the sample proportion, \hat{p} .

Sampling Distribution of \hat{p}

For a simple random sample of size n with a population proportion p ,

- The shape of the sampling distribution of \hat{p} is approximately normal provided $np(1-p) \geq 10$.
- The mean of the sampling distribution of \hat{p} is $\mu_{\hat{p}} = p$.
- The standard deviation of the sampling distribution of \hat{p} is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

NOTE

The sample size, n , can be no more than 5% of the population size, N . That is, $n \leq 0.05N$.

EXAMPLE 2 Describing the Sampling Distribution of the Sample Proportion

Problem

Based on a study conducted by the Gallup Organization, 76% of Americans believe

Video Solution



EXAMPLE 2 Describing the Sampling Distribution of the Sample Proportion

Problem

Based on a study conducted by the Gallup Organization, 76% of Americans believe that the state of moral values in the United States is getting worse. Suppose we obtain a simple random sample of $n = 60$ Americans and determine which of them believe that the state of moral values in the United States is getting worse. Describe the sampling distribution of the sample proportion for Americans with this belief.

Video Solution



Approach

Recall, *describe the distribution* means to identify the shape, center (mean), and spread (standard deviation) of the distribution. If the sample size is less than 5% of the population size and $np(1 - p)$ is at least 10, the sampling distribution of \hat{p} is approximately normal, with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Solution

The United States has over 310 million people, so the sample of $n = 60$ is less than 5% of the population size. Also, $np(1 - p) = 60(0.76)(1 - 0.76) = 10.944 \geq 10$. The distribution of \hat{p} is approximately normal, with mean $\mu_{\hat{p}} = p = 0.76$ and standard deviation

$$\begin{aligned}\sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.76(1-0.76)}{60}} \\ &= 0.055\end{aligned}$$

8.2 Objective 2 Compute Probabilities of a Sample Mean

November 18, 2016 05:36 PM

EXAMPLE 3 Computing Probabilities of a Sample Proportion

Problem

According to the National Center for Health Statistics, 15% of all Americans have hearing trouble.

- In a random sample of 120 Americans, what is the probability that at most 12% have hearing trouble?
- Suppose that a random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. What might you conclude?

Approach

First, determine whether the sampling distribution is approximately normal by verifying that the sample size is less than 5% of the population size and that $np(1 - p) \geq 10$. Then use the normal model to determine the probabilities.

Solution

Video Solution Technology Step-By-Step



Results and Interpretation

(a) The probability that a random sample of $n = 120$ Americans results in at most 12% having hearing trouble is approximately 0.18 (by hand: 0.1788; technology: 0.1787). This means that about 18 out of 100 random samples of size 120 will result in at most 12% having hearing trouble if the population proportion of Americans with hearing trouble is 0.15.

(b) The probability that a random sample of $n = 120$ Americans results in 26 (or more) having hearing trouble is about 0.02 (by hand: 0.0197; technology: 0.0199). About 2 samples in 100 will result in a sample proportion of 0.217 or more from a population whose proportion is 0.15. We obtained a result that should only happen about 2 times in 100, so the results obtained are unusual. We could make one of two conclusions:

- The population proportion of Americans with hearing trouble who regularly listen to music using headphones is 0.15, and we just happen to randomly select a higher proportion that have hearing trouble.
- The population proportion of Americans with hearing trouble who regularly listen to music using headphones is more than 0.15.

The second conclusion is more reasonable. We conclude that the proportion of Americans who regularly listen using headphones who have hearing trouble is higher than the general population.

Finding Area Under a Normal Curve

1. From the HOME screen, press 2^{nd} VARS to access the DISTRibution menu.
2. Select 2:normalcdf.
3. Enter the lower bound, upper bound, μ , and σ . For example, to find the area to the left of $x = 35$ under the normal curve with $\mu = 40$ and $\sigma = 10$, enter $-1\text{E}99$ for lower, 35 for upper, 40 for μ , and 10 for σ . Highlight Paste and hit ENTER. Hit ENTER a second time.

NOTE

When there is no lower bound, enter $-1\text{E}99$. When there is no upper bound, enter $1\text{E}99$. The E shown is scientific notation; it is selected by pressing 2^{nd} then press the comma key.

Finding Area Under a Normal Curve

1. Select Stat, highlight Calculators, and select Normal.
2. Enter the mean and the standard deviation.
 - If you want to compute $P(X \leq x)$ or $P(X \geq x)$, select Standard. Then select the \leq or \geq from the pull-down menu and enter the value of x . Click Compute.
 - If you want to compute $P(a \leq X \leq b)$, select Between. Then enter the values of a and b . Click Compute.

8.2.16

November 18, 2016 05:47 PM

8.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/18/16

Instructor: Matthew Naby
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 8.2 Interactive Assignment

According to a study conducted by a statistical organization, the proportion of Americans who are satisfied with the way things are going in their lives is 0.84. Suppose that a random sample of 107 Americans is obtained. Complete parts (a) through (c).

(a) Describe the sampling distribution of \hat{p} .

Let p be the proportion of Americans who are satisfied with the way things are going in their lives, n be the size of the sample. For a simple random sample of size n with a population proportion p , the shape of the sampling distribution is approximately normal provided $np(1-p) \geq 10$. The mean of the sampling distribution is $\mu_{\hat{p}} = p$.

The standard deviation of the sampling distribution of \hat{p} is given by the following formula.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Determine n and p .

$$n = 107$$
$$p = 0.84$$

Calculate $\sigma_{\hat{p}}$.

$$\begin{aligned}\sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.84 \cdot (1-0.84)}{107}} \\ &= 0.035441\end{aligned}$$

Find $np(1-p)$.

$$\begin{aligned}np(1-p) &= 107(0.84)(1-0.84) \\ &\approx 14.381\end{aligned}$$

Thus, the distribution of \hat{p} can be approximated by a normal distribution with $\mu_{\hat{p}} = 0.84$ and $\sigma_{\hat{p}} = 0.035441$.

(b) Using the distribution from part (a), what is the probability that at least 88 Americans in the sample are satisfied with their lives?

Let \hat{p} be the sample proportion, x be the number of successes, and n be the size of the sample. The value of \hat{p} is given by the formula below.

$$\hat{p} = \frac{x}{n}$$

Let x_1 be the value from part (b). Then the problem is to find probability $P\left(\hat{p} > \frac{x_1}{n}\right)$, where n is the size of the sample. It equals the area under the corresponding normal curve to the right of point $\hat{p}_1 = \frac{x_1}{n}$.

Calculate \hat{p}_1 .

$$\begin{aligned}\hat{p}_1 &= \frac{x_1}{n} \\ &= \frac{88}{107} \\ &\approx 0.82\end{aligned}$$

Convert \hat{p}_1 to the corresponding z-score z_1 .

$$\begin{aligned}z_1 &= \frac{\hat{p}_1 - \mu_p}{\sigma_p} \\ &= \frac{0.82 - 0.84}{0.035441} \\ &\approx -0.56\end{aligned}$$

Therefore, since $np(1-p) \geq 10$ the probability $P(z > z_1)$ equals the area under the standard normal curve to the right of $z_1 = -0.56$. Calculate this probability.

$$\text{Area} \approx 0.7123$$

(c) Using the distribution from part (a), what is the probability that 82 or fewer Americans in the sample are satisfied with their lives?

Let x_2 be the value from part (c). Then the problem is to find probability $P\left(\hat{p} < \frac{x_2}{n}\right)$, where n is the size of the sample. It equals the area under the corresponding normal curve to the left of point $\hat{p}_2 = \frac{x_2}{n}$.

Calculate \hat{p}_2 .

$$\begin{aligned}\hat{p}_2 &= \frac{x_2}{n} \\ &= \frac{82}{107} \\ &\approx 0.77\end{aligned}$$

Convert \hat{p}_2 to the corresponding z-score z_2 .

$$\begin{aligned}z_2 &= \frac{\hat{p}_2 - \mu_p}{\sigma_p} \\ &= \frac{0.77 - 0.84}{0.035441} \\ &= -1.98\end{aligned}$$

Therefore, since $np(1-p) \geq 10$ the probability $P(z < z_2)$ equals the area under the standard normal curve to the left of $z_2 = -1.98$. Calculate this probability.

$$\text{Area} \approx 0.0239$$

8.2.18-T

November 18, 2016 06:14 PM

Section 8.2 Homework-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/18/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: Section 8.2 Homework

According to a survey in a country, 28% of adults do not have any credit cards. Suppose a simple random sample of 400 adults is obtained.

(a) Describe the sampling distribution of \hat{p} , the sample proportion of adults who do not have a credit card.

For a simple random sample size n such that $n \leq 0.05N$, the shape of the sampling distribution of \hat{p} is approximately normal provided $np(1-p) \geq 10$.

The sample size 400 is 5% of 8,000. It is safe to assume there are more than 8,000 adults in the country so the sample size is less than 5% of the population size.

Evaluate $np(1-p)$.

$$\begin{aligned} np(1-p) &= (400)(0.28)(1-0.28) \\ &= \boxed{80.64} \quad (\text{Round to two decimal places as needed.}) \end{aligned}$$

Choose the phrase that best describes the shape of the sampling distribution of \hat{p} below.

- A. Approximately normal because $n \leq 0.05N$ and $np(1-p) \geq 10$.
- B. Not normal because $n \leq 0.05N$ and $np(1-p) < 10$.
- C. Not normal because $n \leq 0.05N$ and $np(1-p) \geq 10$.
- D. Approximately normal because $n \leq 0.05N$ and $np(1-p) < 10$.

The mean of the sampling distribution of \hat{p} is the same as the population proportion.

$$\mu_{\hat{p}} = \boxed{.28}$$

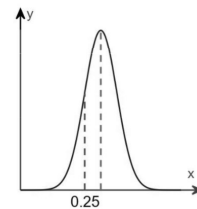
Determine the standard deviation of the sampling distribution of \hat{p} .

$$\begin{aligned} \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.28(1-0.28)}{400}} \\ &= \boxed{0.022} \quad (\text{Round to three decimal places as needed.}) \end{aligned}$$

(b) In a random sample of 400 adults, what is the probability that less than 25% have no credit cards?

The normal curve with $\hat{p} = 0.25$ is shown to the right. Which area corresponds to $P(\hat{p} < 0.25)$?

- The area to the left of $\hat{p} = 0.25$.
- The area to the right of $\hat{p} = 0.25$.



Use technology to find the area to the left of $\hat{p} = 0.25$. Use the Tech Help button for further assistance.

$$P(\hat{p} < 0.25) = \boxed{0.0863} \quad (\text{Round to four decimal places as needed.})$$

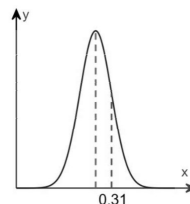
(c) Would it be unusual if a random sample of 400 adults results in 124 or more having no credit cards?

To determine the probability that 124 or more adults in a sample of 400 have no credit cards, first find the sample proportion.

$$\hat{p} = \frac{124}{400} = \text{.31} \quad (\text{Round to two decimal places as needed.})$$

The normal curve with $\hat{p} = 0.31$ is shown to the right. Which area corresponds to $P(\hat{p} \geq 0.31)$?

- The area to the right of $\hat{p} = 0.31$.
 The area to the left of $\hat{p} = 0.31$.



Use technology to find the area to the right of $\hat{p} = 0.31$. Use the Tech Help button for further assistance.

$$P(\hat{p} \geq 0.31) = \text{0.0863} \quad (\text{Round to four decimal places as needed.})$$

A result is considered unusual if the probability of such a sample being selected is less than 5%. Is this result unusual?

- A. The result is not unusual because the probability that $\hat{p} \geq 0.31$ is greater than 5%.
 B. The result is not unusual because the probability that $\hat{p} \geq 0.31$ is less than 5%.
 C. The result is unusual because the probability that $\hat{p} \geq 0.31$ is less than 5%.
 D. The result is unusual because the probability that $\hat{p} \geq 0.31$ is greater than 5%.

YOU ANSWERED: 0.0224

8.2.23

November 18, 2016 06:20 PM

Section 8.2 Homework-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/18/16

Instructor: Matthew Naby
Course: MTH 243: Introduction to
Probability and Statistics

Assignment: Section 8.2 Homework

A researcher studying public opinion of proposed Social Security changes obtains a simple random sample of 25 adult Americans and asks them whether or not they support the proposed changes. To say that the distribution of the sample proportion of adults who respond yes, is approximately normal, how many more adult Americans does the researcher need to sample in the following cases?

- (a) 15% of all adult Americans support the changes
(b) 20% of all adult Americans support the changes

The sampling distribution is approximately normal if $np(1-p) \geq 10$ is correct, where p is the percent of Americans who support the changes. The new sample size n must suit this condition. Then find the difference between n and the initial size of the sample.

- (a) Check this condition for the distribution concerned.

$$np(1-p) = 10$$

Obtain the following minimal required size of the sample.

$$n = \frac{10}{p(1-p)}$$

Calculate n .

$$\begin{aligned} n &= \frac{10}{p(1-p)} \\ &= \frac{10}{0.15(1-0.15)} \\ &= 79 \end{aligned}$$

(Round up to the nearest integer.)

Let $n_0 = 25$ be the initial size of the sample. The researcher must ask $n - n_0$ more American adults.

$$\begin{aligned} n - n_0 &= 79 - 25 \\ &= 54 \end{aligned}$$

The researcher must ask 54 more American adults.

- (b) Using $np(1-p) = 10$, obtain the following minimal required size of the sample.

$$n = \frac{10}{p(1-p)}$$

Calculate n .

$$\begin{aligned} n &= \frac{10}{p(1-p)} \\ &= \frac{10}{0.20(1-0.20)} \\ &= 63 \end{aligned}$$

(Round up to the nearest integer.)

Let $n_0 = 25$ be the initial size of the sample. The researcher must ask $n - n_0$ more American adults.

$$\begin{aligned}n - n_0 &= 63 - 25 \\ &= \boxed{38}\end{aligned}$$

The researcher must ask 38 more American adults.

