What Are Variables?

In statistics, a **variable** has two defining characteristics:

* A variable is an attribute that describes a person, place, thing, or idea.
* The value of the variable can "vary" from one entity to another.

For example, a person's *hair color* is a potential variable, which could have the value of "blond" for one person and "brunette" for another.

Qualitative vs. Quantitative Variables

Variables can be classified as **qualitative** (aka, categorical) or **quantitative** (aka, numeric).

* Qualitative. Qualitative variables take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of qualitative or categorical variables.
* Quantitative. Quantitative variables are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

In algebraic equations, quantitative variables are represented by symbols (e.g., *x*, *y*, or *z*).

Discrete vs. Continuous Variables

Quantitative variables can be further classified as **discrete** or **continuous**. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete variable.

Some examples will clarify the difference between discrete and continouous variables.

* Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
* Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.3 heads. Therefore, the number of heads must be a discrete variable.

Univariate vs. Bivariate Data

Statistical data are often classified according to the number of variables being studied.

* **Univariate data**. When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.
* **Bivariate data**. When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

Test Your Understanding

**Problem 1**

Which of the following statements are true?

I. All variables can be classified as quantitative or categorical variables.   
II. Categorical variables can be continuous variables.   
III. Quantitative variables can be discrete variables.

(A) I only   
(B) II only   
(C) III only   
(D) I and II   
(E) I and III

**Solution**

The correct answer is (E). All variables can be classified as quantitative or categorical variables. Discrete variables are indeed a category of quantitative variables. Categorical variables, however, are not numeric. Therefore, they cannot be classified as continuous variables.

Populations and Samples

The study of statistics revolves around the study of data sets. This lesson describes two important types of data sets - **populations** and **samples**. Along the way, we introduce simple random sampling, the main method used in this tutorial to select samples.

Population vs Sample

The main difference between a population and sample has to do with how observations are assigned to the data set.

* A **population** includes all of the [elements](http://stattrek.com/Help/Glossary.aspx?Target=element) from a set of data.
* A **sample** consists of one or more observations from the population.

Depending on the sampling method, a sample can have fewer observations than the population, the same number of observations, or more observations. More than one sample can be derived from the same population.

Other differences have to do with nomenclature, notation, and computations. For example,

* A a measurable characteristic of a population, such as a [mean](http://stattrek.com/Help/Glossary.aspx?Target=Mean) or [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation), is called a **parameter**; but a measurable characteristic of a sample is called a **statistic**.
* We will see in future lessons that the mean of a population is denoted by the symbol μ; but the mean of a sample is denoted by the symbol x.
* We will also learn in future lessons that the formula for the standard deviation of a population is different from the formula for the standard deviation of a sample.

What is Simple Random Sampling?

A **sampling method** is a procedure for selecting sample elements from a population. **Simple random sampling** refers to a sampling method that has the following properties.

* The population consists of *N* objects.
* The sample consists of *n* objects.
* All possible samples of *n* objects are equally likely to occur.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define a [confidence interval](http://stattrek.com/statistics/dictionary.aspx?definition=confidence_interval) around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the *N* population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects *n* numbers. Population members having the selected numbers are included in the sample.

Random Number Generator

In practice, the lottery method described above can be cumbersome, particularly with large sample sizes. As an alternative, use Stat Trek's Random Number Generator. With the Random Number Generator, you can select up to 1000 random numbers quickly and easily. This tool is provided at no cost - free!! To access the Random Number Generator, simply click on the button below. It can also be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Random Number Generator](http://stattrek.com/Tables/Random.aspx) |

Sampling With Replacement and Without Replacement

Suppose we use the lottery method described above to select a simple random sample. After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can selected only one time.

When a population element can be selected more than one time, we are **sampling with replacement**. When a population element can be selected only one time, we are **sampling without replacement**.

Test Your Understanding

**Problem 1**

Which of the following statements are true?

I. The mean of a population is denoted by x.   
II. Sample size is never bigger than population size.   
III. The population mean is a statistic.

(A) I only.   
(B) II only.   
(C) III only.   
(D) All of the above.   
(E) None of the above.

**Solution**

The correct answer is (E), none of the above.

The mean of a population is denoted by μ; not x. When sampling with replacement, sample size can be greater than population size. And the population mean is a *parameter*; the sample mean is a statistic.

The Mean and Median: Measures of Central Tendency

The **mean** and the **median** are summary measures used to describe the most "typical" value in a set of values.

Statisticians refer to the mean and median as **measures of central tendency**.

The Mean and the Median

The difference between the mean and median can be illustrated with an example. Suppose we draw a sample of five women and measure their weights. They weigh 100 pounds, 100 pounds, 130 pounds, 140 pounds, and 150 pounds.

* To find the **median**, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values. Thus, in the sample of five women, the median value would be 130 pounds; since 130 pounds is the middle weight.
* The **mean** of a sample or a population is computed by adding all of the observations and dividing by the number of observations. Returning to the example of the five women, the mean weight would equal (100 + 100 + 130 + 140 + 150)/5 = 620/5 = 124 pounds. In the general case, the mean can be calculated, using one of the following equations:

Population mean = μ = ΣX / N OR Sample mean = x = Σx / n

where ΣX is the sum of all the population observations, N is the number of population observations, Σx is the sum of all the sample observations, and n is the number of sample observations.

When statisticians talk about the mean of a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population), they use the Greek letter μ to refer to the mean score. When they talk about the mean of a [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample), statisticians use the symbol x to refer to the mean score.

The Mean vs. the Median

As measures of central tendency, the mean and the median each have advantages and disadvantages. Some pros and cons of each measure are summarized below.

* The median may be a better indicator of the most typical value if a set of scores has an **outlier**. An outlier is an extreme value that differs greatly from other values.
* However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

To illustrate these points, consider the following example. Suppose we examine a sample of 10 households to estimate the typical family income. Nine of the households have incomes between $20,000 and $100,000; but the tenth household has an annual income of $1,000,000,000. That tenth household is an outlier. If we choose a measure to estimate the income of a typical household, the mean will greatly over-estimate the income of a typical family (because of the outlier); while the median will not.

Effect of Changing Units

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of central tendency are affected when we change units.

* If you add a constant to every value, the mean and median increase by the same constant. For example, suppose you have a set of scores with a mean equal to 5 and a median equal to 6. If you add 10 to every score, the new mean will be 5 + 10 = 15; and the new median will be 6 + 10 = 16.
* Suppose you multiply every value by a constant. Then, the mean and the median will also be multiplied by that constant. For example, assume that a set of scores has a mean of 5 and a median of 6. If you multiply each of these scores by 10, the new mean will be 5 \* 10 = 50; and the new median will be 6 \* 10 = 60.

Test Your Understanding

**Problem 1**

Four friends take an IQ test. Their scores are 96, 100, 106, 114. Which of the following statements is true?

I. The mean is 103.   
II. The mean is 104.   
III. The median is 100.   
IV. The median is 106.

(A) I only   
(B) II only   
(C) III only   
(D) IV only   
(E) None is true

**Solution**

The correct answer is (B). The mean score is computed from the equation:

Mean score = Σx / n = (96 + 100 + 106 + 114) / 4 = 104

Since there are an even number of scores (4 scores), the median is the average of the two middle scores. Thus, the median is (100 + 106) / 2 = 103.

How to Measure Variability in Statistics

Statisticians use summary measures to describe the amount of variability or spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, and standard deviation.

The Range

The **range** is the difference between the largest and smallest values in a [set](http://stattrek.com/Help/Glossary.aspx?Target=Set) of values.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. For this set of numbers, the range would be 11 - 1 or 10.

The Interquartile Range (IQR)

The **interquartile range** (IQR) is a measure of variability, based on dividing a data set into [quartiles](http://stattrek.com/Help/Glossary.aspx?Target=Quartile).

Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

* Q1 is the "middle" value in the *first* half of the rank-ordered data set.
* Q2 is the [median](http://stattrek.com/Help/Glossary.aspx?Target=Median) value in the set.
* Q3 is the "middle" value in the *second* half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. Q1 is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, Q1 = (3 + 4)/2 or Q1 = 3.5. Q3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, Q3 = (6 + 7)/2 or Q3 = 6.5. The interquartile range is Q3 minus Q1, so IQR = 6.5 - 3.5 = 3.

An Alternative Definition for IQR

In some texts, the interquartile range is defined differently. It is defined as the difference between the largest and smallest values in the middle 50% of a set of data.

To compute an interquartile range using this definition, first remove observations from the lower quartile. Then, remove observations from the upper quartile. Then, from the remaining observations, compute the difference between the largest and smallest values.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. After we remove observations from the lower and upper quartiles, we are left with: 4, 5, 5, 6. The interquartile range (IQR) would be 6 - 4 = 2.

When the data set is large, the two definitions usually produce the same (or very close) results. However, when the data set is small, the definitions can produce different results.

The Variance

In a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population), **variance** is the average squared deviation from the population mean, as defined by the following formula:

σ2 = Σ ( Xi - μ )2 / N

where σ2 is the population variance, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

Observations from a [simple random sample](http://stattrek.com/Help/Glossary.aspx?Target=simple_random_sampling) can be used to estimate the variance of a population. For this purpose, sample variance is defined by slightly different formula, and uses a slightly different notation:

*s*2 = Σ ( xi - x )2 / ( n - 1 )

where *s*2 is the sample variance, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample. Using this formula, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate an unknown population variance, based on data from a simple random sample, this is the formula to use.

The Standard Deviation

The **standard deviation** is the square root of the variance. Thus, the standard deviation of a population is:

σ=σ 2 − − √ =∑(X i −μ) 2 N − − − − − − − − − − √ σ=σ2=∑(Xi-μ)2N

where σ is the population standard deviation, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

Statisticians often use [simple random samples](http://stattrek.com/statistics/dictionary.aspx?definition=simple_random_sampling) to estimate the standard deviation of a population, based on sample data. Given a simple random sample, the best estimate of the standard deviation of a population is:

s=s 2 − − √ =∑(x i −x − ) 2 n−1 − − − − − − − − − − −  ⎷   s=s2=∑(xi-x-)2n-1

where *s* is the sample standard deviation, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample.

Effect of Changing Units

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of variability are affected when we change units.

* If you add a constant to every value, the distance between values does not change. As a result, all of the measures of variability (range, interquartile range, standard deviation, and variance) remain the same.
* On the other hand, suppose you multiply every value by a constant. This has the effect of multiplying the range, interquartile range (IQR), and standard deviation by that constant. It has an even greater effect on the variance. It multiplies the variance by the square of the constant.

Test Your Understanding

**Problem 1**

A population consists of four observations: {1, 3, 5, 7}. What is the variance?

(A) 2   
(B) 4   
(C) 5   
(D) 6   
(E) None of the above

**Solution**

The correct answer is (C). First, we need to compute the population mean.

μ = ( 1 + 3 + 5 + 7 ) / 4 = 4

Then we plug all of the known values into formula for the variance of a population, as shown below:

σ2 = Σ ( Xi - μ )2 / N   
σ2 = [ ( 1 - 4 )2 + ( 3 - 4 )2 + ( 5 - 4 )2 + ( 7 - 4 )2 ] / 4   
σ2 = [ ( -3 )2 + ( -1 )2 + ( 1 )2 + ( 3 )2 ] / 4   
σ2 = [ 9 + 1 + 1 + 9 ] / 4 = 20 / 4 = 5

**Note:** Sometimes, students are unsure about whether the denominator in the formula for the variance should be N or (n - 1). We use N to compute the variance of a population, based on population data; and we use (n - 1) to estimate the variance of a population, based on sample data. In this problem, we are computing the variance of a population based on population data, so this solution uses N in the denominator.

**Problem 2**

A simple random sample consists of four observations: {1, 3, 5, 7}. Based on these sample observations, what is the best estimate of the standard deviation of the population?

(A) 2   
(B) 2.58   
(C) 6   
(D) 6.67   
(E) None of the above

**Solution**

The correct answer is (B). First, we need to compute the sample mean.

x = ( 1 + 3 + 5 + 7 ) / 4 = 4

Then we plug all of the known values into formula for the standard deviation of a sample, as shown below:

*s* = sqrt [ Σ ( xi - x )2 / ( n - 1 ) ]   
*s* = sqrt { [ ( 1 - 4 )2 + ( 3 - 4 )2 + ( 5 - 4 )2 + ( 7 - 4 )2 ] / ( 4 - 1 ) }   
*s* = sqrt { [ ( -3 )2 + ( -1 )2 + ( 1 )2 + ( 3 )2 ] / 3 }   
*s* = sqrt { [ 9 + 1 + 1 + 9 ] / 3 } = sqrt (20 / 3) = sqrt ( 6.67 ) = 2.58

**Note:** This problem asked us to estimate the standard deviation of a population, based on sample data. To do this, we used (n - 1) in the denominator of the standard deviation formula. If the problem had asked us to compute the standard deviation of a population based on population data, we would have used N in the denominator.

Measures of Position: Percentiles, Quartiles, z-Scores

Statisticians often talk about the **position** of a value, relative to other values in a [set](http://stattrek.com/Help/Glossary.aspx?Target=Set) of observations. The most common measures of position are percentiles, quartiles, and standard scores (aka, z-scores).

Percentiles

Assume that the [elements](http://stattrek.com/Help/Glossary.aspx?Target=Elements) in a data set are rank ordered from the smallest to the largest. The values that divide a rank-ordered set of elements into 100 equal parts are called **percentiles**.

An element having a percentile rank of Pi would have a greater value than i percent of all the elements in the set. Thus, the observation at the 50th percentile would be denoted P50, and it would be greater than 50 percent of the observations in the set. An observation at the 50th percentile would correspond to the [median](http://stattrek.com/Help/Glossary.aspx?Target=Median) value in the set.

Quartiles

**Quartiles** divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

Note the relationship between quartiles and percentiles. Q1 corresponds to P25, Q2 corresponds to P50, Q3 corresponds to P75. Q2 is the median value in the set.

Standard Scores (z-Scores)

A **standard score** (aka, a **z-score**) indicates how many [standard deviations](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) an element is from the mean. A standard score can be calculated from the following formula.

z = (X - μ) / σ

where z is the z-score, X is the value of the element, μ is the mean of the population, and σ is the standard deviation.

Here is how to interpret z-scores.

* A z-score less than 0 represents an element less than the mean.
* A z-score greater than 0 represents an element greater than the mean.
* A z-score equal to 0 represents an element equal to the mean.
* A z-score equal to 1 represents an element that is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean; etc.
* A z-score equal to -1 represents an element that is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean; etc.

Test Your Understanding

**Problem 1**

A national achievement test is administered annually to 3rd graders. The test has a mean score of 100 and a standard deviation of 15. If Jane's z-score is 1.20, what was her score on the test?

(A) 82   
(B) 88   
(C) 100   
(D) 112   
(E) 118

**Solution**

The correct answer is (E). From the z-score equation, we know

z = (X - μ) / σ

where z is the z-score, X is the value of the element, μ is the mean of the population, and σ is the standard deviation.

Solving for Jane's test score (X), we get

X = ( z \* σ) + 100 = ( 1.20 \* 15) + 100 = 18 + 100 = 118

Sets and Subsets

The lesson introduces the important topic of sets, a simple idea that recurs throughout the study of probability and statistics.

Set Definitions

* A **set** is a well-defined collection of objects.
* Each object in a set is called an **element** of the set.
* Two sets are **equal** if they have exactly the same elements in them.
* A set that contains no elements is called a **null set** or an **empty** set.
* If every element in Set *A* is also in Set *B*, then Set *A* is a **subset** of Set *B*.

Set Notation

* A set is usually denoted by a capital letter, such as *A, B,* or *C*.
* An element of a set is usually denoted by a small letter, such as *x, y,* or *z*.
* A set may be described by listing all of its elements enclosed in braces. For example, if Set *A* consists of the numbers 2, 4, 6, and 8, we may say: *A* = {2, 4, 6, 8}.
* The null set is denoted by {} or ∅.
* Sets may also be described by stating a rule. We could describe Set *A* from the previous example by stating: Set *A* consists of all the even single-digit positive integers.

Set Operations

Suppose we have four sets - W, X, Y, and Z. Let these sets be defined as follows: W = {2}; X = {1, 2}; Y= {2, 3, 4}; and Z = {1, 2, 3, 4}.

* The **union** of two sets is the set of elements that belong to one or both of the two sets. Thus, set Z is the union of sets X and Y.
* Symbolically, the union of X and Y is denoted by X ∪ Y.
* The **intersection** of two sets is the set of elements that are common to both sets. Thus, set W is the intersection of sets X and Y.
* Symbolically, the intersection of X and Y is denoted by X ∩ Y.

Sample Problems

1. Describe the set of vowels.   
   If A is the set of vowels, then A could be described as *A* = {a, e, i, o, u}.
2. Describe the set of positive integers.   
   Since it would be impossible to list *all* of the positive integers, we need to use a rule to describe this set. We might say *A* consists of all integers greater than zero.
3. Set *A* = {1, 2, 3} and Set *B* = {3, 2, 1}. Is Set *A* equal to Set *B*?   
   Yes. Two sets are equal if they have the same elements. The order in which the elements are listed does not matter.
4. What is the set of men with four arms?   
   Since all men have two arms at most, the set of men with four arms contains no elements. It is the null set (or empty set).
5. Set *A* = {1, 2, 3} and Set *B* = {1, 2, 4, 5, 6}. Is Set *A* a subset of Set *B*?   
   Set *A* would be a subset of Set *B* if every element from Set *A* were also in Set *B*. However, this is not the case. The number 3 is in Set *A*, but not in Set *B*. Therefore, Set *A* is not a subset of Set *B*.

What is a Statistical Experiment?

All **statistical experiments** have three things in common:

* The experiment can have more than one possible outcome.
* Each possible outcome can be specified in advance.
* The outcome of the experiment depends on chance.

A coin toss has all the attributes of a statistical experiment. There is more than one possible outcome. We can specify each possible outcome (i.e., heads or tails) in advance. And there is an element of chance, since the outcome is uncertain.

The Sample Space

* A **sample space** is a set of elements that represents all possible outcomes of a statistical experiment.
* A **sample point** is an element of a sample space.
* An **event** is a subset of a sample space - one or more sample points.

Types of events

* Two events are **mutually exclusive** if they have no sample points in common.
* Two events are **independent** when the occurrence of one does not affect the probability of the occurrence of the other.

Test Your Understanding

1. Suppose I roll a die. Is that a statistical experiment?   
   Yes. Like a coin toss, rolling dice is a statistical experiment. There is more than one possible outcome. We can specify each possible outcome in advance. And there is an element of chance.
2. When you roll a single die, what is the sample space.   
   The sample space is all of the possible outcomes - an integer between 1 and 6.
3. Which of the following are sample points when you roll a die - 3, 6, and 9?   
   The numbers 3 and 6 are sample points, because they are in the sample space. The number 9 is not a sample point, since it is outside the sample space; with one die, the largest number that you can roll is 6.
4. Which of the following sets represent an event when you roll a die?   
   A. {1}  
   B. {2, 4,}  
   C. {2, 4, 6}  
   D. All of the above   
     
   The correct answer is D. Remember that an event is a subset of a sample space. The sample space is any integer from 1 to 6. Each of the sets shown above is a subset of the sample space, so each represents an event.
5. Consider the events listed below. Which are mutually exclusive?   
     