9.1 Estimating a Population Proportion

November 24, 2016 07:26 AM

9.1 Objective 1 Obtain a Point Estimate for the Population Proportion

November 24, 2016 07:36 AM

Suppose we want to estimate the proportion of adult Americans who believe that the amount they pay in federal income taxes is fair. It is unreasonable to expect that we could survey every adult American. Instead, we use a sample of adult Americans to arrive at an estimate of the proportion. We call this estimate a *point estimate*.

DEFINITION

A **point estimate** is the value of a statistic that estimates the value of a parameter.

EXAMPLE 1 Obtaining a Point Estimate of a Population Proportion

Problem

The Gallup Organization conducted a poll in April 2013 in which a simple random sample of 1016 Americans aged 18 and older were asked, "Do you regard the income tax that you will have to pay this year as fair?" Of the 1016 adult Americans surveyed, 558 said yes. Obtain a point estimate for the proportion of Americans aged 18 and older who believe that the amount of income tax they pay is fair.

Approach

The point estimate of the population proportion is $\hat{p} = \frac{x}{n}$, where $x = 558$ and $n = 1016$.

Solution

Substituting into the formula, we get $\hat{p} = \frac{x}{n} = \frac{558}{1016} = 0.549 = 54.9\%$.

We estimate that 54.9% of Americans aged 18 and older believe that the amount of income tax they pay is fair.

9.1 Objective 2 - Construct and Interpret a Confidence Interval for the Population Proportion

November 24, 2016 07:36 AM

Based on the point estimate of Example 1, can we conclude that a majority (more than 50%) of the U.S. adult population believes that the amount of income tax they pay is fair? Or is it possible that less than a majority of adult Americans believe that their income tax is fair and we just happened to sample folks who do believe that they pay a fair amount in taxes?

After all, statistics such as \hat{p} vary from sample to sample. So a different random sample of adult Americans might result in a different point estimate of the population proportion, such as $\hat{p} = 0.498$.

If the method used to select the adult Americans was done appropriately, both point estimates would be good guesses of the population proportion. Due to variability in the sample proportion, we report a range (or *interval*) of values, including a measure of the likelihood that the interval includes the unknown population proportion.

To understand the idea of this interval, consider the following situation.

Suppose you were asked to guess the proportion of students on your campus who use Facebook. If a survey of 80 students results in 60 who use Facebook, then $\hat{p} = 0.75$. From this, you might guess that the proportion of *all* students on your campus who use Facebook is 0.75, but because you did not survey every student on campus, your estimate may be incorrect.

To account for this error, you could adjust your guess by stating that the proportion of students on your campus who use Facebook is 0.75, give or take 0.05 (the *margin of error*). Mathematically, we write this as 0.75 ± 0.05 . If asked how confident you are that the proportion is between 0.70 and 0.80, you might respond, "I am 90% confident that the proportion of students on my campus who use Facebook is between 0.70 and 0.80."

If you want an interval for which your confidence increases to, say, 95%, what do you think will happen to the interval? More confidence that the interval will capture the unknown population proportion requires that your interval increase to, say, 0.65 to 0.85.

In statistics, we construct an interval for a population parameter based on a guess (the point estimate) along with a level of confidence. The level of confidence plays a role in the width of the interval. See Figure 1.

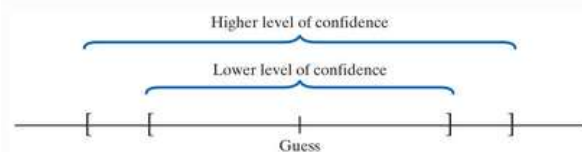


Figure 1

DEFINITION

- A **confidence interval** for an unknown parameter consists of an interval of numbers based on a point estimate.
- The **level of confidence** represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained. The level of confidence is denoted $(1 - \alpha) \cdot 100\%$.

For example, a 95% level of confidence ($\alpha = 0.05$) implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, then we will expect 95 of the intervals to include the parameter and 5 to not include the parameter.

Confidence interval estimates for the population proportion are of the form

$$\text{point estimate} \pm \text{margin of error}$$

For example, in the poll in which Gallup was estimating the proportion of adult Americans who believe that the amount of income tax they pay is fair, the point estimate is $\hat{p} = 0.549$. From the report, Gallup indicated that the margin of error was 0.04. Gallup also reported the level of confidence to be 95%. So Gallup is 95% confident that the proportion of adult Americans who believe that the amount they pay in income tax is fair is between 0.509 ($= 0.549 - 0.04$) and 0.589 ($= 0.549 + 0.04$).

Here are some unanswered questions at this point in the discussion:

- Why does the level of confidence represent the expected proportion of intervals that contain the parameter if a large number of different samples is obtained?
- How is the margin of error determined?

To help answer these questions, let's review what we know about the model that describes the sampling distribution of the sample proportion, \hat{p} .

The Sampling Distribution of the Sample Proportion

- The shape of the distribution of all possible sample proportions is approximately normal provided $np(1 - p) \geq 10$ and the sample size is no more than 5% of the population size. That is, $n \leq 0.05N$.
- The mean of the distribution of the sample proportions equals the population proportion. That is, $\mu_{\hat{p}} = p$.
- The standard deviation of the distribution of the sample proportion (the standard error) is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Because the distribution of the sample proportion is approximately normal, we know that 95% of all sample proportions will lie within 1.96 standard errors of the population proportion, p , and 2.5% of the sample proportions will lie in each tail.

See Figure 2.

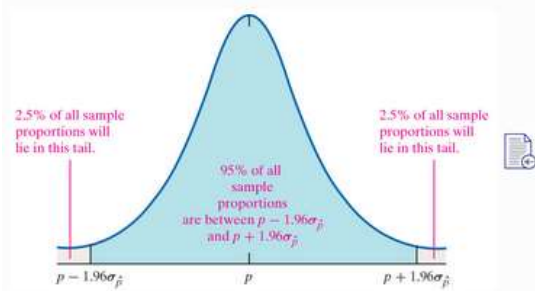


Figure 2

The 1.96 comes from the fact that $z_{0.025}$ is the z -value such that 2.5% of the area under the standard normal curve is to its right. Recall that $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.

From Figure 2, we see that 95% of all sample proportions satisfy the inequality

$$p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}$$

parameter - 1.96 standard error < point estimate < parameter + 1.96 standard error

With a little algebraic manipulation, we can rewrite this inequality with p in the middle and obtain the following:

$$\hat{p} - 1.96\sigma_{\hat{p}} < p < \hat{p} + 1.96\sigma_{\hat{p}}$$

point estimate - 1.96 standard error < parameter < point estimate + 1.96 standard error

This inequality states that 95% of all sample proportions will result in confidence interval estimates that contain the population proportion, while 5% of all sample proportions will result in confidence interval estimates that do not contain the population proportion. It is common to write the 95% confidence interval as

$$\hat{p} \pm 1.96\sigma_{\hat{p}}$$

point estimate \pm 1.96 standard error
point estimate \pm margin of error

So the margin of error for a 95% confidence interval for the population proportion is $1.96\sigma_{\hat{p}}$. This determines the width of the interval.

To visually illustrate the idea of a confidence interval, consider the sampling distribution of \hat{p} shown in Figure 3. For the sample proportion \hat{p}_1 , the 95% confidence interval includes the population proportion p . For the sample proportion \hat{p}_2 (shown in blue), the 95% confidence interval does not include the population proportion. What is the difference between the two sample proportions? The sample proportion \hat{p}_1 is within 1.96 standard errors of the population proportion, while the sample proportion \hat{p}_2 is more than 1.96 standard errors from the population proportion.

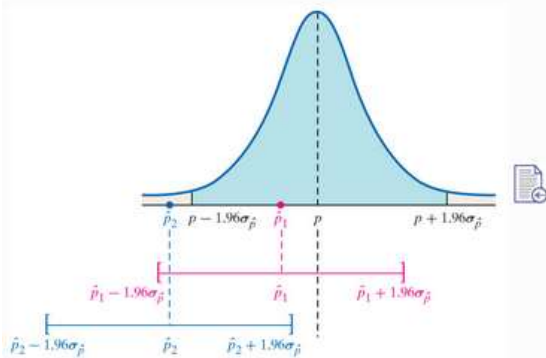


Figure 3

We can express this as follows:

- For a 95% confidence interval, any sample proportion that lies within 1.96 standard errors of the population proportion will result in a confidence interval that includes p . This will happen in 95% of all possible samples.
- Any sample proportion that is more than 1.96 standard errors from the population proportion will result in a confidence interval that does not contain p . This will happen in 5% of all possible samples (those sample proportions in the tails of the distribution).

Whether a confidence interval contains the population parameter depends solely on the value of the sample statistic.

Any sample statistic that is in the tails of the sampling distribution will result in a confidence interval that does not include the population parameter.

Reinforcing the Meaning of Level of Confidence

Be sure you understand the meaning of level of confidence in a 95% confidence interval.



A 95% confidence interval does not mean that there is a 95% probability that the interval contains the parameter (such as p or μ). The 95% in a 95% confidence interval represents the proportion of all samples that will result in intervals that include the population proportion.

In practice, we construct only one confidence interval based on one sample. We do not know whether the sample results in a confidence interval that includes the parameter, but we do know that if we construct a 95% confidence interval, it will include the unknown parameter 95% of the time.

Constructing any $(1 - \alpha) \cdot 100\%$ Confidence Interval

We need a method for constructing any $(1 - \alpha) \cdot 100\%$ confidence interval. When $\alpha = 0.05$, we are constructing a 95% confidence interval.

We generalize

$$p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}$$

parameter - 1.96 standard error < point estimate < parameter + 1.96 standard error

by first noting that $(1 - \alpha) \cdot 100\%$ of all sample proportions are in the interval as shown in Figure 4.

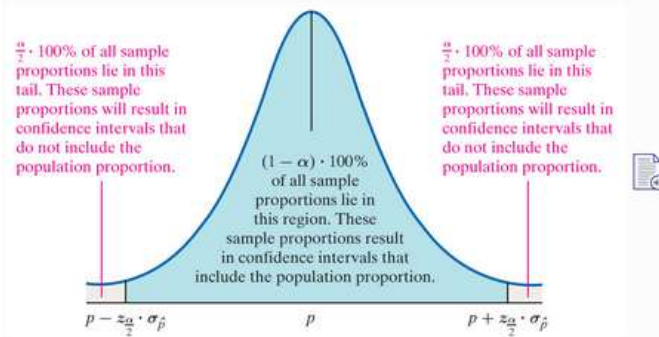


Figure 4

We rewrite this inequality with p in the middle and obtain

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

So $(1 - \alpha) \cdot 100\%$ of all sample proportions will result in confidence intervals that contain the population proportion. The sample proportions that are in the tails of the distribution in Figure 4 will not result in confidence intervals that contain the population proportion.

The value $z_{\frac{\alpha}{2}}$ is called the **critical value** of the distribution. It represents the number of standard deviations the sample statistic can be from the parameter and still result in an interval that includes the parameter.

Table 1 shows some of the common critical values used in the construction of confidence intervals. Notice that higher levels of confidence correspond to higher critical values. After all, if your level of confidence that the interval includes the unknown parameter increases, the width of your interval (through the margin of error) should increase.

TABLE 1

Level of Confidence, $(1 - \alpha) \cdot 100\%$	Area in Each tail, $\frac{\alpha}{2}$	Critical Value, $z_{\frac{\alpha}{2}}$
90%	0.05	1.645
95%	0.025	1.96
99%	0.005	2.575

Interpretation of a Confidence Interval

A $(1 - \alpha) \cdot 100\%$ confidence interval indicates that $(1 - \alpha) \cdot 100\%$ of all simple random samples of size n from the population whose parameter is unknown will result in an interval that contains the parameter.

IN
OTHER
WORDS



For example, a 90% confidence interval for a parameter suggests that 90% of all possible samples will result in an interval that includes the unknown parameter and 10% of the samples will result in an interval that does not capture the parameter.

EXAMPLE 2 Interpreting a Confidence Interval

Problem

The Gallup Organization conducted a poll in April 2013 in which a simple random sample of 1016 Americans aged 18 and older were asked, "Do you regard the income tax that you will have to pay this year as fair?" We learned from Example 1 that the proportion of those surveyed who responded yes was 0.549. Gallup reported its "survey methodology" as follows:

Results are based on telephone interviews with a random sample of 1016 national adults, aged 18 and older. For results based on the total sample of national adults, one can say with 95% confidence that margin of sampling error is 4 percentage points.

Determine and interpret the confidence interval for the proportion of Americans aged 18 and older who believe the amount of federal income tax they have to pay is fair.

Approach

Confidence intervals for a proportion are of the form point estimate \pm margin of error. So add and subtract the margin of error from the point estimate to obtain the confidence interval. Interpret the confidence interval, "We are 95% confident that the proportion of Americans aged 18 and older who believe that the income tax they will have to pay this year is fair is between *lower bound and upper bound*."

Solution

The point estimate is 0.549, and the margin of error is 0.04. The confidence interval is 0.549 ± 0.04 . Therefore, the lower bound of the confidence interval is $0.549 - 0.04 = 0.509$ and the upper bound of the confidence interval is $0.549 + 0.04 = 0.589$. We are 95% confident that the proportion of Americans aged 18 and older who believe that the income tax they will have to pay this year is fair is between 0.509 and 0.589.

A Word of Caution about Interpreting the Level of Confidence

An extremely important point is that the level of confidence refers to the confidence in the *method*, not in the specific interval. A 90% confidence interval means that the method "works" (that is, the interval includes the unknown parameter) for 90% of all samples. We do not know whether the sample statistic we obtained is one of the 90% with an interval that includes the parameter or one of the 10% whose interval does not include the parameter.

A 90% level of confidence does *not tell us* that there is a 90% probability that the parameter lies between the lower and upper bound.

We are now prepared to present a method for constructing a confidence interval about the population proportion, p .

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for a Population Proportion

Suppose that a simple random sample of size n is taken from a population or the data are the result of a randomized experiment. A $(1 - \alpha) \cdot 100\%$ confidence interval for p is given by the following quantities:

$$\text{Lower bound : } \hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{Upper bound : } \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

NOTE

It must be the case that $n\hat{p}(1 - \hat{p}) \geq 10$ and $n \leq 0.05N$ to construct this interval. Use \hat{p} in place of p in the standard deviation. This is because p is unknown, and \hat{p} is the best point estimate of p .