   A. {1}  
   B. {2, 4,}  
   C. {2, 4, 6}   
   Two events are mutually exclusive, if they have no sample points in common. Events A and B are mutually exclusive, and Events A and C are mutually exclusive; since they have no points in common. Events B and C have common sample points, so they are not mutually exclusive.
6. Suppose you roll a die two times. Is each roll of the die an independent event?   
   Yes. Two events are independent when the occurrence of one has no effect on the probability of the occurrence of the other. Neither roll of the die affects the outcome of the other roll; so each roll of the die is independent.

Combinations, Permutations, and Counting Events

The solution to many statistical experiments involves being able to count the number of points in a sample space. Counting points can be hard, tedious, or both.

Fortunately, there are ways to make the counting task easier. This lesson focuses on three rules of counting that can save both time and effort - combinations, permutations, and event multiples.

Combinations

Sometimes, we want to count all of the possible ways that a single set of objects can be selected - without regard to the order in which they are selected.

* In general, *n* objects can be arranged in *n*(*n* - 1)(*n* - 2) ... (3)(2)(1) ways. This product is represented by the symbol *n*!, which is called **n factorial**. (By convention, 0! = 1.)
* A **combination** is a selection of all or part of a set of objects, *without* regard to the order in which they were selected. This means that XYZ is considered the same combination as ZYX.
* The number of combinations of *n* objects taken *r* at a time is denoted by nCr.

**Rule 1.** The number of combinations of *n* objects taken *r* at a time is

nCr = n(n - 1)(n - 2) ... (n - r + 1)/r! = n! / r!(n - r)!

**Example 1**  
How many different ways can you select 2 letters from the set of letters: X, Y, and Z? (Hint: In this problem, order is NOT important; i.e., XY is considered the same selection as YX.)

*Solution:* One way to solve this problem is to list all of the possible selections of 2 letters from the set of X, Y, and Z. They are: XY, XZ, and YZ. Thus, there are 3 possible combinations.

Another approach is to use Rule 1. Rule 1 tells us that the number of combinations is n! / r!(n - r)!. We have 3 distinct objects so n = 3. And we want to arrange them in groups of 2, so r = 2. Thus, the number of combinations is 3! / 2!(3 - 2)! or 3! /2!1!. This is equal to (3)(2)(1)/(2)(1)(1) = 3.

Combination and Permutation Calculator

Use Stat Trek's Combination and Permutation Calculator to (what else?) compute combinations and permutations. The calculator is free and easy to use. It can be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Combination and Permutation Calculator](http://stattrek.com/online-calculator/combinations-permutations.aspx) |

**Example 2**  
Five-card stud is a poker game, in which a player is dealt 5 cards from an ordinary deck of 52 playing cards. How many distinct poker hands could be dealt? (Hint: In this problem, the order in which cards are dealt is NOT important; For example, if you are dealt the ace, king, queen, jack, ten of spades, that is the same as being dealt the ten, jack, queen, king, ace of spades.)

*Solution:* For this problem, it would be impractical to list all of the possible poker hands. However, the number of possible poker hands can be easily calculated using Rule 1.

Rule 1 tells us that the number of combinations is n! / r!(n - r)!. We have 52 cards in the deck so n = 52. And we want to arrange them in groups of 5, so r = 5. Thus, the number of combinations is 52! / 5!(52 - 5)! or 52! / 5!47!. This is equal to 2,598,960 distinct poker hands.

Permutations

Often, we want to count all of the possible ways that a single set of objects can be arranged. For example, consider the letters X, Y, and Z. These letters can be arranged a number of different ways (XYZ, XZY, YXZ, etc.) Each of these arrangements is a permutation.

* A **permutation** is an arrangement of all or part of a set of objects, *with* regard to the order of the arrangement. This means that XYZ is considered a different permutation than ZYX.
* The number of permutations of *n* objects taken *r* at a time is denoted by nPr.

**Rule 2.** The number of permutations of *n* objects taken *r* at a time is

nPr = n(n - 1)(n - 2) ... (n - r + 1) = n! / (n - r)!

**Example 1**  
How many different ways can you arrange the letters X, Y, and Z? (Hint: In this problem, order is important; i.e., XYZ is considered a different arrangement than YZX.)

*Solution:* One way to solve this problem is to list all of the possible permutations of X, Y, and Z. They are: XYZ, XZY, YXZ, YZX, ZXY, and ZYX. Thus, there are 6 possible permutations.

Another approach is to use Rule 2. Rule 2 tells us that the number of permutations is n! / (n - r)!. We have 3 distinct objects so n = 3. And we want to arrange them in groups of 3, so r = 3. Thus, the number of permutations is 3! / (3 - 3)! or 3! / 0!. This is equal to (3)(2)(1)/1 = 6.

**Example 2**  
In horse racing, a trifecta is a type of bet. To win a trifecta bet, you need to specify the horses that finish in the top three spots in the exact order in which they finish. If eight horses enter the race, how many different ways can they finish in the top three spots?

*Solution:* Rule 2 tells us that the number of permutations is n! / (n - r)!. We have 8 horses in the race. so n = 8. And we want to arrange them in groups of 3, so r = 3. Thus, the number of permutations is 8! / (8 - 3)! or 8! / 5!. This is equal to (8)(7)(6) = 336 distinct trifecta outcomes. With 336 possible permutations, the trifecta is a difficult bet to win.

**Note:** Combinations and permutations are related according to the following formulas:

nPr = nCr \* r! and nCr = nPr / r!

Event Multiples

The third rule of counting deals with event multiples. An **event multiple** occurs when two or more *independent* events are grouped together. The third rule of counting helps us determine how many ways an event multiple can occur.

**Rule 3.** Suppose we have k independent events. Event 1 can be performed in n1 ways; Event 2, in n2 ways; and so on up to Event k (which can be performed in nk ways). The number of ways that these events can be performed together is equal to n1n2**. . .** nk ways.

**Example 1**  
How many sample points are in the sample space when a coin is flipped 4 times?

*Solution:* Each coin flip can have one of two outcomes - heads or tails. Therefore, the four coin flips can land in (2)(2)(2)(2) = 16 ways.

Event Counter

Use Stat Trek's Event Counter to count event multiples. The Event Counter is free and easy to use. It can be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Event Counter](http://stattrek.com/Tools/EventCounter.aspx) |

**Example 2**  
A business man has 4 dress shirts and 7 ties. How many different shirt/tie outfits can he create?

*Solution:* For each outfit, he can choose one of four shirts and one of seven ties. Therefore, the business man can create (4)(7) = 28 different shirt/tie outfits.

What is Probability?

The **probability** of an event refers to the likelihood that the event will occur.

How to Interpret Probability

Mathematically, the probability that an event will occur is expressed as a number between 0 and 1. Notationally, the probability of event A is represented by P(A).

* If P(A) equals zero, event A will almost definitely not occur.
* If P(A) is close to zero, there is only a small chance that event A will occur.
* If P(A) equals 0.5, there is a 50-50 chance that event A will occur.
* If P(A) is close to one, there is a strong chance that event A will occur.
* If P(A) equals one, event A will almost definitely occur.

In a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical%20experiment), the sum of probabilities for all possible outcomes is equal to one. This means, for example, that if an experiment can have three possible outcomes (A, B, and C), then P(A) + P(B) + P(C) = 1.

How to Compute Probability: Equally Likely Outcomes

Sometimes, a statistical experiment can have *n* possible outcomes, each of which is equally likely. Suppose a subset of *r* outcomes are classified as "successful" outcomes.

The probability that the experiment results in a successful outcome (S) is:

P(S) = ( Number of successful outcomes ) / ( Total number of equally likely outcomes ) = r / n

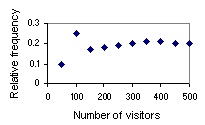
Consider the following experiment. An urn has 10 marbles. Two marbles are red, three are green, and five are blue. If an experimenter randomly selects 1 marble from the urn, what is the probability that it will be green?

In this experiment, there are 10 equally likely outcomes, three of which are green marbles. Therefore, the probability of choosing a green marble is 3/10 or 0.30.

How to Compute Probability: Law of Large Numbers

One can also think about the probability of an event in terms of its *long-run* relative frequency. The relative frequency of an event is the number of times an event occurs, divided by the total number of trials.

P(A) = ( Frequency of Event A ) / ( Number of Trials )



For example, a merchant notices one day that 5 out of 50 visitors to her store make a purchase. The next day, 20 out of 50 visitors make a purchase. The two relative frequencies (5/50 or 0.10 and 20/50 or 0.40) differ. However, summing results over many visitors, she might find that the probability that a visitor makes a purchase gets closer and closer 0.20.

The scatterplot (above right) shows the relative frequency as the number of trials (in this case, the number of visitors) increases. Over many trials, the relative frequency converges toward a stable value (0.20), which can be interpreted as the probability that a visitor to the store will make a purchase.

The idea that the relative frequency of an event will converge on the probability of the event, as the number of trials increases, is called the **law of large numbers**.

Test Your Understanding

**Problem**

A coin is tossed three times. What is the probability that it lands on heads *exactly* one time?

(A) 0.125   
(B) 0.250   
(C) 0.333   
(D) 0.375   
(E) 0.500

**Solution**

The correct answer is (D). If you toss a coin three times, there are a total of eight possible outcomes. They are: HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT. Of the eight possible outcomes, three have exactly one head. They are: HTT, THT, and TTH. Therefore, the probability that three flips of a coin will produce *exactly* one head is 3/8 or 0.375.

How to Solve Probability Problems

You can solve many simple probability problems just by knowing two simple rules:

* The probability of any sample point can range from 0 to 1.
* The sum of probabilities of all sample points in a [sample space](http://stattrek.com/Help/Glossary.aspx?Target=Sample_space) is equal to 1.

The following sample problems show how to apply these rules to find (1) the probability of a sample point and (2) the probability of an event.

Probability of a Sample Point

The **probability** of a [sample point](http://stattrek.com/Help/Glossary.aspx?Target=Sample_point) is a measure of the likelihood that the sample point will occur.

**Example 1**  
Suppose we conduct a simple [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment). We flip a coin one time. The coin flip can have one of two equally-likely outcomes - heads or tails. Together, these outcomes represent the sample space of our experiment. Individually, each outcome represents a sample point in the sample space. What is the probability of each sample point?

*Solution:* The sum of probabilities of all the sample points must equal 1. And the probability of getting a head is equal to the probability of getting a tail. Therefore, the probability of each sample point (heads or tails) must be equal to 1/2.

**Example 2**  
Let's repeat the experiment of Example 1, with a die instead of a coin. If we toss a fair die, what is the probability of each sample point?

*Solution:* For this experiment, the sample space consists of six sample points: {1, 2, 3, 4, 5, 6}. Each sample point has equal probability. And the sum of probabilities of all the sample points must equal 1. Therefore, the probability of each sample point must be equal to 1/6.

Probability of an Event

The probability of an [event](http://stattrek.com/Help/Glossary.aspx?Target=Event) is a measure of the likelihood that the event will occur. By convention, statisticians have agreed on the following rules.

* The probability of any event can range from 0 to 1.
* The probability of event A is the sum of the probabilities of all the sample points in event A.
* The probability of event A is denoted by P(A).

Thus, if event A were very unlikely to occur, then P(A) would be close to 0. And if event A were very likely to occur, then P(A) would be close to 1.

**Example 1**  
Suppose we draw a card from a deck of playing cards. What is the probability that we draw a spade?

*Solution:* The sample space of this experiment consists of 52 cards, and the probability of each sample point is 1/52. Since there are 13 spades in the deck, the probability of drawing a spade is

P(Spade) = (13)(1/52) = 1/4

**Example 2**  
Suppose a coin is flipped 3 times. What is the probability of getting two tails and one head?

*Solution:* For this experiment, the sample space consists of 8 sample points.

S = {TTT, TTH, THT, THH, HTT, HTH, HHT, HHH}

Each sample point is equally likely to occur, so the probability of getting any particular sample point is 1/8. The event "getting two tails and one head" consists of the following subset of the sample space.

A = {TTH, THT, HTT}

The probability of Event A is the sum of the probabilities of the sample points in A. Therefore,

P(A) = 1/8 + 1/8 + 1/8 = 3/8

Rules of Probability

Often, we want to compute the probability of an event from the known probabilities of other events. This lesson covers some important rules that simplify those computations.

Definitions and Notation

Before discussing the rules of probability, we state the following definitions:

* Two [events](http://stattrek.com/Help/Glossary.aspx?Target=Event) are **mutually exclusive** or **disjoint** if they cannot occur at the same time.
* The probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A, given Event B, is denoted by the symbol P(A|B).
* The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by P(A').
* The probability that Events A and B *both* occur is the probability of the **intersection** of A and B. The probability of the intersection of Events A and B is denoted by P(A ∩ B). If Events A and B are mutually exclusive, P(A ∩ B) = 0.
* The probability that Events A or B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by P(A ∪ B) .
* If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are **independent**.

Probability Calculator

Use the Probability Calculator to compute the probability of an event from the known probabilities of other events. The Probability Calculator is free and easy to use. It can be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Probability Calculator](http://stattrek.com/Tools/ProbabilityCalculator.aspx) |

Rule of Subtraction

In a [previous lesson](http://stattrek.com/AP-Statistics-3/Probability.aspx?Tutorial=stat), we learned two important properties of probability:

* The probability of an event ranges from 0 to 1.
* The sum of probabilities of all possible events equals 1.

The rule of subtraction follows directly from these properties.

**Rule of Subtraction** The probability that event A will occur is equal to 1 minus the probability that event A will not occur.

P(A) = 1 - P(A')

Suppose, for example, the probability that Bill will graduate from college is 0.80. What is the probability that Bill will not graduate from college? Based on the rule of subtraction, the probability that Bill will not graduate is 1.00 - 0.80 or 0.20.

Rule of Multiplication

The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events (Event A and Event B) both occur.

**Rule of Multiplication** The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.

P(A ∩ B) = P(A) P(B|A)

**Example**  
An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *without replacement* from the urn. What is the probability that both of the marbles are black?

*Solution:* Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

* In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, P(A) = 4/10.
* After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, P(B|A) = 3/9.

Therefore, based on the rule of multiplication:

P(A ∩ B) = P(A) P(B|A)   
P(A ∩ B) = (4/10) \* (3/9) = 12/90 = 2/15

Rule of Addition

The rule of addition applies to the following situation. We have two events, and we want to know the probability that either event occurs.

**Rule of Addition** The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

P(A ∪ B) = P(A) + P(B) - P(A ∩ B)

Note: Invoking the fact that P(A ∩ B) = P( A )P( B | A ), the Addition Rule can also be expressed as

P(A ∪ B) = P(A) + P(B) - P(A)P( B | A )

**Example**  
A student goes to the library. The probability that she checks out (a) a work of fiction is 0.40, (b) a work of non-fiction is 0.30, and (c) both fiction and non-fiction is 0.20. What is the probability that the student checks out a work of fiction, non-fiction, or both?

*Solution:* Let F = the event that the student checks out fiction; and let N = the event that the student checks out non-fiction. Then, based on the rule of addition:

P(F ∪ N) = P(F) + P(N) - P(F ∩ N)   
P(F ∪ N) = 0.40 + 0.30 - 0.20 = 0.50

Test Your Understanding

**Problem 1**

An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn *with replacement* from the urn. What is the probability that both of the marbles are black?

(A) 0.16   
(B) 0.32   
(C) 0.36   
(D) 0.40   
(E) 0.60

**Solution**

The correct answer is A. Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:

* In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, P(A) = 4/10.
* After the first selection, we replace the selected marble; so there are still 10 marbles in the urn, 4 of which are black. Therefore, P(B|A) = 4/10.

Therefore, based on the rule of multiplication:

P(A ∩ B) = P(A) P(B|A)   
P(A ∩ B) = (4/10)\*(4/10) = 16/100 = 0.16

**Problem 2**

A card is drawn randomly from a deck of ordinary playing cards. You win $10 if the card is a spade or an ace. What is the probability that you will win the game?

(A) 1/13   
(B) 13/52   
(C) 4/13   
(D) 17/52   
(E) None of the above.

**Solution**

The correct answer is C. Let S = the event that the card is a spade; and let A = the event that the card is an ace. We know the following:

* There are 52 cards in the deck.
* There are 13 spades, so P(S) = 13/52.
* There are 4 aces, so P(A) = 4/52.
* There is 1 ace that is also a spade, so P(S ∩ A) = 1/52.

Therefore, based on the rule of addition:

P(S ∪ A) = P(S) + P(A) - P(S ∩ A)   
P(S ∪ A) = 13/52 + 4/52 - 1/52 = 16/52 = 4/13

Bayes Theorem (aka, Bayes Rule)

Bayes' theorem (also known as Bayes' rule) is a useful tool for calculating [conditional probabilities](http://stattrek.com/Help/Glossary.aspx?Target=Conditional_probability). Bayes' theorem can be stated as follows:

**Bayes' theorem.** Let A1, A2, ... , An be a set of mutually exclusive events that together form the sample space S. Let B be any event from the same sample space, such that P(B) > 0. Then,

|  |  |
| --- | --- |
| P( Ak | B ) = | P( Ak∩ B )  P( A1 ∩ B ) + P( A2 ∩ B ) + . . . + P( An ∩ B ) |

Note: Invoking the fact that P( Ak ∩ B ) = P( Ak )P( B | Ak ), Baye's theorem can also be expressed as

|  |  |
| --- | --- |
| P( Ak | B ) = | P( Ak ) P( B | Ak )  P( A1 ) P( B | A1 ) + P( A2 ) P( B | A2 ) + . . . + P( An ) P( B | An ) |

Unless you are a world-class statiscian, Bayes' theorem (as expressed above) can be intimidating. However, it really is easy to use. The remainder of this lesson covers material that can help you understand when and how to apply Bayes' theorem effectively.

When to Apply Bayes' Theorem

Part of the challenge in applying Bayes' theorem involves recognizing the types of problems that warrant its use. You should consider Bayes' theorem when the following conditions exist.