EXAMPLE 3 Constructing a Confidence Interval for a Population Proportion

Problem

In the Parent–Teen Cell Phone Survey conducted by Princeton Survey Research Associates International, 800 randomly sampled 16- to 17-year-olds living in the United States were asked whether they have ever used their cell phone to text while driving. Of the 800 teenagers surveyed, 272 indicated that they text while driving. Obtain a 95% confidence interval for the proportion of 16- to 17-year-olds who text while driving.

Approach

It is important to verify the requirements for constructing the confidence interval first. It must be the case that $n\hat{p}(1 - \hat{p}) > 10$ and the sample size is no more than 5% of the population size ($n \leq 0.05N$). Then, construct the confidence interval either by hand or using technology.

Solution

Video Solution Technology Step-by-Step



Interpretation

We are 95% confident that the proportion of 16- to 17-year olds who text while driving is between 0.307 and 0.373.

It is important to remember the correct interpretation of a confidence interval. The statement “95% confident” means that if 1000 samples of size 800 were taken, we would expect 950 of the intervals to contain the parameter p , while 50 will not. Unfortunately, we cannot know whether the interval we constructed is one of the 950 intervals that contains p or one of the 50 that does not contain p .

NOTE

We report the interval as *Lower bound: 0.307; Upper bound: 0.373*. Some texts will use interval notation, as in $(0.307, 0.373)$.

Obtaining a Confidence Interval for a Population Proportion

1. Press STAT, highlight TESTS, and select A: 1-propZInt....
2. Enter the values of x and n .
3. Enter the confidence level following C-Level:
4. Highlight Calculate and press ENTER.

Obtaining a Confidence Interval for a Population Proportion

1. If you have raw data, enter it into the spreadsheet. Name the column variable.
2. Select Stat, highlight Proportion Stats, select One sample, and then choose either **with data** or **with summary**.
3. If you chose **with data**, select the column that has the observations and choose which outcome represents a success. If you chose **with summary**, enter the number of successes and the number of trials. Choose the confidence interval radio button. Enter the level of confidence. Leave the method as Standard-Wald. Click Calculate!.

The Effect of Level of Confidence on the Margin of Error

The width of the interval is determined by the margin of error.

DEFINITION

The margin of error, E , in a $(1 - \alpha) \cdot 100\%$ confidence interval for a population proportion is given by

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

EXAMPLE 4 The Role of the Level of Confidence in the Margin of Error

Problem

In the Parent–Teen Cell Phone Survey conducted by Princeton Survey Research Associates International, 800 randomly sampled 16- to 17-year-olds living in the United States were asked whether they have ever used their cell phone to text while driving. Of the 800 teenagers surveyed, 272 indicated that they text while driving. From the last example, we concluded that we are 95% confident that the proportion of 16- to 17-year olds who text while driving is between 0.307 and 0.373. Determine the effect on the margin of error by increasing the level of confidence from 95% to 99%.

Approach

We would expect the margin of error to increase with a larger level of confidence. Construct the confidence interval either by hand or using technology.

Solution

Video Solution Technology Step-by-Step



Result

The margin of error for the 95% confidence interval found in Example 3 is 0.033, and the margin of error for the 99% confidence interval is 0.043. So increasing the level of confidence increases the margin of error, resulting in a wider confidence interval.

IN
OTHER
WORDS



The Effect of Sample Size on the Margin of Error

We know that larger sample sizes produce more precise estimates (the [Law of Large Numbers](#)). Given that the margin of error is

$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, we can see that increasing the sample size n decreases the standard error, so the margin of error decreases. Therefore, larger sample sizes will result in narrower confidence intervals.

To illustrate the idea, suppose the [survey](#) conducted in Example 3 resulted in the same sample proportion of 16- to 17-year-old teenagers who text while driving, but the sample size is only 200. The margin of error would be

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{0.34(1-0.34)}{200}} = 0.066$$

The margin of error in Example 3 was 0.033. So a sample size that is one-fourth the original size causes the margin of error to double. Put another way, if the sample size is quadrupled, the margin of error will be cut in half.

Law of Large Numbers

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.

In estimating a parameter, this law suggests that as the sample size increases, the statistic gets closer to the parameter.

9.1 Objective 3 Determine the Sample Size Necessary for Estimating a Population Proportion within a Specified Margin of Error

November 24, 2016 08:10 AM

Sample Size Needed for Estimating the Population Proportion

- The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval for p with a margin of error E is given by

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$$

(rounded up to the next integer) where \hat{p} is a prior estimate of p .

- If a prior estimate of p is unavailable, the sample size required is

$$n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$$

rounded up to the next integer. The margin of error should always be expressed as a decimal when using these formulas.

EXAMPLE 5 Determining Sample Size

Problem

An economist wants to know if the proportion of the U.S. population who commutes to work via car-pooling is on the rise. What size sample should be obtained if the economist wants an estimate within 2 percentage points of the true proportion with 90% confidence?

Video
Solution



Technology
Step-By-Step



Part A

Part B

Assume that the economist uses the estimate of 10% obtained from the American Community Survey.

Approach

Since $1 - \alpha = 0.9$, we know that $\alpha = 0.10$. Use $E = 0.02$, $z_{\frac{\alpha}{2}} = z_{\frac{0.1}{2}} = z_{0.05} = 1.645$, and $\hat{p} = 0.10$ (the prior estimate).

Solution

Using the formula assuming that a prior estimate is available, $n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$, we obtain

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = 0.10(1 - 0.10) \left(\frac{1.645}{0.02} \right)^2 = 608.9$$

Round this value up to **609**. So the economist must survey **609** randomly selected residents of the United States.

Part A

Part B

Assume that the economist does not use any prior estimates.

Approach

Since $1 - \alpha = 0.9$, we know $\alpha = 0.10$. Use $E = 0.02$, $z_{\frac{\alpha}{2}} = z_{\frac{0.1}{2}} = z_{0.05} = 1.645$.

Solution

Using the formula assuming that a prior estimate is available, $n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$, we obtain

$$n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = 0.25 \left(\frac{1.645}{0.02} \right)^2 = 1691.3$$

Round this value up to **1692**. So the economist must survey **1692** randomly selected residents of the United States.

The effect of not having a prior estimate of p is that the sample size required more than doubled!

9.1.11

November 24, 2016 08:37 AM

Student: Kacey Howell
Date: 11/24/16

Instructor: Matthew Nability
Course: MTH 243: Introduction to
 Probability and Statistics

Assignment: Section 9.1 Homework

Determine the point estimate of the population proportion, the margin of error for the following confidence interval, and the number of individuals in the sample with the specified characteristic, x , for the sample size provided.

Lower bound = 0.117, upper bound = 0.573, $n = 1200$

Confidence interval estimates for the population mean are of the form below.

Point estimate \pm margin of error

Let \hat{p} be the point estimate of the population proportion, and E be the margin of error. Variables u and b are upper and lower bounds of the confidence interval accordingly. The values of u and b are given by the formulas below.

$$u = \hat{p} + E$$

$$b = \hat{p} - E$$

Thus, the value of \hat{p} is given by the following formula.

$$\hat{p} = \frac{u + b}{2}$$

Calculate \hat{p} .

$$\begin{aligned}\hat{p} &= \frac{u + b}{2} \\ &= \frac{0.573 + 0.117}{2} \\ &= \text{.345}\end{aligned}$$

(Round to the nearest thousandth as needed.)

The value of E is given by the following formula.

$$E = \frac{u - b}{2}$$

Calculate E .

$$\begin{aligned}E &= \frac{u - b}{2} \\ &= \frac{0.573 - 0.117}{2} \\ &= \text{.228}\end{aligned}$$

(Round to the nearest thousandth as needed.)

The point estimate for the population proportion is $\hat{p} = \frac{x}{n}$ where x is the number of individuals in the sample with the specified characteristic and n is the sample size. Thus, $x = n\hat{p}$.

Substitute the values of n and \hat{p} into this formula and find x .

$$\begin{aligned}x &= n\hat{p} \\ &= 1200(0.345) \\ &= \boxed{414}\end{aligned}$$

(Round to the nearest integer as needed.)

The point estimate of the population proportion is 0.345, the margin error is 0.228 and the number of individuals in the sample with the specified characteristic is 414.

9.1.35

November 24, 2016 08:20 AM

9.1 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell Date: 11/24/16	Instructor: Matthew Nabity Course: MTH 243: Introduction to Probability and Statistics	Assignment: 9.1 Interactive Assignment
---	---	---

A researcher wishes to estimate the percentage of adults who support abolishing the penny. What size sample should be obtained if he wishes the estimate to be within 4 percentage points with 95% confidence if

- (a) he uses a previous estimate of 36%?
(b) he does not use any prior estimates?

(a) The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval for p with a margin of error E , with a previous

estimate of p , is given by $n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$.

Express the margin of error and the previous estimate as decimals instead of percentages.

$$\begin{array}{ll} E = 4\% & \hat{p} = 36\% \\ E = \text{0.04} & \hat{p} = \text{.36} \end{array}$$

Because a 95% confidence interval is requested, $\alpha = 0.05$. Use technology to find the critical value.

$$\begin{array}{l} z_{\alpha/2} = z_{0.05/2} \\ z_{0.05/2} = z_{0.025} \\ z_{0.025} = \text{1.96} \quad (\text{Round to three decimal places as needed.}) \end{array}$$

Substitute the critical value, $z = 1.96$ into $n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2$, with margin of error $E = 0.04$.

$$\begin{array}{l} n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2 \\ = 0.36(1 - 0.36) \left(\frac{1.96}{0.04} \right)^2 \\ = \text{554} \quad (\text{Round up to the nearest integer.}) \end{array}$$

The actual value for n is 553.19, which is rounded up to the nearest whole number to guarantee that the desired specifications are met, so he must sample 554 adults.

(b) The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval for p with a margin of error E , without a previous

estimate of p , is given by $n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$.

Substitute the same critical value found above into $n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$, with margin of error $E = 0.04$.

$$\begin{array}{l} n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2 \\ = 0.25 \left(\frac{1.96}{0.04} \right)^2 \\ = \text{601} \quad (\text{Round up to the nearest integer.}) \end{array}$$

The actual value for n is 600.25, which is rounded up to the nearest whole number to guarantee that the desired specifications are met, so he must sample 601 adults.

9.2 Estimating a Population Mean

November 24, 2016 09:12 AM

Before getting started, review the following:

1. Degrees of Freedom

For the sample standard deviation, we call $n - 1$ the **degrees of freedom** because the first $n - 1$ observations have freedom to be whatever value they wish, but the n th observation has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

9.2 Objective 1 Obtain a Point Estimate for the Population Mean

November 24, 2016 09:18 AM

Remember, the goal of statistical inference is to use information obtained from a sample and generalize the results to the population being studied. As with estimating the population proportion, the first step is to obtain a point estimate of the parameter.

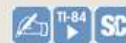
The point estimate of the population mean, μ , is the sample mean, \bar{x} .

EXAMPLE 1 Computing a Point Estimate of the Population Mean

Problem

The website fuelconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 2 show the reported miles per gallon of 2011 Ford Focus automobiles for 16 different owners. Obtain a point estimate of the population mean miles per gallon of a 2011 Ford Focus.

Video Solution



Technology Step-By-Step



TABLE 2

35.7	37.2	34.1	38.9
32.0	41.3	32.5	37.1
37.3	38.8	38.2	39.6
32.2	40.9	37.0	36.0



Approach

Treat the 16 entries as a simple random sample of all 2011 Ford Focus automobiles. To find the point estimate of the population mean, compute the sample mean miles per gallon of the 16 cars.

Solution

The sample mean is

$$\begin{aligned}\bar{x} &= \frac{35.7+32.0+\cdots+36.0}{16} \\ &= \frac{588.8}{16} \\ &= 36.8 \text{ miles per gallon}\end{aligned}$$

The point estimate of μ is 36.8 miles per gallon.

Remember to round statistics to one more decimal point than the raw data, if necessary.

9.2 Objective 2 State the Properties of Student's t-Distribution

November 24, 2016 09:21 AM

In Example 1, a different random sample of 16 cars would likely result in a different point estimate of μ . For this reason, we want to construct a confidence interval for the population mean, just as we did for the population proportion.

A confidence interval for the population mean is of the form point estimate \pm margin of error (just like the confidence interval for a population proportion). To determine the margin of error, we need to know the sampling distribution of the sample mean.

Recall that the distribution of \bar{x} is approximately normal if the population from which the sample is drawn is normal or the sample size is sufficiently large. In addition, the distribution of \bar{x} has the same mean as the parent population, $\mu_{\bar{x}} = \mu$, and a standard deviation equal to the parent population's standard deviation divided by the square root of the sample size,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

This presents a problem because we need to know the population standard deviation, σ , to construct this interval. It does not seem likely that we would know the population standard deviation but not know the population mean. So what can we do?

Student's t -distribution

Suppose that a simple random sample of size n is taken from a population. If the population from which the sample is drawn follows a normal distribution, the distribution of

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

follows Student's t -distribution with $n - 1$ degrees of freedom, where \bar{x} is the sample mean and s is the sample standard deviation.

Properties of the t -Distribution

1. The t -distribution is different for different degrees of freedom.
2. The t -distribution is centered at 0 and is symmetric about 0.
3. The area under the curve is 1. The area under the curve to the right of 0 equals the area under the curve to the left of 0, which equals $1/2$.
4. As t increases or decreases without bound, the graph approaches, but never equals, zero.
5. The area in the tails of the t -distribution is a little greater than the area in the tails of the standard normal distribution because we are using s as an estimate of σ , thereby introducing further variability into the t -statistic.
6. As the sample size n increases, the density curve of t gets closer to the standard normal density curve. This result occurs because as the sample size increases, the values of s get closer to the value of σ by the Law of Large Numbers. See Figure 5, which shows the t -distribution for samples of size $n = 5$ and $n = 15$, along with the standard normal density curve.

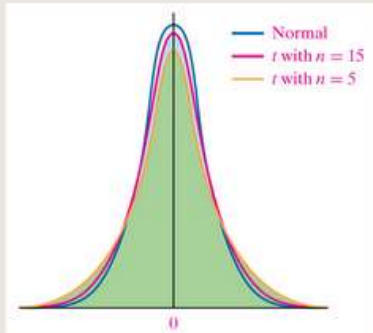


Figure 5

9.2 Objective 3 Determine t-Values

November 24, 2016 09:26 AM

Recall that the notation z_α is used to represent the z -score whose area under the normal curve to the right of z_α is α . Similarly, we let t_α represent the t -value whose area under the t -distribution to the right of t_α is α . See Figure 6.

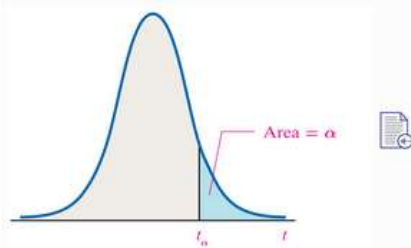


Figure 6

The shape of the t -distribution depends on the sample size, n . Therefore, the value of t_α depends not only on α , but also on the degrees of freedom, $n - 1$. In Table VII, the far left column gives the degrees of freedom (df). The top row represents the area under the t -distribution to the right of some t -value.

t-Distribution Area in Right Tail												
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.065	1.326	1.725	2.086	2.198	2.528	2.847	3.153	3.553	3.850

EXAMPLE 2 Finding t -Values

Problem

Find the t -value such that the area under the t -distribution to the right of the t -value is 0.10, assuming 15 degrees of freedom (df). That is, find $t_{0.10}$ with 15 degrees of freedom.

Video Solution



Approach

Step 1 Draw a t -distribution with the unknown t -value labeled. Shade the area under the curve to the right of the t -value, as in Figure 6.

Step 2 Find the row in Table VII corresponding to 15 degrees of freedom and the column corresponding to an area in the right tail of 0.10. Identify where the row and column intersect. This is the unknown t -value.

Solution

Step 1 Figure 7 shows the graph of the t -distribution with 15 degrees of freedom. The unknown value of t is labeled, and the area under the curve to the right of t is shaded.

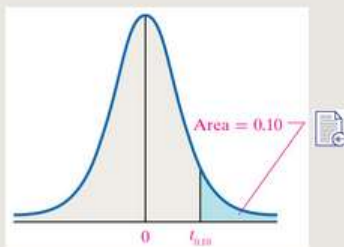
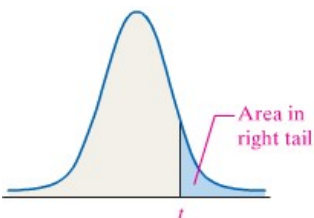


Figure 7

Step 2 A portion of Table VII is shown in Figure 8. We have enclosed the row that represents 15 degrees of freedom and the column that represents the area 0.10 in the right tail. The value where the row and column intersect is the t -value we are seeking. The value of $t_{0.10}$ with 15 degrees of freedom is 1.341; that is, the area under the t -distribution to the right of $t = 1.341$ with 15 degrees of freedom is 0.10. The critical value of z with an area of 0.10 to the right is smaller—approximately 1.28 (Look it up in Table V). This is because the t -distribution has more spread (or more area in the tails) than the z -distribution.

df	Area in Right Tail											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

Figure 8



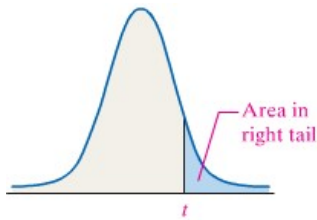


Table VI

**t-Distribution
Area in Right Tail**

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.044	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.042	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

Area in Right Tail

Using Table VII When the Degrees of Freedom Are Not Listed

If the degrees of freedom we desire are not listed in Table VII, choose the closest number in the “df” column.

For example, if we have 43 degrees of freedom, we use 40 degrees of freedom from Table VII.

In addition, the last row of Table VII lists the z -values from the standard normal distribution. Use these values when the degrees of freedom are more than 1000 because the t -distribution starts to behave like the standard normal distribution as n increases.

9.2 Objective 4 Construct and Interpret a Confidence Interval for a Population Mean

November 24, 2016 09:35 AM

We are now ready to construct a confidence interval for a population mean.

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for μ

Provided

- sample data come from a simple random sample or randomized experiment
- sample size is small relative to the population size ($n \leq 0.05N$)
- the data come from a population that is normally distributed, or the sample size is large

A $(1 - \alpha) \cdot 100\%$ confidence interval for μ is given by

$$\text{Lower bound : } \bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad \text{Upper bound : } \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2}}$ is the critical value with $n - 1$ degrees of freedom.

Because this confidence interval uses the t -distribution, it is often referred to as the t -interval.

A Robust Procedure

Notice that a confidence interval about μ can be computed for non-normal populations even though Student's t -distribution requires a normal population. This is because the procedure for constructing the confidence interval is **robust**—it is accurate despite minor departures from normality.

If a data set has outliers, the confidence interval is not accurate because neither the sample mean nor the sample standard deviation is **resistant** to outliers. Sample data should always be inspected for serious departures from normality and for outliers. This is easily done with **normal probability plots** and **boxplots**.

EXAMPLE 3 Constructing a Confidence Interval about a Population Mean

Problem

The website fuelconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 3 show the reported miles per gallon of 2011 Ford Focus automobiles for 16 different owners. Treat the sample as a simple random sample of all 2011 Ford Focus automobiles. Construct a 95% confidence interval for the mean miles per gallon of a 2011 Ford Focus. Interpret the interval.

TABLE 3

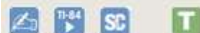
35.7	37.2	34.1	38.9
32.0	41.3	32.5	37.1
37.3	38.8	38.2	39.6
32.2	40.9	37.0	36.0

Approach

Verify the model requirements for constructing a confidence interval about a mean are satisfied by drawing a normal probability plot and boxplot. Then, find the confidence interval either by hand or using technology.

Solution

Video Solution Technology Step-By-Step



Interpretation

We are 95% confident that the mean miles per gallon of all 2011 Ford Focus cars is between 35.24 (Technology: 35.25) and 38.36 (Technology: 38.35) miles per gallon.

TI-83/84 Plus

StatCrunch

Constructing a Confidence Interval for the Population Mean

1. If necessary, enter the raw data into L1.
2. Press STAT, highlight TESTS, and select 8: TInterval...
3. If the data are raw, highlight DATA. Make sure List is set to L1 and Freq to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
4. Enter the confidence level following C-Level:
5. Highlight Calculate and press ENTER.

TI-83/84 Plus

StatCrunch

Constructing a Confidence Interval for the Population Mean

1. If necessary, enter the raw data into column var1. Name the column.
2. Select Stat, highlight T Stats, highlight One Sample. Choose **With Data** if you have raw data, choose **With Summary** if you have summarized data.
3. If you chose **With Data**, highlight the column that contains the data in the "Select column(s):" drop-down menu. If you chose **With Summary**, enter the sample mean, sample standard deviation, and sample size. Choose the confidence interval radio button. Enter the level of confidence. Click Compute!

In Example 3, the critical value is $t_{0.025} = 2.131$ for 15 degrees of freedom, whereas $z_{0.025} = 1.96$. The t -distribution gives a larger critical value, so the width of the interval is wider. This larger critical value using Student's t -distribution is necessary to account for the increased variability due to using s as an estimate of σ .

Remember, 95% confidence refers to our confidence in the method. If we obtained 100 samples of size $n = 16$ from the population of 2011 Ford Focuses, we would expect about 95 of the samples to result in confidence intervals that include μ . We do not know whether the interval in Example 3 includes μ or does not include μ .

When Model Requirements Fail

What should we do if the requirements to compute a t -interval are not met?

Based on the results of Activity 1, we could increase the sample size beyond 30 observations, or we could try to use *nonparametric procedures*. Nonparametric procedures typically do not require normality, and the methods are resistant to outliers. A third option is to use resampling methods, such as bootstrapping. Neither of these options are presented in this course.

9.2 Objective 5 Determine the Sample Size Necessary for Estimating a Population Mean within a Given Margin of Error

November 24, 2016 10:02 AM

The margin of error in constructing a confidence interval about the population mean is $E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$.

Solving this for n , we obtain $n = \left(\frac{t_{\frac{\alpha}{2}} \cdot s}{E} \right)^2$.

The problem with this formula is that the critical value $t_{\frac{\alpha}{2}}$ requires that we know the sample size to determine the degrees of freedom, $n - 1$. Obviously, if we do not know n , we cannot know the degrees of freedom.

The solution to this problem lies in the fact that the t -distribution approaches the standard normal z -distribution as the sample size increases. To convince yourself of this, look at the last few rows of [Table VII](#) and compare them to the corresponding z -scores for 95% or 99% confidence. Now, if we use z in place of t and a sample standard deviation, s , from previous or pilot studies, in place of σ , we can write the margin of error formula $E = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ and solve it for n to obtain a formula for determining sample size.

Determining the Sample Size n

The sample size required to estimate the population mean, μ , with a level of confidence $(1 - \alpha) \cdot 100\%$ within a specified margin of error, E , is given by

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2$$

where n is rounded up to the nearest whole number.

EXAMPLE 4 Determining Sample Size

Problem

We again consider the problem of estimating the miles per gallon of a 2011 Ford Focus. How large a sample is required to estimate the mean miles per gallon within 0.5 miles per gallon with 95% confidence? Note: The sample standard deviation is $s = 2.92$ mpg.

Video Solution



Technology Step-By-Step



Approach

Use $n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E}\right)^2$ with $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, $s = 2.92$, and $E = 0.5$ to find the required sample size.

Solution

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E}\right)^2 = \left(\frac{1.96 \cdot 2.92}{0.5}\right)^2 = 131.02$$

Round 131.02 up to 132. A sample size of $n = 132$ results in an interval estimate of the population mean miles per gallon of a 2011 Ford Focus with a margin of error of 0.5 mile per gallon with 95% confidence.

StatCrunch

Determining Sample Size When Estimating a Population Mean

1. Select Stat, highlight Z Stats, highlight One Sample, and highlight Power/Sample size.
2. Click on the "Confidence Interval Width" tab. Enter the Confidence level and standard deviation. The width is the difference between the lower bound and the upper bound in the confidence interval. Therefore, the width is two times the margin of error. Clear any entry in the sample size cell. Click Compute.

9.2.21

November 24, 2016 09:53 AM

9.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/24/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 9.2 Interactive Assignment

A simple random sample of size n is drawn. The sample mean, \bar{x} , is found to be 18.6, and the sample standard deviation, s , is found to be 4.2.

¹ Click the icon to view the table of areas under the t-distribution.

(a) Construct a 95% confidence interval about μ if the sample size, n , is 35.

Suppose a simple random sample of size n is taken from a population with unknown mean μ and unknown standard deviation σ . A $(1 - \alpha) \cdot 100\%$ confidence interval is given by the following formulas, where $t_{\alpha/2}$ is computed with $n - 1$ degrees of freedom.

$$\text{Lower bound} = \bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

$$\text{Upper bound} = \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Start by determining α .

$\alpha =$

Next determine the degrees of freedom.

There are degrees of freedom.

Now determine the critical t-value, $t_{\alpha/2}$. Find the t-value associated with a right tail area of 0.025 with 34 degrees of freedom.

$$t_{\frac{0.05}{2}} = t_{0.025}$$

$t_{0.025} =$ (Round to three decimal places as needed.)

The margin of error, E , for a confidence interval is given by the following formula.

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Calculate the margin of error.

$E = 2.032 \cdot \frac{4.2}{\sqrt{35}} =$ (Round to three decimal places as needed.)

Now subtract the margin of error from the sample mean to find the lower bound of the confidence interval.

$18.6 - 1.443 =$ (Round to two decimal places as needed.)

Now add the margin of error to the sample mean to find the upper bound of the confidence interval.

$18.6 + 1.443 =$ (Round to two decimal places as needed.)

Thus, the 95% confidence interval is (17.16, 20.04).

(b) Construct a 95% confidence interval about μ if the sample size, n , is 51. How does increasing the sample size affect the margin of error, E ?

First determine the critical t-value, $t_{\alpha/2}$. Find the t-value associated with a right tail area of 0.025 with 50 degrees of freedom.

$t_{0.025} =$ (Round to three decimal places as needed.)

$$t_{0.025} = 2.009 \quad (\text{Round to three decimal places as needed.})$$

1 of 4

11/24/2016 09:53 AM

9.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Now calculate the margin of error.

$$E = 2.009 \cdot \frac{4.2}{\sqrt{51}} = 1.182 \quad (\text{Round to three decimal places as needed.})$$

What is the 95% confidence interval?

(17.42, 19.78)

(Type your answer in interval notation. Round to two decimal places as needed.)

How does increasing the sample size affect the margin of error? Recall that the margin of error for the sample of size 35 is 1.443 and 1.182 for the sample of size 51.

- A. The margin of error does not change.
- B. The margin of error increases.
- C. The margin of error decreases.

(c) Construct a 99% confidence interval about μ if the sample size, n , is 35. Compare the results to those obtained in part (a). How does increasing the level of confidence affect the size of the margin of error, E ?

First determine the critical t -value, $t_{\alpha/2}$. Find the t -value associated with a right tail area of 0.005 with 34 degrees of freedom.

$$t_{0.005} = 2.728 \quad (\text{Round to three decimal places as needed.})$$

Now calculate the margin of error.

$$E = 2.728 \cdot \frac{4.2}{\sqrt{35}} = 1.937 \quad (\text{Round to three decimal places as needed.})$$

What is the 99% confidence interval?

(16.66, 20.54)

(Type your answer in interval notation. Round to two decimal places as needed.)

How does increasing the level of confidence affect the size of the margin of error, E ? Recall that the margin of error for the 95% confidence interval is 1.443 and the margin of error for the 99% confidence interval is 1.937.