* The [sample space](http://stattrek.com/Help/Glossary.aspx?Target=Sample_space) is partitioned into a [set](http://stattrek.com/Help/Glossary.aspx?Target=Set) of [mutually exclusive](http://stattrek.com/Help/Glossary.aspx?Target=Mutually_exclusive) events { A1, A2, . . . , An }.
* Within the sample space, there exists an [event](http://stattrek.com/Help/Glossary.aspx?Target=Event) B, for which P(B) > 0.
* The analytical goal is to compute a conditional probability of the form: P( Ak | B ).
* You know at least one of the two sets of probabilities described below.
  + P( Ak ∩ B ) for each Ak
  + P( Ak ) and P( B | Ak ) for each Ak

Bayes Rule Calculator

Use the Bayes Rule Calculator to compute conditional probability, when Bayes' theorem can be applied. The calculator is free, and it is easy to use. It can be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Bayes Rule Calculator](http://stattrek.com/Tools/BayesRuleCalculator.aspx) |

Sample Problem

Bayes' theorem can be best understood through an example. This section presents an example that demonstrates how Bayes' theorem can be applied effectively to solve statistical problems.

**Example 1**  
Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

*Solution:* The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. Notation for these events appears below.

* Event A1. It rains on Marie's wedding.
* Event A2. It does not rain on Marie's wedding.
* Event B. The weatherman predicts rain.

In terms of probabilities, we know the following:

* P( A1 ) = 5/365 =0.0136985 [It rains 5 days out of the year.]
* P( A2 ) = 360/365 = 0.9863014 [It does not rain 360 days out of the year.]
* P( B | A1 ) = 0.9 [When it rains, the weatherman predicts rain 90% of the time.]
* P( B | A2 ) = 0.1 [When it does not rain, the weatherman predicts rain 10% of the time.]

We want to know P( A1 | B ), the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes' theorem, as shown below.

|  |  |
| --- | --- |
| P( A1 | B ) = | P( A1 ) P( B | A1 )  P( A1 ) P( B | A1 ) + P( A2 ) P( B | A2 ) |
| P( A1 | B ) = | (0.014)(0.9) / [ (0.014)(0.9) + (0.986)(0.1) ] |
| P( A1 | B ) = | 0.111 |

Note the somewhat unintuitive result. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, there is a good chance that Marie will not get rained on at her wedding.

What is a Random Variable?

When the value of a [variable](http://stattrek.com/Help/Glossary.aspx?Target=Variable) is determined by a chance event, that variable is called a **random variable**.

Discrete vs. Continuous Random Variables

Random variables can be [discrete](http://stattrek.com/Help/Glossary.aspx?Target=Discrete%20variable) or [continuous](http://stattrek.com/Help/Glossary.aspx?Target=Continuous%20variable).

* **Discrete**. Within a range of numbers, discrete variables can take on only certain values. Suppose, for example, that we flip a coin and count the number of heads. The number of heads will be a value between zero and plus infinity. Within that range, though, the number of heads can be only certain values. For example, the number of heads can only be a whole number, not a fraction. Therefore, the number of heads is a discrete variable. And because the number of heads results from a random process - flipping a coin - it is a discrete random variable.
* **Continuous**. Continuous variables, in contrast, can take on any value within a range of values. For example, suppose we randomly select an individual from a population. Then, we measure the age of that person. In theory, his/her age can take on any value between zero and plus infinity, so age is a continuous variable. In this example, the age of the person selected is determined by a chance event; so, in this example, age is a continuous random variable.

Discrete Variables: Finite vs. Infinite

Some references state that continuous variables can take on an infinite number of values, but discrete variables cannot. This is incorrect.

* In some cases, discrete variables can take on only a finite number of values. For example, the number of aces dealt in a poker hand can take on only five values: 0, 1, 2, 3, or 4.
* In other cases, however, discrete variables can take on an infinite number of values. For example, the number of coin flips that result in heads could be infinitely large.

When comparing discrete and continuous variables, it is more correct to say that continuous variables can always take on an infinite number of values; whereas some discrete variables can take on an infinite number of values, but others cannot.

Test Your Understanding

**Problem 1**

Which of the following is a discrete random variable?

I. The average height of a randomly selected group of boys.   
II. The annual number of sweepstakes winners from New York City.   
III. The number of presidential elections in the 20th century.

(A) I only   
(B) II only   
(C) III only   
(D) I and II   
(E) II and III

**Solution**

The correct answer is B.

The annual number of sweepstakes winners results from a random process, but it can only be a whole number - not a fraction; so it is a discrete random variable. The average height of a randomly-selected group of boys could take on any value between the height of the smallest and tallest boys, so it is not a discrete variable. And the number of presidential elections in the 20th century does not result from a random process; so it is not a random variable.

What is a Random Variable?

When the value of a [variable](http://stattrek.com/Help/Glossary.aspx?Target=Variable) is determined by a chance event, that variable is called a **random variable**.

Discrete vs. Continuous Random Variables

Random variables can be [discrete](http://stattrek.com/Help/Glossary.aspx?Target=Discrete%20variable) or [continuous](http://stattrek.com/Help/Glossary.aspx?Target=Continuous%20variable).

* **Discrete**. Within a range of numbers, discrete variables can take on only certain values. Suppose, for example, that we flip a coin and count the number of heads. The number of heads will be a value between zero and plus infinity. Within that range, though, the number of heads can be only certain values. For example, the number of heads can only be a whole number, not a fraction. Therefore, the number of heads is a discrete variable. And because the number of heads results from a random process - flipping a coin - it is a discrete random variable.
* **Continuous**. Continuous variables, in contrast, can take on any value within a range of values. For example, suppose we randomly select an individual from a population. Then, we measure the age of that person. In theory, his/her age can take on any value between zero and plus infinity, so age is a continuous variable. In this example, the age of the person selected is determined by a chance event; so, in this example, age is a continuous random variable.

Discrete Variables: Finite vs. Infinite

Some references state that continuous variables can take on an infinite number of values, but discrete variables cannot. This is incorrect.

* In some cases, discrete variables can take on only a finite number of values. For example, the number of aces dealt in a poker hand can take on only five values: 0, 1, 2, 3, or 4.
* In other cases, however, discrete variables can take on an infinite number of values. For example, the number of coin flips that result in heads could be infinitely large.

When comparing discrete and continuous variables, it is more correct to say that continuous variables can always take on an infinite number of values; whereas some discrete variables can take on an infinite number of values, but others cannot.

Test Your Understanding

**Problem 1**

Which of the following is a discrete random variable?

I. The average height of a randomly selected group of boys.   
II. The annual number of sweepstakes winners from New York City.   
III. The number of presidential elections in the 20th century.

(A) I only   
(B) II only   
(C) III only   
(D) I and II   
(E) II and III

**Solution**

The correct answer is B.

The annual number of sweepstakes winners results from a random process, but it can only be a whole number - not a fraction; so it is a discrete random variable. The average height of a randomly-selected group of boys could take on any value between the height of the smallest and tallest boys, so it is not a discrete variable. And the number of presidential elections in the 20th century does not result from a random process; so it is not a random variable.

What is a Probability Distribution?

A **probability distribution** is a table or an equation that links each possible value that a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable) can assume with its probability of occurrence.

Discrete Probability Distributions

The probability distribution of a [discrete](http://stattrek.com/Help/Glossary.aspx?Target=Discrete%20variable) random variable can always be represented by a table. For example, suppose you flip a coin two times. This simple exercise can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of heads that result from the coin flips. The variable X can take on the values 0, 1, or 2; and X is a discrete random variable.

The table below shows the probabilities associated with each possible value of X. The probability of getting 0 heads is 0.25; 1 head, 0.50; and 2 heads, 0.25. Thus, the table is an example of a probability distribution for a discrete random variable.

|  |  |
| --- | --- |
| **Number of heads, x** | **Probability, P(x)** |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

**Note:** Given a probability distribution, you can find [cumulative probabilities](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability). For example, the probability of getting 1 or fewer heads [ P(X < 1) ] is P(X = 0) + P(X = 1), which is equal to 0.25 + 0.50 or 0.75.

Continuous Probability Distributions

The probability distribution of a [continuous](http://stattrek.com/Help/Glossary.aspx?Target=Continuous%20variable) random variable is represented by an equation, called the **probability density function** (pdf). All probability density functions satisfy the following conditions:

* The random variable Y is a function of X; that is, y = f(x).
* The value of y is greater than or equal to zero for all values of x.
* The total area under the curve of the function is equal to one.

The charts below show two continuous probability distributions. The chart on the left shows a probability density function described by the equation y = 1 over the range of 0 to 1 and y = 0 elsewhere. The chart on the right shows a probability density function described by the equation y = 1 - 0.5x over the range of 0 to 2 and y = 0 elsewhere. The area under the curve is equal to 1 for both charts.

|  |  |
| --- | --- |
| http://stattrek.com/Images/Sp26.jpg | http://stattrek.com/Images/Sp27.jpg |
| y = 1 | y = 1 - 0.5x |

The probability that a continuous random variable falls in the interval between *a* and *b* is equal to the area under the pdf curve between *a* and *b*. For example, in the first chart above, the shaded area shows the probability that the random variable X will fall between 0.6 and 1.0. That probability is 0.40. And in the second chart, the shaded area shows the probability of falling between 1.0 and 2.0. That probability is 0.25.

**Note:** With a continuous distribution, there are an infinite number of values between any two data points. As a result, the probability that a continuous random variable will assume a particular value is always zero. For example, in both of the above charts, the probability that variable X will equal *exactly* 0.4 is zero.

Test Your Understanding

**Problem 1**

The number of adults living in homes on a randomly selected city block is described by the following probability distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of adults, x** | 1 | 2 | 3 | 4 or more |
| **Probability, P(x)** | 0.25 | 0.50 | 0.15 | ??? |

What is the probability that 4 or more adults reside at a randomly selected home?