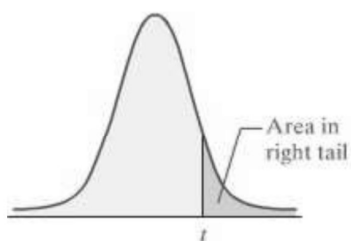
- A. The margin of error increases.
- B. The margin of error does not change.
- C. The margin of error decreases.

(d) If the sample size is 14, what conditions must be satisfied to compute the confidence interval?

In order to compute a confidence interval the sample size must be large ($n \geq 30$) or the sample data must come from a population that is normally distributed with no outliers.

1: Table of t -Distribution Areas

9.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>
**Table VI**
**t-Distribution
Area in Right Tail**

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356

31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307
50	0.670	0.840	1.047	1.300	1.676	2.000	2.100	2.400	2.670	2.937	3.261

3 of 4

11/24/2016 09:53 AM

9.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

YOU ANSWERED: 20.043

.05

B.

C.

9.2.45

November 24, 2016 10:08 AM

9.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 11/24/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 9.2 Interactive Assignment

A doctor wants to estimate the mean HDL cholesterol of all 20- to 29-year-old females. How many subjects are needed to estimate the mean HDL cholesterol within 3 points with 99% confidence assuming $s = 16.9$ based on earlier studies? Suppose the doctor would be content with 95% confidence. How does the decrease in confidence affect the sample size required?

The sample size required to estimate the population mean, with a level of confidence $(1 - \alpha) \cdot 100\%$ with a specified margin of

error, E , is given by $n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$.

Identify the margin of error and the sample standard deviation.

$$E = \text{3}$$

$$s = \text{16.9}$$

(Type integers or decimals.)

Because a 99% confidence interval is requested, use $\alpha = 0.01$. Use technology to find the critical value.

$$z_{\alpha/2} = z_{0.01/2}$$

$$z_{0.01/2} = z_{0.005}$$

$$z_{0.005} = \text{2.576} \quad (\text{Round to three decimal places as needed.})$$

Substitute the critical value found above, the margin of error, and the standard deviation into $n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$.

$$\begin{aligned} n &= \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2 \\ &= \left(\frac{2.576 \cdot 16.9}{3} \right)^2 \\ &= \text{211} \quad (\text{Round up to the nearest subject.}) \end{aligned}$$

Therefore, a 99% confidence level with a margin of error of 3 points requires 211 subjects.

Because a 95% confidence interval is requested, use $\alpha = 0.05$. Use technology to find the critical value.

$$z_{\alpha/2} = z_{0.05/2}$$

$$z_{0.05/2} = z_{0.025}$$

$$z_{0.025} = \text{1.96} \quad (\text{Round to three decimal places as needed.})$$

Substitute the critical value found above, the margin of error, and the standard deviation into $n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$.

$$\begin{aligned} n &= \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2 \\ &= \left(\frac{1.96 \cdot 16.9}{3} \right)^2 \\ &= \text{122} \quad (\text{Round up to the nearest subject.}) \end{aligned}$$

Therefore, a 95% confidence level with a margin of error of 3 points requires 122 subjects.

How does the decrease in confidence affect the sample size required? Recall that a 99% confidence level with a margin of error of 3 points requires 211 subjects.

- A. The sample size is the same for all levels of confidence.
- B. Decreasing the confidence level decreases the sample size needed.
- C. Decreasing the confidence level increases the sample size needed.

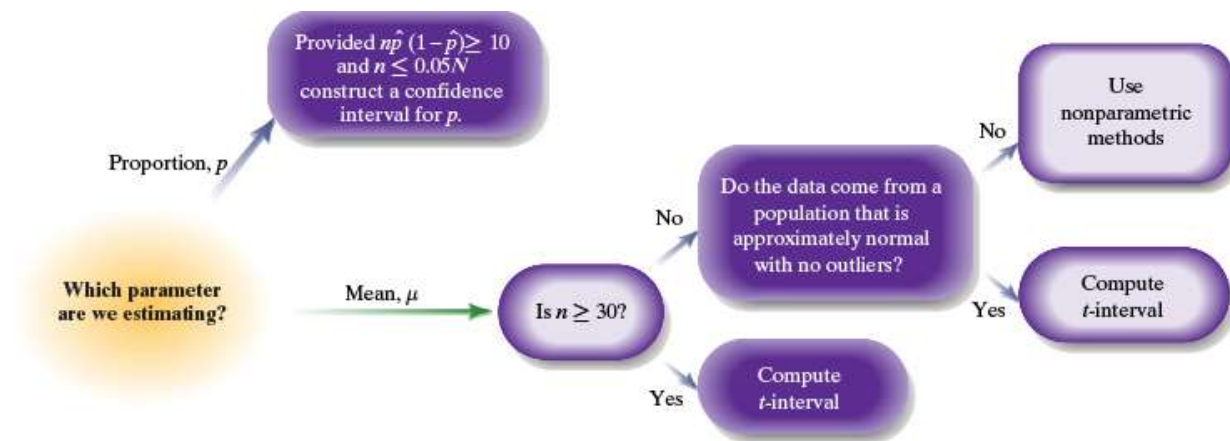
YOU ANSWERED: C.

9.3 Putting it Together: Which Procedure Do I Use?

November 25, 2016 08:02 AM

9.3 Objective 1 Determine the Appropriate Confidence Interval to Construct

November 25, 2016 08:48 AM



10.1 The Language of Hypothesis Testing

November 28, 2016 09:58 AM



Well done!



Although sample data is used to test the null hypothesis, it cannot be stated with 100% certainty that the null hypothesis is true. It can only be determined whether the sample data supports or does not support the null hypothesis.

Next Question

10.1 Objective 1 Determine the Null and Alternative Hypothesis

November 28, 2016 10:03 AM

Left- and right-tailed tests are referred to as **one-tailed tests**. Notice that in the left-tailed test, the direction of the inequality sign in the alternative hypothesis points to the left ($<$), whereas in the right-tailed test, the direction of the inequality sign in the alternative hypothesis points to the right ($>$). In all three tests, the null hypothesis contains a statement of equality.

The statement we are trying to gather evidence for, which is dictated by the researcher before any data are collected, determines the structure of the alternative hypothesis (two-tailed, left-tailed, or right-tailed). For example, the label on a can of soda states that the can contains 12 ounces of liquid. A consumer advocate would be concerned if the mean contents were less than 12 ounces, so the alternative hypothesis is

$H_1 : \mu < 12$ ounces. However, a quality control engineer for the soda

manufacturer would be concerned if there was too little or too much soda in the can, so the alternative hypothesis would be

$H_1 : \mu \neq 12$ ounces. In both cases, the null hypothesis is a statement of no difference between the manufacturer's assertion on the label and the actual mean contents of the can, so the null hypothesis is $H_0 : \mu = 12$ ounces.

IN
OTHER
WORDS



EXAMPLE 1 Forming Hypotheses

Problem

For each situation, determine the null and alternative hypotheses. State whether the test is two-tailed, left-tailed, or right-tailed.

Video Solution



Part A

Part B

Part C

The Medco pharmaceutical company has just developed a new antibiotic for children. Two percent of children taking competing antibiotics experience headaches as a side effect. A researcher for the Food and Drug Administration wants to know if the percentage of children taking the new antibiotic and experiencing headaches as a side effect is more than 2%.

Approach

We must determine the parameter to be tested, the statement of no change or no difference (status quo), and the statement for which we are attempting to gather evidence.

Solution

The hypothesis deals with a population proportion, p . If the new drug is no different from competing drugs, then the proportion of individuals taking it who experience a headache will be 0.02, so the null hypothesis is $H_0 : p = 0.02$. We want to determine whether the proportion of individuals who experience a headache is more than 0.02, so the alternative hypothesis is $H_1 : p > 0.02$. This is a right-tailed test because the alternative hypothesis contains a $>$ symbol.

IN
OTHER
WORDS



Part A

Part B

Part C

The *Blue Book* value of a used three-year-old Chevy Corvette Z06 is \$56,130. Grant wonders if the mean price of a used three-year-old Chevy Corvette Z06 in the Miami metropolitan area is different from \$56,130.

Approach

We must determine the parameter to be tested, the statement of no change or no difference (status quo), and the statement for which we are attempting to gather evidence.

Solution

The hypothesis deals with a population mean, μ . If the mean price of a three-year-old Corvette Z06 in Miami is no different from the *Blue Book* price, then the population mean in Miami will be \$56,130, so the null hypothesis is $H_0: \mu = \$56,130$. Grant wants to know if the mean price is different from \$56,130, so the alternative hypothesis is $H_1: \mu \neq \$56,130$. This is a two-tailed test because the alternative hypothesis contains a \neq symbol.

Part A

Part B

Part C

The standard deviation of the contents in a 64-ounce bottle of detergent using an old filling machine is 0.23 ounce. The manufacturer wants to know if a new filling machine has less variability.

Approach

We must determine the parameter to be tested, the statement of no change or no difference (status quo), and the statement for which we are attempting to gather evidence.

Solution

The hypothesis deals with a population standard deviation, σ . If the new machine is no different from the old one, then the standard deviation of the amount in the bottles filled by the new machine will be 0.23 ounce, so the null hypothesis is $H_0: \sigma = 0.23$ ounce. The company wants to know if the new machine has less variability than the old machine, so the alternative hypothesis is $H_1: \sigma < 0.23$ ounce. This is a left-tailed test because the alternative hypothesis contains a $<$ symbol.

10.1 Objective 2 Explain Type I and Type II Errors

November 28, 2016 10:10 AM

Sample data is used to decide whether to reject the statement in the null hypothesis. Because this decision is based on incomplete (sample) information, there is always the possibility of making an incorrect decision. In fact, there are four possible outcomes from hypothesis testing.

FOUR OUTCOMES FROM HYPOTHESIS TESTING

1. Reject the null hypothesis when the alternative hypothesis is true. This decision would be correct.
2. Do not reject the null hypothesis when the null hypothesis is true. This decision would be correct.
3. Reject the null hypothesis when the null hypothesis is true. This decision would be incorrect. This type of error is called a **Type I error**.
4. Do not reject the null hypothesis when the alternative hypothesis is true. This decision would be incorrect. This type of error is called a **Type II error**.

IN
OTHER
WORDS



		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

Figure 1

Example – Type I and Type II Errors

The Medco pharmaceutical company has just developed a new antibiotic.

Two percent of children taking competing antibiotics experience headaches as a side effect.

A researcher for the Food and Drug Administration wishes to know if the percentage of children taking the new antibiotic who experience a headache as a side effect is more than 2%.

The researcher conducts a hypothesis test with $H_0: p = 0.02$ and $H_1: p > 0.02$.

Explain what it would mean to make a (a) Type I error and (b) Type II error.

Solution – (a)

A Type I error occurs if we reject the null hypothesis when it is true.

A Type I error is made if the sample evidence leads the researcher to believe that $p > 0.02$ (that is, we reject the null hypothesis) when, in fact, the proportion of children who experience a headache is not greater than 0.02.

Solution – (b)

A Type II error occurs if we do not reject the null hypothesis when the alternative hypothesis is true.

A Type II error is made if the researcher does not reject the null hypothesis that the proportion of children experiencing a headache is equal to 0.02 when, in fact, the proportion of children who experience a headache is more than 0.02.

In other words, the sample evidence led the researcher to believe $p = 0.02$ when in fact the true proportion is some value larger than 0.02.

The Probability of Making a Type I or Type II Error

When we studied how to construct confidence intervals, we learned that we never know whether a **confidence interval** contains the unknown parameter. We only know the likelihood that a confidence interval captures the parameter. Similarly, we never know whether the conclusion of a hypothesis test is correct. However, just as we place a level of confidence in the construction of a confidence interval, we can assign probabilities to making Type I or Type II errors when testing hypotheses. The following notation is commonplace:

$$\alpha = P(\text{Type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$$
$$\beta = P(\text{Type II error}) = P(\text{not rejecting } H_0 \text{ when } H_1 \text{ is true})$$

The symbol β is the Greek letter **beta** (pronounced "BAY tah"). The probability of making a Type I error, α , is chosen by the researcher *before* the sample data are collected. This probability is called the *level of significance*.

DEFINITION

The level of significance, α , is the probability of making a Type I error.

The choice of the level of significance depends on the consequences of making a Type I error. If the consequences are severe, the level of significance should be small (say, $\alpha = 0.01$). However, if the consequences are not severe, a higher level of significance can be chosen (say, $\alpha = 0.05$ or $\alpha = 0.10$).

Why is the level of significance not always set at $\alpha = 0.01$? Reducing the probability of making a Type I error increases the probability of making a Type II error, β . Using our court analogy from the video explaining [Figure 1](#), a jury is instructed that the prosecution must provide proof of guilt "beyond all reasonable doubt." This implies that we are choosing to make α small so that the probability of convicting an innocent person is very small. The consequence of the small α , however, is a large β , which means many guilty defendants will go free. For now, we are content to recognize the inverse relation between α and β . (As one goes up, the other goes down.)

IN
OTHER
WORDS



10.1 Objective 3 State Conclusions to Hypothesis Tests

November 28, 2016 10:18 AM

Once the decision is made whether to reject the null hypothesis, the researcher must state his or her conclusion. It is important to recognize that we never *accept* the null hypothesis. Watch the Caution video.



So sample evidence can never prove the null hypothesis to be true. By not rejecting the null hypothesis, we are saying that the evidence indicates that the null hypothesis *could* be true or that the sample evidence is consistent with the statement in the null hypothesis.

EXAMPLE 3 Stating the Conclusion

Problem

The Medco pharmaceutical company has just developed a new antibiotic. Two percent of children taking competing antibiotics experience a headache as a side effect. A researcher for the Food and Drug Administration believes that the proportion of children taking the new antibiotic who experience a headache as a side effect is more than 0.02. So the null hypothesis is $H_0 : p = 0.02$ and the alternative hypothesis is $H_1 : p > 0.02$.

Video Solution



Part A

Part B

Suppose the sample evidence indicates that the null hypothesis is rejected. State the conclusion.

Approach

When the null hypothesis is rejected, we say that there is sufficient evidence to support the statement in the alternative hypothesis. When the null hypothesis is *not* rejected, we say that there is *not* sufficient evidence to support the statement in the alternative hypothesis. We never say that the null hypothesis is true!

Solution

The statement in the alternative hypothesis is that the proportion of children taking the new antibiotic who experience a headache as a side effect is more than 0.02. Because the null hypothesis ($p = 0.02$) is rejected, there is sufficient evidence to conclude that the proportion of children who experience a headache as a side effect is more than 0.02.

Suppose the sample evidence indicates that the null hypothesis is not rejected. State the conclusion.

Approach

When the null hypothesis is rejected, we say that there is sufficient evidence to support the statement in the alternative hypothesis. When the null hypothesis is *not* rejected, we say that there is *not* sufficient evidence to support the statement in the alternative hypothesis. We never say that the null hypothesis is true!

Solution

Because the null hypothesis is not rejected, there is not sufficient evidence to say that the proportion of children who experience a headache as a side effect is more than 0.02 .

10.2 Hypothesis Tests for a Population proportion

November 28, 2016 10:22 AM

10.2 Objective 1 Explain the Logic of Hypothesis Testing

December 7, 2016 09:24 AM

Activity—The Logic of Hypothesis Testing

Suppose a politician wants to know if a majority (more than 50%) of her constituents are in favor of a certain policy. The politician hires a polling firm to obtain a random sample of 500 registered voters in her district and asks them to disclose whether they are in favor of the policy. What would be convincing evidence that the true (that is, population) proportion of registered voters is greater than 50%? In this scenario, we are testing

$$H_0 : p = 0.5 \text{ versus } H_1 : p > 0.5.$$

The applet on the next screen will simulate surveying 500 registered voters, assuming that 50% of all registered voters are in favor of the policy and 50% are against the policy. This is done because we always assume that the statement in the null hypothesis is true until we have evidence to the contrary. So we assume that we are sampling from a population where the proportion of registered voters who are in favor of the policy is 0.5.

In the Logic of Hypothesis Testing Activity, you were likely not convinced that a majority of registered voters were in favor of the policy when the sample proportion was 0.52. But as the sample proportion started increasing (from $\hat{p} = 0.52$ to $\hat{p} = 0.54$ to $\hat{p} = 0.56$), you likely became more convinced that the statement in the null ($p = 0.5$) should be rejected. What is convincing, or *statistically significant*, evidence that the population proportion is greater than 0.5?

DEFINITION

When observed results are unlikely under the assumption that the null hypothesis is true, we say that the result is **statistically significant** and we reject the null hypothesis.

Did you notice that the shape of the distribution of possible outcomes from the applet activity was bell-shaped? This is no coincidence. From our study of sampling distributions, we know that the sample distribution of \hat{p} is approximately normal, with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, provided that the following requirements are satisfied:

1. The sample is a simple random sample.
2. $np(1-p) \geq 10$
3. The sampled values are independent of each other ($n \leq 0.05N$).

Rather than using a simulation to determine whether a sample proportion is statistically significant, we can build a probability model for the distribution of \hat{p} . Because $np(1-p) = 500(0.5)(1-0.5) = 125 \geq 10$ and the sample size ($n = 500$) is less than 5% of the population size (provided there are at least $N = 10,000$ registered voters in the politician's district), we can use a normal model to describe the distribution of \hat{p} . That is, we can describe the variability in \hat{p} that we saw in the simulation activity.

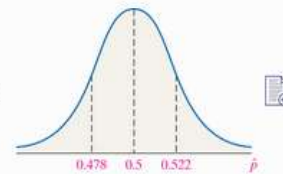


Figure 2

The mean of the distribution of \hat{p} is $\mu_{\hat{p}} = 0.5$ (because we assume that the statement in the null hypothesis is true), and the standard deviation of the distribution of \hat{p} is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{500}} \approx 0.022$.

Figure 2 shows the sampling distribution of the sample proportion for the "Logic of Hypothesis Testing Activity" example.

Now that we have a model that describes the distribution of the sample proportion, we can use it to look at the logic of the P -value approach to test whether a majority of the politician's constituents are in favor of the policy.

A criterion for testing hypotheses using the P -value approach is to determine how likely the observed sample proportion is under the assumption that the statement in the null hypothesis is true. For example, in part (c) from the Logic of Hypothesis Testing Activity, we used simulation to determine the likelihood of obtaining a sample proportion of $\hat{p} = 0.52$ or higher from a population whose proportion is assumed to be 0.5. If a sample proportion of 0.52 or higher is unlikely (or unusual), we have evidence against the statement in the null hypothesis. Otherwise, we do not have sufficient evidence against the statement in the null hypothesis.

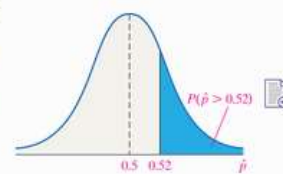


Figure 3

Using the normal model in Figure 2, we can compute the probability of obtaining a sample proportion of 0.52 or higher from a population whose proportion is 0.5. See Figure 3.

We can find the area to the right of 0.52 by hand or by using technology. Using a statistical calculator, we find the area to be 0.1855. Therefore,

$$P(\hat{p} > 0.52) = 0.1855$$

Look back at the results of the "Logic of Hypothesis Testing Activity" from part (c). What proportion of the 1000 simulations resulted in a sample proportion of 0.52 or higher? Is it close to 0.1855?

We also could use the normal model to find the area to the right of 0.54 or 0.56. Again, using a statistical calculator, we find

$$P(\hat{p} > 0.54) = 0.0368 \text{ and } P(\hat{p} > 0.56) = 0.0036$$

Look back at the results of the "Logic of Hypothesis Testing Activity" from parts (D) and (E). What proportion of the 1000 simulations resulted in a sample proportion of 0.54 or higher? 0.56 or higher? Are the results close to the probabilities obtained from the normal model?

We conclude that the likelihood of obtaining a sample statistic can be obtained either through simulation (repeatedly taking samples of size n from the population under the assumption the statement in the null hypothesis is true and determining the number of times we observe a statistic as extreme as or more extreme than the one obtained) or through the use of the normal model. Both approaches give similar results. Why? Because the normal model accurately describes the distribution of the sample proportion! Essentially, simulation represents an empirical probability and the normal model represents a classical probability. Remember, with empirical probability, we estimate the probability of an event by conducting the experiment over and over, whereas classical probability obtains probabilities through counting techniques. (Here we are using the normal model to obtain the probability.)

What do the probabilities (such as $P(\hat{p} > 0.52) = 0.1855$) represent? They are called P -values. For example, $P(\hat{p} > 0.52) = 0.1855$ means that about 18 or 19 samples (of size 500 voters) in 100 will give a sample proportion as high as or higher than the one we obtained if the population proportion really is 0.5. Because these results are not unusual, we say that we do not have enough evidence to reject the statement in the null hypothesis. However, if we observed 280 out of 500 voters in favor of the policy, then $P(\hat{p} > 0.56) = 0.0036$. This means that only about 3 or 4 samples in 1000 will give a sample proportion as high as or higher than the one we obtained if the population proportion really is 0.5. Because it is unlikely to obtain a sample proportion of 0.56 from a population whose proportion is 0.5, these results are unusual. So we take this as evidence against the statement in the null hypothesis.

DEFINITION

A P -value is the probability of observing a sample statistic as extreme as or more extreme than one observed under the assumption that the statement in the null hypothesis is true. Stated another way, the P -value is the likelihood or probability that a sample will result in a statistic such as the one obtained if the null hypothesis is true.

Hypothesis Testing Using the P -Value Approach

If the probability of getting a sample statistic as extreme as or more extreme than the one obtained is small under the assumption that the statement in the null hypothesis is true, reject the null hypothesis.

Figure 4 further illustrates the idea behind the p -value approach. The distribution in blue shows the distribution of the sample proportion, assuming that the statement in the null hypothesis is true. The sample proportion of 0.54 is too far from the assumed population proportion of 0.5. In other words, obtaining a sample proportion of 0.54 or higher from a population whose proportion is 0.5 is unlikely. Therefore, we reject the null hypothesis that $p = 0.5$ and conclude that $p > 0.5$, as indicated by the distribution in red. We do not know what the population proportion of registered voters who are in favor of the policy is, but we have evidence to say that it is greater than 0.5 (a majority).

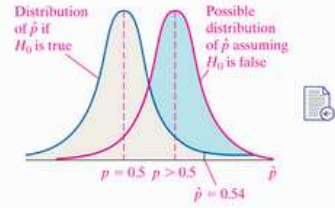


Figure 4

10.2.2 Test Hypothesis about a Population Proportion

December 7, 2016 09:28 AM

Testing Hypotheses Regarding a Population Proportion, p

Use Steps 1–5 shown below provided that

- the sample is obtained by **simple random sampling** or the data result from a randomized **experiment**;
- $np_0(1 - p_0) \geq 10$ where p_0 is the proportion stated in the null hypothesis; and
- the sampled values are independent of each other. This means that the sample size is no more than 5% of the population size ($n \leq 0.05N$).

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$
$H_1: p \neq p_0$	$H_1: p < p_0$	$H_1: p > p_0$

Note: p_0 is the assumed value of the population proportion.

Step 2 Select a level of significance, α , depending on the seriousness of making a **Type I error**.

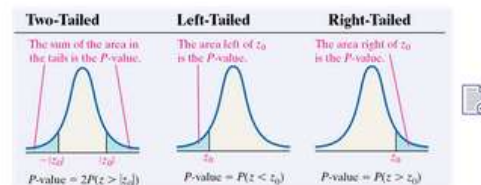
By Hand Step 3

Technology Step 3

Compute the test statistic.

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Use [Table V](#) to find the P -value as shown in the figures.



Notice in the computation of the test statistic that we are using p_0 (the proportion stated in the null hypothesis) in computing the standard error rather than \hat{p} , as we did in constructing confidence intervals about p . This is because H_0 is assumed to be true when we are performing a hypothesis test. So the assumed mean of the distribution of \hat{p} is $\mu_{\hat{p}} = p_0$, and the assumed standard error is $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$.

By Hand Step 3

Technology Step 3

Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value.

Step 4 If $P\text{-value} < \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

EXAMPLE 1 Testing a Hypothesis about a Population Proportion: Left-Tailed Test

Problem

The two major college entrance exams that a majority of colleges accept for student admission are the SAT and ACT. ACT looked at historical records and established 22 as the minimum ACT math score for a student to be considered prepared for college mathematics. (Note: "Being prepared" means that there is a 75% probability of successfully completing College Algebra in college.) An official with the Illinois State Department of Education wonders whether less than half of the students in her state are prepared for College Algebra. She obtains a simple random sample of 500 records of students who have taken the ACT and finds that 219 are prepared for college mathematics (that is, scored at least 22 on the ACT math test). Does this represent significant evidence that less than half of Illinois students are prepared for college mathematics upon graduation from a high school? Use the $\alpha = 0.05$ level of significance. Data from ACT High School Profile Report.

Approach

The problem deals with a hypothesis test of a proportion. We want to determine whether the sample evidence suggests that less than half of the students are prepared for college mathematics. Symbolically, we represent this as $p < \frac{1}{2}$ or $p < 0.5$. Verify the three requirements to perform the hypothesis test: The sample must be a simple random sample or the result of a randomized experiment, $np_0(1 - p_0) \geq 10$, and the sample size cannot be more than 5% of the population size (for independence). Then follow Steps 1 through 5.

TI-83/84 Plus

StatCrunch

Hypothesis Tests Regarding a Population Proportion

1. Press STAT, highlight TESTS, and select 5:1-PropZTest.
2. For the value of p_0 , enter the value of the population proportion stated in the null hypothesis.
3. Enter the number of successes, x , and the sample size, n .
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw and press ENTER.

Hypothesis Tests Regarding a Population Proportion

1. If necessary, enter the raw data into column var1. Name the column.
2. Select **Stat**, highlight **Proportion Statistics**, highlight **One Sample**. Choose **With Data** if you have raw data, choose **With Summary** if you have summarized data.
3. If you chose **With Data**, highlight the column that contains the data in the "Values in:" drop-down menu. Enter the value that represents a success. If you chose **With Summary**, enter the number of successes and the number of observations. Choose the hypothesis test radio button. Enter the value of the proportion stated in the null hypothesis and choose the direction of the alternative hypothesis from the drop-down menu. Click **Compute!**.

Result

The p -value of **0.003** means that *if* the null hypothesis that $p = 0.5$ is true, we expect **219** or fewer successes in **500** trials in less than **1** sample in **100**. Because the results are unusual (the p -value is less than the level of significance, α), we reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that less than half of the Illinois students are prepared for college mathematics. In other words, the data suggest that less than a majority of the students in Illinois are prepared for college mathematics.

EXAMPLE 2 Testing a Hypothesis about a Population Proportion: Two-Tailed Test

Problem

When asked the following question, "Which do you think is more important—protecting the right of Americans to own guns or controlling gun ownership?", 46% of Americans said that protecting the right to own guns is more important. The Pew Research Center surveyed 1267 randomly selected Americans with at least a bachelor's degree and found that 559 believed that protecting the right to own guns is more important. Does this result suggest that the proportion of Americans with at least a bachelor's degree feel differently than the general American population when it comes to gun control? Use the $\alpha = 0.1$ level of significance.

Approach

The problem deals with a hypothesis test of a proportion. Verify the three requirements to perform the hypothesis test. Then follow Steps 1 through 5.

Solution

Video Solution Technology Step-By-Step



Result

The P -value of 0.1738 [Technology: 0.1794] means that if the null hypothesis that $p = 0.46$ is true, we expect the type of results we observed (or more extreme results) in about 17 or 18 out of 100 samples. Because the results are not unusual (the P -value is greater than the level of significance), we do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.1$ level of significance to conclude that Americans with at least a bachelor's degree feel differently than the general American population when it comes to gun control.



Test a Hypothesis Using a Confidence Interval

Recall that the level of confidence $(1 - \alpha) \cdot 100\%$ in a confidence interval represents the percentage of intervals that will contain the unknown parameter if repeated samples are obtained.