(A) 0.10   
(B) 0.15   
(C) 0.25   
(D) 0.50   
(E) 0.90

**Solution**

The correct answer is A. The sum of all the probabilities is equal to 1. Therefore, the probability that four or more adults reside in a home is equal to 1 - (0.25 + 0.50 + 0.15) or 0.10.

# Mean and Variance of Random Variables

Just like [variables](http://stattrek.com/Help/Glossary.aspx?Target=Variable) from a data set, [random variables](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable) are described by measures of central tendency (like the mean) and measures of variability (like variance). This lesson shows how to compute these measures for [discrete](http://stattrek.com/Help/Glossary.aspx?Target=Discrete%20variable) random variables.

## Mean of a Discrete Random Variable

The mean of the discrete random variable X is also called the **expected value** of X. Notationally, the expected value of X is denoted by E(X). Use the following formula to compute the mean of a discrete random variable.

E(X) = μx = Σ [ xi \* P(xi) ]

where xi is the value of the random variable for outcome i, μx is the mean of random variable X, and P(xi) is the probability that the random variable will be outcome i.

**Example 1**

In a recent little league softball game, each player went to bat 4 times. The number of hits made by each player is described by the following probability distribution.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Number of hits, x** | 0 | 1 | 2 | 3 | 4 |
| **Probability, P(x)** | 0.10 | 0.20 | 0.30 | 0.25 | 0.15 |

What is the mean of the probability distribution?

(A) 1.00   
(B) 1.75   
(C) 2.00   
(D) 2.25   
(E) None of the above.

**Solution**

The correct answer is E. The mean of the probability distribution is 2.15, as defined by the following equation.

E(X) = Σ [ xi \* P(xi) ]   
E(X) = 0\*0.10 + 1\*0.20 + 2\*0.30 + 3\*0.25 + 4\*0.15 = 2.15

## Median of a Discrete Random Variable

The median of a discrete random variable is the "middle" value. It is the value of X for which P(X < x) is greater than or equal to 0.5 and P(X > x) is greater than or equal to 0.5.

Consider the problem presented above in Example 1. In Example 1, the median is 2; because P(X < 2) is equal to 0.60, and P(X > 2) is equal to 0.70. The computations are shown below.

P(X < 2) = P(x=0) + P(x=1) + P(x=2) = 0.10 + 0.20 + 0.30 = 0.60   
  
P(X > 2) = P(x=2) + P(x=3) + P(x=4) = 0.30 + 0.25 + 0.15 = 0.70

## Variability of a Discrete Random Variable

The equation for computing the variance of a discrete random variable is shown below.

σ2 = Σ { [ xi - E(x) ]2 \* P(xi) }

where xi is the value of the random variable for outcome i, P(xi) is the probability that the random variable will be outcome i, E(x) is the expected value of the discrete random variable x.

**Example 2**

The number of adults living in homes on a randomly selected city block is described by the following probability distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of adults, x** | 1 | 2 | 3 | 4 |
| **Probability, P(x)** | 0.25 | 0.50 | 0.15 | 0.10 |

What is the standard deviation of the probability distribution?

(A) 0.50   
(B) 0.62   
(C) 0.79   
(D) 0.89   
(E) 2.10

**Solution**

The correct answer is D. The solution has three parts. First, find the expected value; then, find the variance; then, find the standard deviation. Computations are shown below, beginning with the expected value.

E(X) = Σ [ xi \* P(xi) ]   
E(X) = 1\*0.25 + 2\*0.50 + 3\*0.15 + 4\*0.10 = 2.10

Now that we know the expected value, we find the variance.

σ2 = Σ { [ xi - E(x) ]2 \* P(xi) }   
σ2 = (1 - 2.1)2 \* 0.25 + (2 - 2.1)2 \* 0.50 + (3 - 2.1)2 \* 0.15 + (4 - 2.1)2 \* 0.10   
σ2 = (1.21 \* 0.25) + (0.01 \* 0.50) + (0.81) \* 0.15) + (3.61 \* 0.10) = 0.3025 + 0.0050 + 0.1215 + 0.3610 = 0.79

And finally, the standard deviation is equal to the square root of the variance; so the standard deviation is sqrt(0.79) or 0.889.

Independent Random Variables

When a study involves pairs of [random variables](http://stattrek.com/Help/Glossary.aspx?Target=random_variable), it is often useful to know whether or not the random variables are independent. This lesson explains how to assess the independence of random variables.

Independence of Random Variables

If two random variables, X and Y, are **independent**, they satisfy the following conditions.

* P(x|y) = P(x), for all values of X and Y.
* P(x ∩ y) = P(x) \* P(y), for all values of X and Y.

The above conditions are equivalent. If either one is met, the other condition also met; and X and Y are independent. If either condition is not met, X and Y are **dependent**.

**Note:** If X and Y are independent, then the [correlation](http://stattrek.com/Help/Glossary.aspx?Target=Correlation) between X and Y is equal to zero.

Joint Probability Distributions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | X | | |
| 0 | 1 | 2 |
| Y | 3 | 0.1 | 0.2 | 0.2 |
| 4 | 0.1 | 0.2 | 0.2 |

The table on the right shows the joint probability distribution between two discrete random variables - X and Y.

In a joint probability distribution table, numbers in the cells of the table represent the probability that particular values of X and Y occur together. From this table, you can see that the probability that X=0 and Y=3 is 0.1; the probability that X=1 and Y=3 is 0.2; and so on.

You can use tables like this to figure out whether two discrete random variables are independent or dependent. Problem 1 below shows how.

Test Your Understanding

**Problem 1**

The table on the left shows the joint probability distribution between two random variables - X and Y; and the table on the right shows the joint probability distribution between two random variables - A and B.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | | X | | | | 0 | 1 | 2 | | Y | 3 | 0.1 | 0.2 | 0.2 | | 4 | 0.1 | 0.2 | 0.2 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | | A | | | | 0 | 1 | 2 | | B | 3 | 0.1 | 0.2 | 0.2 | | 4 | 0.2 | 0.2 | 0.1 | |

Which of the following statements are true?

I. X and Y are independent random variables.   
II. A and B are independent random variables.

(A) I only   
(B) II only   
(C) I and II   
(D) Neither statement is true.   
(E) It is not possible to answer this question, based on the information given.

**Solution**

The correct answer is A. The solution requires several computations to test the independence of random variables. Those computations are shown below.

X and Y are independent if P(x|y) = P(x), for all values of X and Y. From the probability distribution table, we know the following:

P(x=0) = 0.2; P(x=0 | y=3) = 0.2; P(x=0 | y = 4) = 0.2   
P(x=1) = 0.4; P(x=1 | y=3) = 0.4; P(x=1 | y = 4) = 0.4   
P(x=2) = 0.4; P(x=2 | y=3) = 0.4; P(x=2 | y = 4) = 0.4

Thus, P(x|y) = P(x), for all values of X and Y, which means that X and Y are independent. We repeat the same analysis to test the independence of A and B.

P(a=0) = 0.3; P(a=0 | b=3) = 0.2; P(a=0 | b = 4) = 0.4   
P(a=1) = 0.4; P(a=1 | b=3) = 0.4; P(a=1 | b = 4) = 0.4   
P(a=2) = 0.3; P(a=2 | b=3) = 0.4; P(a=2 | b = 4) = 0.2

Thus, P(a|b) is not equal to P(a), for all values of A and B. For example, P(a=0) = 0.3; but P(a=0 | b=3) = 0.2. This means that A and B are *not* independent.

# Combinations of Random Variables

Sometimes, it is necessary to add or subtract [random variables](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable). When this occurs, it is useful to know the mean and variance of the result.

**Recommendation:** Read the [sample problems](http://stattrek.com/random-variable/combination.aspx?Tutorial=stat#SampleProblems) at the end of the lesson. This lesson introduces some useful equations, and the sample problems show how to apply those equations.

## Sums and Differences of Random Variables: Effect on the Mean

Suppose you have two variables: X with a mean of μx and Y with a mean of μy. Then, the mean of the sum of these variables μx+y and the mean of the difference between these variables μx-y are given by the following equations.

μx+y = μx + μy and μx-y = μx - μy

The above equations for general variables also apply to random variables. If X and Y are random variables, then

E(X + Y) = E(X) + E(Y) and E(X - Y) = E(X) - E(Y)

where E(X) is the expected value (mean) of X, E(Y) is the expected value of Y, E(X + Y) is the expected value of X plus Y, and E(X - Y) is the expected value of X minus Y.

## Sums and Differences of Independent Random Variables: Effect on Variance

Suppose X and Y are independent random variables. Then, the variance of (X + Y) and the variance of (X - Y) are described by the following equations

Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)

where Var(X + Y) is the variance of the sum of X and Y, Var(X - Y) is the variance of the difference between X and Y, Var(X) is the variance of X, and Var(Y) is the variance of Y.

**Note:** The standard deviation (SD) is always equal to the square root of the variance (Var). Thus,

SD(X + Y) = sqrt[ Var(X + Y) ] and SD(X - Y) = sqrt[ Var(X - Y) ]

## Test Your Understanding

**Problem 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | X | | |
| 0 | 1 | 2 |
| Y | 3 | 0.1 | 0.2 | 0.2 |
| 4 | 0.1 | 0.2 | 0.2 |

The table on the right shows the joint probability distribution between two random variables - X and Y. (In a joint probability distribution table, numbers in the cells of the table represent the probability that particular values of X and Y occur together.)

What is the mean of the sum of X and Y?

(A) 1.2   
(B) 3.5   
(C) 4.5   
(D) 4.7   
(E) None of the above.