Two-Tailed Hypothesis Testing Using Confidence Intervals

When testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, if a $(1 - \alpha) \cdot 100\%$ confidence interval contains p_0 , we do not reject the null hypothesis. However, if the confidence interval does not contain p_0 , we conclude that $p \neq p_0$ at the level of significance, α .

EXAMPLE 3 Testing a Hypothesis Using a Confidence Interval

Problem

A 2009 study by Princeton Survey Research Associates International found that 34% of teenagers text while driving. Does a 2012 survey conducted by Consumer Reports, which found that 353 of 1200 randomly selected teens had texted while driving, suggest that the proportion of teens who text while driving has changed since 2009? Use a 95% confidence interval to answer the question.

Approach

Find the 95% confidence interval. If the interval does not include 0.34, reject the null hypothesis $H_0 : p = 0.34$ in favor of $H_1 : p \neq 0.34$.

Solution

Video Solution Technology Step-By-Step



Result

The 95% confidence interval for p based on the Consumer Reports survey has a lower bound of 0.268 and an upper bound of 0.320. Because 0.34 is not within the bounds of the confidence interval, there is sufficient evidence to conclude that the proportion of teens who text while driving has changed since 2009.

The same Consumer Reports article cited in Example 3 states that 75% of teens have friends who text while driving. What does this say about the difficulty in finding truthful responses to questions while conducting a survey?

10.2.19-T

December 7, 2016 09:39 AM

10.2 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 12/7/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to Probability and Statistics

Assignment: 10.2 Interactive Assignment

In a recent survey, 24% of employed U.S. adults reported that basic mathematical skills were critical or very important to their job. The supervisor of the job placement office at a 4-year college thinks this percentage has increased due to increased use of technology in the workplace. She takes a random sample of 250 employed adults and finds that 80 of them feel that basic mathematical skills are critical or very important to their job. Is there sufficient evidence to conclude that the percentage of employed adults who feel basic mathematical skills are critical or very important to their job has increased at the $\alpha = 0.1$ level of significance?

To test the hypothesis, first verify the requirements; that is, the sample must be a simple random sample, $np_0(1 - p_0) \geq 10$, and the sample size cannot be more than 5% of the population size.

As stated above, the sample is a simple random sample. Also, there are more than 150 million adults in the U.S., so the sample size is less than 5% of the population size. Compute $np_0(1 - p_0)$ to verify that $np_0(1 - p_0) \geq 10$.

$$np_0(1 - p_0) = (250)(0.24)(1 - 0.24) = 45.6 \quad (\text{Round to one decimal place as needed.})$$

Thus, $np_0(1 - p_0) = 45.6 \geq 10$ and all of the requirements are satisfied.

Now that all the requirements have been verified, the next step is to determine the null and alternative hypotheses. The statement is that the percentage of employed adults who feel basic mathematical skills are critical or very important to their job has increased from the previous percentage. What are the null and alternative hypotheses?

- A. $H_0: p = 0.24$ versus $H_1: p \neq 0.24$ B. $H_0: p > 0.24$ versus $H_1: p = 0.24$
 C. $H_0: p = 0.24$ versus $H_1: p < 0.24$ D. $H_0: p = 0.24$ versus $H_1: p > 0.24$

Determine the test statistic. For the purposes of this exercise, the test statistic will be found using technology, but the formula below can also be used.

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

What is the test statistic?

$$z_0 = 2.96 \quad (\text{Round to two decimal places as needed.})$$

Now continue with the P-value approach. While either a table of standard normal values or technology can be used to find the P-value, this exercise will use technology.

$$\text{P-value} = 0.002 \quad (\text{Round to three decimal places as needed.})$$

Finally, compare the P-value with α . If $\text{P-value} < \alpha$, reject the null hypothesis.

10.2.21-T

December 7, 2016 09:48 AM

Previously, 47% of parents of children in high school felt it was a serious problem that high school students were not being taught enough math and science. A recent survey found that 396 of 800 parents of children in high school felt it was a serious problem that high school students were not being taught enough math and science. Do parents feel differently today? Use the $\alpha = 0.1$ level of significance.

What are the null and alternative hypotheses?

$H_0: p = 0.47$ versus $H_1: p \neq 0.47$

(Type integers or decimals.)

Calculate the test statistic.

$z_0 = 1.42$ (Round to two decimal places as needed.)

Use technology to find the P-value.

P-value = 0.157 (Round to three decimal places as needed.)

Interpret the P-value. Select the correct choice below and fill in the answer box to complete your choice.

(Type an integer or a decimal.)

- A. The P-value is the probability of observing a sample statistic as extreme or more extreme than the one obtained if the population proportion equals 1.
- B. The P-value is the probability of observing a sample statistic as small or smaller than the one obtained if the population proportion equals .
- C. The P-value is the probability of observing a sample statistic as small or smaller than the one obtained if the population proportion is less than .
- D. The P-value is the probability of observing a sample statistic as extreme or more extreme than the one obtained if the population proportion is greater than .

✘ Sorry, that's not correct.



Sorry, your answer is not correct.

Correct answer: A: 0.47

Your answer: A: 1

OK

10.2.23

December 7, 2016 09:51 AM

Several years ago, 45% of parents who had children in grades K-12 were satisfied with the quality of education the students receive. A recent poll asked 1,065 parents who have children in grades K-12 if they were satisfied with the quality of education the students receive. Of the 1,065 surveyed, 499 indicated that they were satisfied. Construct a 99% confidence interval to assess whether this represents evidence that parents' attitudes toward the quality of education have changed.

What are the null and alternative hypotheses?

$H_0: p = 0.45$ versus $H_1: p \neq 0.45$

(Round to two decimal places as needed.)

Use technology to find the 99% confidence interval.

(0.43 , 0.51)

(Round to two decimal places as needed.)

What is the correct conclusion?

- A. Since the interval contains the proportion stated in the null hypothesis, there is insufficient evidence that parents' attitudes toward the quality of education have changed.
- B. Since the interval does not contain the proportion stated in the null hypothesis, there is sufficient evidence that parents' attitudes toward the quality of education have changed.
- C. Since the interval does not contain the proportion stated in the null hypothesis, there is insufficient evidence that parents' attitudes toward the quality of education have changed.
- D. Since the interval contains the proportion stated in the null hypothesis, there is sufficient evidence that parents' attitudes toward the quality of education have changed.

10.2.3 Test Hypotheses about a Population Proportion Using the Binomial Probability Distribution

December 7, 2016 09:51 AM

For the sampling distribution of \hat{p} to be approximately normal, we require that $np(1-p)$ be at least 10. If this requirement is not satisfied we use the binomial probability formula to determine the P -value.

From <<https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx>>

EXAMPLE 4 Hypothesis Test for a Population Proportion: Small Sample Size

Problem

According to the U.S. Department of Agriculture, 48.9% of males aged 20 to 39 years consume the recommended daily requirement of calcium. After an aggressive "Got Milk" advertising campaign, the USDA conducts a survey of 35 randomly selected males aged 20 to 39 and finds that 21 of them consume the recommended daily allowance (RDA) of calcium. At the $\alpha = 0.10$ level of significance, is there evidence to conclude that the percentage of males aged 20 to 39 who consume the RDA of calcium has increased?

Approach

We use the following steps:

Step 1 Determine the null and alternative hypotheses.

Step 2 Check whether $np_0(1 - p_0)$ is greater than or equal to 10, where p_0 is the proportion stated in the null hypothesis. If it is, then the sampling distribution of \hat{p} is approximately normal and we can use the steps presented earlier. Otherwise, we use Steps 3 and 4, presented next.

Step 3 Compute the P -value. For right-tailed tests, the P -value is the probability of obtaining x or more successes. For left-tailed tests, the P -value is the probability of obtaining x or fewer successes. For two-tailed tests, the P -value is 2 times the probability of obtaining x or more successes if $\hat{p} > p$ and 2 times the probability of obtaining x or fewer successes if $\hat{p} < p$. The P -value is always computed with the proportion given in the null hypothesis. Remember, we assume that the null is true until we have evidence to the contrary.

Step 4 If the P -value is less than the level of significance, α , reject the null hypothesis.

Solution

Step 1 The status quo, or no-change, proportion of 20- to 39-year-old males who consume the recommended daily requirement of calcium is 0.489. We want to know whether the advertising campaign increased this proportion. Therefore,

$$H_0 : p = 0.489 \text{ and } H_1 : p > 0.489$$

Step 2 From the null hypothesis, we have $p_0 = 0.489$. There were $n = 35$ individuals surveyed, so $np_0(1 - p_0) = 35(0.489)(1 - 0.489) = 8.75$. Because $np_0(1 - p_0) < 10$, the sampling distribution of \hat{p} is not approximately normal.

Video Solution



Solution

Step 1 The status quo, or no-change, proportion of 20- to 39-year-old males who consume the recommended daily requirement of calcium is 0.489. We want to know whether the advertising campaign increased this proportion. Therefore,

Video Solution



$$H_0 : p = 0.489 \text{ and } H_1 : p > 0.489$$

Step 2 From the null hypothesis, we have $p_0 = 0.489$. There were $n = 35$ individuals surveyed, so $np_0(1 - p_0) = 35(0.489)(1 - 0.489) = 8.75$. Because $np_0(1 - p_0) < 10$, the sampling distribution of \hat{p} is not approximately normal.

Step 3 Let the random variable X represent the number of individuals who consume the daily requirement of calcium. We have $x = 21$ successes in $n = 35$ trials, so $\hat{p} = \frac{21}{35} = 0.6$. We want to judge whether the larger proportion is due to an increase in the population proportion or to sampling error. We obtained $x = 21$ successes in the survey, and this is a right-tailed test; so the P -value is $P(X \geq 21)$.

$$P\text{-value} = P(X \geq 21) = 1 - P(X < 21) = 1 - P(X \leq 20) = 0.1261$$

Step 4 The P -value is greater than the level of significance ($0.1261 > 0.10$), so we do not reject H_0 . There is not sufficient evidence (at the $\alpha = 0.1$ level of significance) to conclude that the proportion of 20- to 39-year-old males who consume the recommended daily allowance of calcium has increased. Notice that the sample proportion, 0.6, is much larger than the proportion stated in the null hypothesis, 0.489. Yet, we were not able to reject the null hypothesis. This serves as a reminder that small sample hypothesis tests require substantial evidence against the null in order to reject the null.

10.3 Hypothesis Tests for a Population Mean

December 7, 2016 09:56 AM

Student's t -Distribution

Suppose that a simple random sample of size n is taken from a population. If the population from which the sample is drawn follows a normal distribution, the distribution of

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

follows Student's t -distribution with $n - 1$ degrees of freedom, where \bar{x} is the sample mean and s is the sample standard deviation.

10.3.1 Test Hypotheses About a Mean

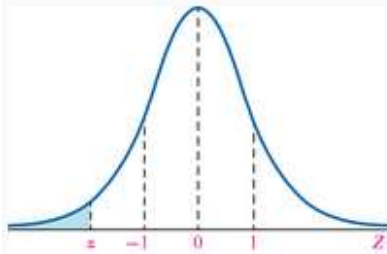
December 7, 2016 09:57 AM

We know that \bar{x} is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ provided the population from which the sample was drawn is normally distributed or the sample size is sufficiently large (because of the Central Limit Theorem). So $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ follows a standard normal distribution. However, it is unreasonable to expect to know σ without knowing μ . Recall that this problem was resolved by William Gosset, who determined that $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ follows Student's t -distribution with $n - 1$ degrees of freedom. We use this distribution to perform hypothesis tests on a mean.

The Central Limit Theorem

Regardless of the shape of the underlying population, the sampling distribution of \bar{x} becomes approximately normal as the sample size, n , increases.

The standard normal distribution is a normal distribution with mean 0 and standard deviation 1.



Testing Hypotheses Regarding a Mean

Use Steps 1–5 shown below provided that

- the sample is obtained by simple random sampling or the data result from a randomized experiment;
- the sample has no outliers, and the population from which the sample is drawn is normally distributed or the sample size, n , is large ($n \geq 30$); and
- the sampled values are independent of each other. This means that the sample size is no more than 5% of the population size ($n \leq 0.05N$).

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

Note: μ_0 is the assumed value of the population mean.

Step 2 Select a level of significance, α , depending on the seriousness of making a Type I error.

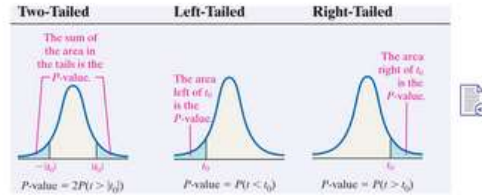
By Hand Step 3

Technology Step 3

Compute the test statistic.

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Use Table VII to approximate the P -value.



Step 4 If $P\text{-value} < \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

By Hand Step 3

Technology Step 3

Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value.

Notice that the procedure just presented requires that the population from which the sample was drawn be normal or that the sample size be large ($n \geq 30$). The procedure is robust, so minor departures from normality will not adversely affect the results of the test. However, if the data include outliers, the procedure should not be used.

We verify these assumptions by constructing normal probability plots (to assess normality) and boxplots (to discover whether there are outliers). If the normal probability plot indicates that the data do not come from a normal population or if the boxplot reveals outliers, nonparametric tests (not covered in this course) should be performed.

Before we look at a couple of examples, it is important to understand that we cannot find exact P -values using the t -distribution table (Table VII) because the table provides t -values only for certain areas. However, we can use the table to calculate lower and upper bounds on the P -value. To find exact P -values, we use statistical software or a graphing calculator with advanced statistical features.

EXAMPLE 1 Testing a Hypothesis about a Population Mean: Large Sample

Problem

The mean height of American males is 69.5 inches. The heights of the 43 male U.S. presidents (Washington through Obama) have a mean of 70.78 inches and a standard deviation of 2.77 inches. Treating the 43 presidents as a simple random sample, determine whether there is evidence to suggest that U.S. presidents are taller than the average American male. Use the $\alpha = 0.05$ level of significance. (Note: Grover Cleveland was elected to two nonconsecutive terms, so technically there have been 44 presidents of the United States.)

Approach

After verifying the requirements, follow Steps 1 through 5 for testing a hypothesis about a mean.

Solution

Video Solution Technology Step-By-Step



Result

The P -value of 0.0021 (by hand: $0.001 < P\text{-value} < 0.0025$) means that if the null hypothesis that $\mu = 69.5$ inches is true, we expect a sample mean of 70.78 inches or higher in about 2 out of 1000 samples. The results we obtained are not consistent with the assumption that the mean height of this population is 69.5 inches. Put another way, because the P -value is less than the level of significance, $\alpha = 0.05$ ($0.0021 < 0.05$), we reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that U.S. presidents are taller than the typical American male.

TI-83/84 Plus

StatCrunch

Hypothesis Tests Regarding a Population Mean

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 2:T-Test.
3. If the data are raw, highlight DATA; make sure that List is set to L1 and Freq is set to 1. If summary statistics are known, highlight STATS and enter the summary statistics. For the value of μ_0 , enter the value of the mean stated in the null hypothesis.
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw and press ENTER.

TI-83/84 Plus

StatCrunch

Hypothesis Tests Regarding a Population Mean

1. If necessary, enter the raw data into column var1. Name the column.
2. Select Stat, highlight T Statistics, highlight One Sample. Choose **With Data** if you have raw data, choose **With Summary** if you have summarized data.
3. If you chose **With Data**, highlight the column that contains the data in the "Select column(s):" drop-down menu. If you chose **With Summary**, enter the sample mean, sample standard deviation, and sample size. Choose the hypothesis test radio button. Enter the value of the mean stated in the null hypothesis and choose the direction of the alternative hypothesis from the drop-down menu. Click Compute!.

EXAMPLE 2 Testing a Hypothesis about a Population Mean: Small Sample

Problem

The “fun size” of a Snickers bar is supposed to weigh 20 grams. Because the penalty for selling candy bars under their advertised weight is severe, the manufacturer calibrates the machine so that the mean weight is 20.1 grams. The quality control engineer at Mars, the Snickers manufacturer, is concerned about the calibration. He obtains a random sample of 11 candy bars, weighs them, and obtains the data in Table 1. Should the machine be shut down and calibrated? Because shutting down the plant is expensive, he decides to conduct the test at the $\alpha = 0.01$ level of significance.

TABLE 1

19.68	20.66	19.56
19.98	20.65	19.61
20.55	20.36	21.02
21.50	19.74	

Data from Michael Carlisle, student at Joliet Junior College

Approach

After verifying the requirements, follow Steps 1 through 5 for testing a hypothesis about a mean.

Result

The P -value of 0.323 (by hand: $0.30 < P\text{-value} < 0.40$) means that if the null hypothesis that $\mu = 20.1$ grams is true, we expect 32 out of 100 samples to result in a sample mean as extreme as or more extreme than the one obtained. The results we obtained are not inconsistent with the assumption that the mean weight is 20.1 grams. In other words, the results obtained are not unusual under the assumption that the mean weight is 20.1 grams ($0.323 > 0.01$); so we do not reject the null hypothesis that $\mu = 20.1$ grams. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that the Snickers bars have a mean weight different from 20.1 grams at the $\alpha = 0.01$ level of significance. The machine should not be shut down.

10.3.15-T

December 7, 2016 10:04 AM

Researchers wanted to measure the effect of alcohol on the development of the hippocampal region in adolescents. The researchers randomly selected 13 adolescents with alcohol use disorders. They wanted to determine whether the hippocampal volumes in the alcoholic adolescents were less than the normal volume of 9.02 cm^3 . An analysis of the sample data revealed that the hippocampal volume is approximately normal with $\bar{x} = 8.46$ and $s = 0.9$. Conduct an appropriate test at the $\alpha = 0.05$ level of significance.

Choose the correct hypotheses.

$$H_0: \mu = 9.02$$

$$H_1: \mu < 9.02$$

Find the test statistic.

$$t_0 = -2.24$$

(Round to two decimal places as needed.)

Find the P-value.

$$\text{The P-value is } 0.022.$$

(Round to three decimal places as needed.)

What conclusion can be drawn?

- A. Do not reject H_0 . There is not sufficient evidence to conclude that the hippocampal volumes in the alcoholic adolescents are greater than the normal volume of 9.02 cm^3 .
- B. Reject H_0 . There is not sufficient evidence to conclude that the hippocampal volumes in the alcoholic adolescents are less than the normal volume of 9.02 cm^3 .
- C. Reject H_0 . There is sufficient evidence to conclude that the hippocampal volumes in the alcoholic adolescents are less than the normal volume of 9.02 cm^3 .
- D. Do not reject H_0 . There is sufficient evidence to conclude that the hippocampal volumes in the alcoholic adolescents are greater than the normal volume of 9.02 cm^3 .

10.3.2 Explain the Difference Between Statistical Significance and Practical Significance

December 7, 2016 10:08 AM

When a large sample size is used in a hypothesis test, the results may be **statistically significant** even though the difference between the sample statistic and mean stated in the null hypothesis may have no *practical significance*.

DEFINITION

Practical significance refers to the idea that although small differences between the statistic and parameter stated in the null hypothesis are statistically significant, the difference may not be large enough to cause concern or be considered important.

Results are **statistically significant** if the difference between the observed result and the statement made in the null hypothesis is unlikely to occur due to chance alone.

EXAMPLE 3 Statistical versus Practical Significance

Problem

According to the American Community Survey, the mean travel time to work in Collin County, Texas, is 27.6 minutes. The Department of Transportation reprogrammed all the traffic lights in Collin County in an attempt to reduce travel time. To determine whether there is evidence that travel time has decreased as a result of the reprogramming, the Department of Transportation obtains a random sample of 2500 commuters, records their travel time to work, and finds a sample mean of 27.3 minutes with a standard deviation of 8.5 minutes. Does this result suggest that travel time has decreased at the $\alpha = 0.05$ level of significance?

Approach

After verifying the requirements, follow Steps 1 through 5 for testing a hypothesis about a mean.

Solution

Video Solution



Technology Step-By-Step



Result

The P -value is 0.0389 (by hand: $0.025 < P\text{-value} < 0.05$). Because the P -value is less than the level of significance, $\alpha = 0.05$, we reject the null hypothesis that $\mu = 27.6$ minutes. There is sufficient evidence to conclude that the mean travel time to work has decreased. While the difference between 27.3 and 27.6 minutes is statistically significant, it has no practical meaning. After all, is 0.3 minute (18 seconds) really going to make anyone feel better about his or her commute to work?

The reason that the results from Example 3 were statistically significant had to do with the large sample size. The moral of the story is this:

Large sample sizes can lead to results that are statistically significant, whereas the difference between the statistic and parameter in the null hypothesis is not enough to be considered practically significant.



13.3.31-T

December 7, 2016 10:23 AM

10.3 Interactive Assignment-Kacey Howell

<https://xlitemprod.pearsoncmg.com/api/v1/print/math>

Student: Kacey Howell
Date: 12/7/16

Instructor: Matthew Nabity
Course: MTH 243: Introduction to
Probability and Statistics

Assignment: 10.3 Interactive
Assignment

6. A math teacher claims that she has developed a review course that increases the scores of students on the math portion of a college entrance exam. Based on data from the administrator of the exam, scores are normally distributed with $\mu = 514$. The teacher obtains a random sample of 2000 students, puts them through the review class, and finds that the mean math score of the 2000 students is 519 with a standard deviation of 111. Complete parts (a) through (d) below.

(a) State the null and alternative hypotheses. Let μ be the mean score. Choose the correct answer below.

- A. $H_0: \mu < 514, H_1: \mu > 514$
 B. $H_0: \mu = 514, H_1: \mu > 514$
 C. $H_0: \mu > 514, H_1: \mu \neq 514$
 D. $H_0: \mu = 514, H_1: \mu \neq 514$

(b) Test the hypothesis at the $\alpha = 0.10$ level of significance. Is a mean math score of 519 statistically significantly higher than 514? Conduct a hypothesis test using the P-value approach.

Find the test statistic.

$t_0 =$

(Round to two decimal places as needed.)

Find the P-value.

The P-value is .

(Round to three decimal places as needed.)

Is the sample mean statistically significantly higher?

- Yes
 No

(c) Do you think that a mean math score of 519 versus 514 will affect the decision of a school admissions administrator? In other words, does the increase in the score have any practical significance?

- Yes, because every increase in score is practically significant.
 No, because the score became only 0.97% greater.

(d) Test the hypothesis at the $\alpha = 0.10$ level of significance with $n = 375$ students. Assume that the sample mean is still 519 and the sample standard deviation is still 111. Is a sample mean of 519 significantly more than 514? Conduct a hypothesis test using the P-value approach.

Find the test statistic.

$t_0 =$

(Round to two decimal places as needed.)

Find the P-value.

The P-value is .

(Round to three decimal places as needed.)

Is the sample mean statistically significantly higher?

- No

What do you conclude about the impact of large samples on the P-value?

- A. As n increases, the likelihood of not rejecting the null hypothesis increases. However, large samples tend to overemphasize practically insignificant differences.
- B. As n increases, the likelihood of rejecting the null hypothesis increases. However, large samples tend to overemphasize practically insignificant differences.
- C. As n increases, the likelihood of not rejecting the null hypothesis increases. However, large samples tend to overemphasize practically significant differences.
- D. As n increases, the likelihood of rejecting the null hypothesis increases. However, large samples tend to overemphasize practically significant differences.

YOU ANSWERED: 2.014

Yes, because every increase in score is practically significant.

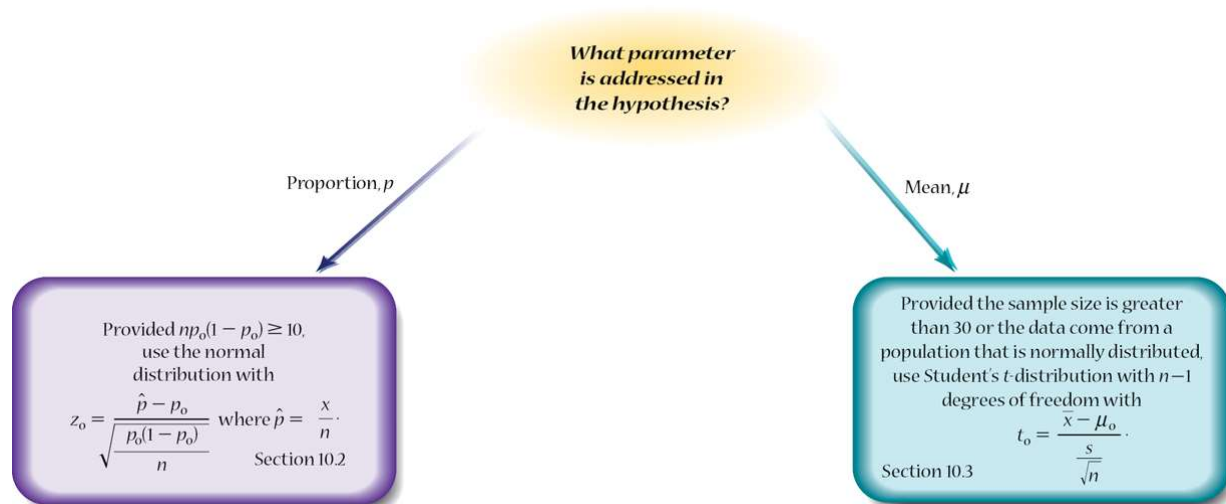
0.59

10.4 Putting it together: Which Procedure do I use?

December 9, 2016 08:53 AM

10.4.1 Determine the Appropriate Hypothesis Test to Perform

December 9, 2016 08:55 AM



10.4.7-T

December 9, 2016 09:01 AM

10.4.7-T

0.57 of 1 Point

Question Help

A psychologist obtains a random sample of 20 mothers in the first trimester of their pregnancy. The mothers are asked to play Mozart in the house at least 30 minutes each day until they give birth. After 5 years, the child is administered an IQ test. It is known that IQs are normally distributed with a mean of 100. If the IQs of the 20 children in the study result in a sample mean of 104.2 and sample standard deviation of 15, is there evidence that the children have higher IQs? Use the $\alpha = 0.10$ level of significance. Complete parts (a) through (d).

(a) Determine the null and alternative hypotheses.

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

(b) Calculate the P-value.

P-value = 0.113 (Round to three decimal places as needed.)

(c) State the conclusion for the test.

Choose the correct answer below.

- A. Do not reject H_0 because the P-value is greater than the $\alpha = 0.10$ level of significance.
- B. Do not reject H_0 because the P-value is less than the $\alpha = 0.10$ level of significance.
- C. Reject H_0 because the P-value is less than the $\alpha = 0.10$ level of significance.
- D. Reject H_0 because the P-value is greater than the $\alpha = 0.10$ level of significance.



10.4.7-T

0.57 of 1 Point

Question Help

A psychologist obtains a random sample of 20 mothers in the first trimester of their pregnancy. The mothers are asked to play Mozart in the house at least 30 minutes each day until they give birth. After 5 years, the child is administered an IQ test. It is known that IQs are normally distributed with a mean of 100. If the IQs of the 20 children in the study result in a sample mean of 104.2 and sample standard deviation of 15, is there evidence that the children have higher IQs? Use the $\alpha = 0.10$ level of significance. Complete parts (a) through (d).

(a) Determine the null and alternative hypotheses.

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

(b) Calculate the P-value.

P-value = 0.113 (Round to three decimal places as needed.)

(c) State the conclusion for the test.

Choose the correct answer below.

- A. Do not reject H_0 because the P-value is greater than the $\alpha = 0.10$ level of significance.
- B. Do not reject H_0 because the P-value is less than the $\alpha = 0.10$ level of significance.
- C. Reject H_0 because the P-value is less than the $\alpha = 0.10$ level of significance.
- D. Reject H_0 because the P-value is greater than the $\alpha = 0.10$ level of significance.

(d) State the conclusion in context of the problem.

There sufficient evidence at the $\alpha = 0.10$ level of significance to conclude that mothers who listen to Mozart have children with higher IQs.