**Solution**

The correct answer is D. The solution requires three computations: (1) find the mean (expected value) of X, (2) find the mean (expected value) of Y, and (3) find the sum of the means. Those computations are shown below, beginning with the mean of X.

E(X) = Σ [ xi \* P(xi) ]   
E(X) = 0 \* (0.1 + 0.1) + 1 \* (0.2 + 0.2) + 2 \* (0.2 + 0.2) = 0 + 0.4 + 0.8 = 1.2

Next, we find the mean of Y.

E(Y) = Σ [ yi \* P(yi) ]   
E(Y) = 3 \* (0.1 + 0.2 + 0.2) + 4 \* (0.1 + 0.2 + 0.2) = (3 \* 0.5) + (4 \* 0.5) = 1.5 + 2 = 3.5

And finally, the mean of the sum of X and Y is equal to the sum of the means. Therefore,

E(X + Y) = E(X) + E(Y) = 1.2 + 3.5 = 4.7

**Note:** A similar approach is used to find differences between means. The difference between X and Y is E(X - Y) = E(X) - E(Y) = 1.2 - 3.5 = -2.3; and the difference between Y and X is E(Y - X) = E(Y) - E(X) = 3.5 - 1.2 = 2.3

**Problem 2**

Suppose X and Y are independent random variables. The variance of X is equal to 16; and the variance of Y is equal to 9. Let Z = X - Y.

What is the standard deviation of Z?

(A) 2.65   
(B) 5.00   
(C) 7.00   
(D) 25.0   
(E) It is not possible to answer this question, based on the information given.

**Solution**

The correct answer is B. The solution requires us to recognize that Variable Z is a combination of two independent random variables. As such, the variance of Z is equal to the variance of X plus the variance of Y.

Var(Z) = Var(X) + Var(Y) = 16 + 9 = 25

The standard deviation of Z is equal to the square root of the variance. Therefore, the standard deviation is equal to the square root of 25, which is 5.

Linear Transformations of Random Variables

Sometimes, it is necessary to apply a linear transformation to a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable). This lesson explains how to make a linear transformation and how to compute the mean and variance of the result.

What is a Linear Transformation?

A **linear transformation** is a change to a variable characterized by one or more of the following operations: adding a constant to the variable, subtracting a constant from the variable, multiplying the variable by a constant, and/or dividing the variable by a constant.

When a linear transformation is applied to a random variable, a new random variable is created. To illustrate, let X be a random variable, and let *m* and *b* be constants. Each of the following examples show how a linear transformation of X defines a new random variable Y.

* Adding a constant: Y = X + b
* Subtracting a constant: Y = X - b
* Multiplying by a constant: Y = mX
* Dividing by a constant: Y = X/m
* Multiplying by a constant and adding a constant: Y = mX + b
* Dividing by a constant and subtracting a constant: Y = X/m - b

**Note:** Suppose X and Z are variables, and the [correlation](http://stattrek.com/Help/Glossary.aspx?Target=Correlation) between X and Z is equal to *r*. If a new variable Y is created by applying a linear transformation to X, then the correlation between Y and Z will also equal *r*.

How Linear Transformations Affect the Mean and Variance

Suppose a linear transformation is applied to the random variable X to create a new random variable Y. Then, the mean and variance of the new random variable Y are defined by the following equations.

Y = mX + b and Var(Y) = m2 \* Var(X)

where *m* and *b* are constants, Y is the mean of Y, X is the mean of X, Var(Y) is the variance of Y, and Var(X) is the variance of X.

**Note:** The standard deviation (SD) of the transformed variable is equal to the square root of the variance. That is, SD(Y) = sqrt[ Var(Y) ].

Test Your Understanding

**Problem 1**

The average salary for an employee at Acme Corporation is $30,000 per year. This year, management awards the following bonuses to every employee.

* A Christmas bonus of $500.
* An incentive bonus equal to 10 percent of the employee's salary.

What is the mean bonus received by employees?

(A) $500   
(B) $3,000   
(C) $3,500   
(D) None of the above.   
(E) There is not enough information to answer this question.

**Solution**

The correct answer is C. To compute the bonus, management applies the following [linear transformation](http://stattrek.com/Help/Glossary.aspx?Target=Linear%20transformation) to the each employee's salary.

Y = mX + b   
Y = 0.10 \* X + 500

where Y is the transformed variable (the bonus), X is the original variable (the salary), *m* is the multiplicative constant 0.10, and *b* is the additive constant 500.

Since we know that the mean salary is $30,000, we can compute the mean bonus from the following equation.

Y = mX + b   
Y = 0.10 \* $30,000 + $500 = $3,500

**Problem 2**

The average salary for an employee at Acme Corporation is $30,000 per year, with a variance of 4,000,000. This year, management awards the following bonuses to every employee.

* A Christmas bonus of $500.
* An incentive bonus equal to 10 percent of the employee's salary.

What is the standard deviation of employee bonuses?

(A) $200   
(B) $3,000   
(C) $40,000   
(D) None of the above.   
(E) There is not enough information to answer this question.

**Solution**

The correct answer is A. To compute the bonus, management applies the following [linear transformation](http://stattrek.com/Help/Glossary.aspx?Target=Linear%20transformation) to the each employee's salary.

Y = mX + b   
Y = 0.10 \* X + 500

where Y is the transformed variable (the bonus), X is the original variable (the salary), *m* is the multiplicative constant 0.10, and *b* is the additive constant 500.

Since we know the variance of employee salaries, we can compute the variance of employee bonuses from the following equation.

Var(Y) = m2 \* Var(X) = (0.1)2 \* 4,000,000 = 40,000

where Var(Y) is the variance of employee bonuses, and Var(X) is the variance of employee salaries.

And finally, since the standard deviation is equal to the square root of the variance, the standard deviation of employee bonuses is equal to the square root of 40,000 or $200.

Simple Random Sampling

Simple random sampling is the most widely-used probability sampling method, probably because it is easy to implement and easy to analyze.

Key Definitions

To understand simple random sampling, you need to first understand a few key definitions.

* The total [set](http://stattrek.com/Help/Glossary.aspx?Target=Set) of observations that can be made is called the **population**.
* A **sample** is a set of observations drawn from a population.
* A **parameter** is a measurable characteristic of a population, such as a [mean](http://stattrek.com/Help/Glossary.aspx?Target=Mean) or [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation).
* A **statistic** is a measurable characteristic of a sample, such as a mean or standard deviation.
* A **sampling method** is a procedure for selecting sample elements from a population.
* A **random number** is a number determined totally by chance, with no predictable relationship to any other number.
* A **random number table** is a list of numbers, composed of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Numbers in the list are arranged so that each digit has no predictable relationship to the digits that preceded it or to the digits that followed it. In short, the digits are arranged randomly. The numbers in a random number table are random numbers.

Simple Random Sampling

**Simple random sampling** refers to a sampling method that has the following properties.

* The population consists of *N* objects.
* The sample consists of *n* objects.
* All possible samples of *n* objects are equally likely to occur.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define a [confidence interval](http://stattrek.com/statistics/dictionary.aspx?definition=confidence_interval) around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the *N* population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects *n* numbers. Population members having the selected numbers are included in the sample.

Random Number Generator

In practice, the lottery method described above can be cumbersome, particularly with large sample sizes. As an alternative, use Stat Trek's Random Number Generator. With the Random Number Generator, you can select up to random numbers quickly and easily. This tool is provided at no cost - free!! To access the Random Number Generator, simply click on the button below. It can also be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

Sampling With Replacement and Without Replacement

Suppose we use the lottery method described above to select a simple random sample. After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can selected only one time.

When a population element can be selected more than one time, we are **sampling with replacement**. When a population element can be selected only one time, we are **sampling without replacement**.

Measures of Central Tendency

Researchers are often interested in defining a value that best describes some attribute of the population. Often this attribute is a measure of central tendency or a proportion.

Measures of Central Tendency

Several different measures of central tendency are defined below.

* The **mode** is the most frequently appearing value in the population or sample. Suppose we draw a sample of five women and measure their weights. They weigh 100 pounds, 100 pounds, 130 pounds, 140 pounds, and 150 pounds. Since more women weigh 100 pounds than any other weight, the mode would equal 100 pounds.
* To find the **median**, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values. Thus, in the sample of five women, the median value would be 130 pounds; since 130 pounds is the middle weight.
* The **mean** of a sample or a population is computed by adding all of the observations and dividing by the number of observations. Returning to the example of the five women, the mean weight would equal (100 + 100 + 130 + 140 + 150)/5 = 620/5 = 124 pounds.

Proportions and Percentages

When the focus is on the degree to which a population possesses a particular attribute, the measure of interest is a percentage or a proportion.

* A **proportion** refers to the fraction of the total that possesses a certain attribute. For example, we might ask what proportion of women in our sample weigh less than 135 pounds. Since 3 women weigh less than 135 pounds, the proportion would be 3/5 or 0.60.
* A **percentage** is another way of expressing a proportion. A percentage is equal to the proportion times 100. In our example of the five women, the percent of the total who weigh less than 135 pounds would be 100 \* (3/5) or 60 percent.

Notation

Of the various measures, the mean and the proportion are most important. The notation used to describe these measures appears below:

* X: Refers to a population mean.
* x: Refers to a sample mean.
* P: The proportion of elements in the population that has a particular attribute.
* p: The proportion of elements in the *sample* that has a particular attribute.
* Q: The proportion of elements in the population that does not have a specified attribute. Note that Q = 1 - P.
* q: The proportion of elements in the *sample* that does not have a specified attribute. Note that q = 1 - p.

Note that capital letters refer to population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter), and lower-case letters refer to sample [statistics](http://stattrek.com/Help/Glossary.aspx?Target=Statistic).

How to Measure Variability in a Data Set

In this lesson, we discuss three measures that are used to quantify the amount of variation in a data set - the range, the variance, and the standard deviation.

For example, consider a population of elements {5, 5 ,5, 5}. Here, each of the values in the data set are equal, so there is no variation. The set {3, 5, 5, 7}, on the other hand, has some variation since some some elements in the data set have different values.

Notation

The following notation is helpful, when we talk about variability.

* σ2: The variance of the population.
* σ: The standard deviation of the population.
* s2: The variance of the sample.
* s: The standard deviation of the sample.
* μ: The population [mean](http://stattrek.com/Help/Glossary.aspx?Target=Mean).
* x: The sample mean.
* N: Number of observations in the population.
* n: Number of observations in the sample.
* P: The proportion of elements in the population that has a particular attribute.
* p: The proportion of elements in the sample that has a particular attribute.
* Q: The proportion of elements in the population that does not have a specified attribute. Note that Q = 1 - P.
* q: The proportion of elements in the sample that does not have a specified attribute. Note that q = 1 - p.

Note that capital letters refer to population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter), and lower-case letters refer to sample [statistics](http://stattrek.com/Help/Glossary.aspx?Target=Statistic).

The Range

The **range** is the simplest measure of variation. It is difference between the biggest and smallest random variable.

Range = Maximum value - Minimum value

Therefore, the range of the four random variables (3, 5, 5, 7} would be 7 minus 3 or 4.

Variance of the Mean

It is important to distinguish between the variance of a population mean and the variance of a sample mean. They have different notation, and they are computed differently. The variance of a population mean is denoted by σ2; and the variance of a sample mean, by *s*2.

The **variance** of a population mean is the average squared deviation from the population mean, as defined by the following formula:

σ2 = Σ ( Xi - μ )2 / N

where σ2 is the population variance, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

The variance of a sample mean is defined by slightly different formula:

*s*2 = Σ ( xi - x )2 / ( n - 1 )

where *s*2 is the sample variance, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample. If you are working with a simple random sample, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you want to estimate the unknown population variance, based on known data from a simple random sample, use this formula.

**Example 1**  
  
A population consists of four observations: {1, 3, 5, 7}. What is the variance?

*Solution:* First, we need to compute the population mean.

μ = ( 1 + 3 + 5 + 7 ) / 4 = 4

Then we plug all of the known values in to formula for the variance of a population, as shown below:

σ2 = Σ ( Xi - μ )2 / N

σ2 = [ ( 1 - 4 )2 + ( 3 - 4 )2 + ( 5 - 4 )2 + ( 7 - 4 )2 ] / 4

σ2 = [ ( -3 )2 + ( -1 )2 + ( 1 )2 + ( 3 )2 ] / 4

σ2 = [ 9 + 1 + 1 + 9 ] / 4 = 20 / 4 = 5

**Example 2**  
  
A simple random sample consists of four observations: {1, 3, 5, 7}. What is the best estimate of the population variance?

*Solution:* This problem is handled exactly like the previous problem, except that we use the formula for calculating sample variance, rather than the formula for calculating population variance.

*s*2 = Σ ( xi - x )2 / ( n - 1 )

*s*2 = [ ( 1 - 4 )2 + ( 3 - 4 )2 + ( 5 - 4 )2 + ( 7 - 4 )2 ] / ( 4 - 1 )

*s*2 = [ ( -3 )2 + ( -1 )2 + ( 1 )2 + ( 3 )2 ] / 3

*s*2 = [ 9 + 1 + 1 + 9 ] / 3 = 20 / 3 = 6.667

Standard Deviation of the Mean

The **standard deviation** is the square root of the variance. It is important to distinguish between the standard deviation of a population and the standard deviation of a sample. They have different notation, and they are computed differently. The standard deviation of a population is denoted by σ; and the standard deviation of a sample, by *s*.

The standard deviation of a population mean is defined by the following formula:

σ=σ 2 − − √ =∑(X i −μ) 2 N − − − − − − − − − − √ σ=σ2=∑(Xi-μ)2N

where σ is the population standard deviation, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

The standard deviation of a sample mean is defined by slightly different formula:

s=s 2 − − √ =∑(x i −x − ) 2 n−1 − − − − − − − − − − −  ⎷   s=s2=∑(xi-x-)2n-1

where *s* is the sample standard deviation, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample.

Variance of a Proportion

The variance formulas introduced in the previous section can be used with confidence for any random variable - even proportions. However, for proportions the formulas can be expressed in a form that is easier to compute.

When all of the elements of the population are known, the variance of a population proportion is defined by the following formula:

σ2 = PQ

where P is the population proportion and Q equals 1 - P.

When the population proportion is estimated from sample data, the variance of the sample proportion is estimated by slightly different formula:

*s*2 = pq

where p is the sample estimate of the true proportion, and q is equal to 1 - p. Given a simple random sample, this sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate the unknown population variance, based on known data from a simple random sample, this is the formula to use.

Standard Deviation of a Proportion

The standard deviation of a proportion is the square root of the variance of the proportion. Thus, the standard deviation of a population proportion is:

σ=σ 2 − − √ =PQ − − − √ σ=σ2=PQ

where P is the population proportion and Q equals 1 - P.

And, using sample data, the standard deviation of a population proportion can be estimated from the following formula:

s=s 2 − − √ =pq − − √ s=s2=pq

where p is the sample proportion and q equals 1 - p.

# Sampling Distributions

Suppose that we draw all possible samples of size *n* from a given population. Suppose further that we compute a [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic) (e.g., a mean, proportion, standard deviation) for each sample. The [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of this statistic is called a **sampling distribution**. And the standard deviation of this statistic is called the **standard error**.

## Variability of a Sampling Distribution

The variability of a sampling distribution is measured by its [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) or its [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation). The variability of a sampling distribution depends on three factors:

* N: The number of observations in the population.
* n: The number of observations in the sample.
* The way that the random sample is chosen.

If the population size is much larger than the sample size, then the sampling distribution has roughly the same standard error, whether we sample [with](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_with_replacement) or [without replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_without_replacement). On the other hand, if the sample represents a significant fraction (say, 1/20) of the population size, the standard error will be meaningfully smaller, when we sample without replacement.

## Sampling Distribution of the Mean

Suppose we draw all possible samples of size *n* from a population of size *N*. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

We know the following about the sampling distribution of the mean. The mean of the sampling distribution (μx) is equal to the mean of the population (μ). And the standard error of the sampling distribution (σx) is determined by the standard deviation of the population (σ), the population size (N), and the sample size (n). These relationships are shown in the equations below:

μx = μ and σx = [ σ / sqrt(n) ] \* sqrt[ (N - n ) / (N - 1) ]

In the standard error formula, the factor sqrt[ (N - n ) / (N - 1) ] is called the finite population correction or fpc. When the population size is very large relative to the sample size, the fpc is approximately equal to one; and the standard error formula can be approximated by:

σx = σ / sqrt(n).

You often see this "approximate" formula in introductory statistics texts. As a general rule, it is safe to use the approximate formula when the sample size is no bigger than 1/20 of the population size.

## Sampling Distribution of the Proportion

In a population of size *N*, suppose that the probability of the occurrence of an event (dubbed a "success") is P; and the probability of the event's non-occurrence (dubbed a "failure") is Q. From this population, suppose that we draw all possible samples of size *n*. And finally, within each sample, suppose that we determine the proportion of successes *p* and failures *q*. In this way, we create a sampling distribution of the proportion.

We find that the mean of the sampling distribution of the proportion (μp) is equal to the probability of success in the population (P). And the standard error of the sampling distribution (σp) is determined by the standard deviation of the population (σ), the population size, and the sample size. These relationships are shown in the equations below:

μp = P

σp = [ σ / sqrt(n) ] \* sqrt[ (N - n ) / (N - 1) ]

σp = sqrt[ PQ/n ] \* sqrt[ (N - n ) / (N - 1) ]

where σ = sqrt[ PQ ].

Like the formula for the standard error of the mean, the formula for the standard error of the proportion uses the finite population correction, sqrt[ (N - n ) / (N - 1) ]. When the population size is very large relative to the sample size, the fpc is approximately equal to one; and the standard error formula can be approximated by:

σp = sqrt[ PQ/n ]

You often see this "approximate" formula in introductory statistics texts. As a general rule, it is safe to use the approximate formula when the sample size is no bigger than 1/20 of the population size.

## Central Limit Theorem

The **central limit theorem** states that the sampling distribution of the mean of any [independent](http://stattrek.com/help/glossary.aspx?target=independent), [random variable](http://stattrek.com/help/glossary.aspx?target=random_variable) will be normal or nearly normal, if the sample size is large enough.

How large is "large enough"? The answer depends on two factors.

* Requirements for accuracy. The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.
* The shape of the underlying population. The more closely the original population resembles a normal distribution, the fewer sample points will be required.

In practice, some statisticians say that a sample size of 30 is large enough when the population distribution is roughly bell-shaped. Others recommend a sample size of at least 40. But if the original population is distinctly not normal (e.g., is badly skewed, has multiple peaks, and/or has outliers), researchers like the sample size to be even larger.

## T-Distribution vs. Normal Distribution

The t distribution and the normal distribution can both be used with statistics that have a bell-shaped distribution. This suggests that we might use either the t-distribution or the normal distribution to analyze sampling distributions. Which should we choose?

Guidelines exist to help you make that choice. Some focus on the population standard deviation.

* If the population standard deviation is known, use the normal distribution
* If the population standard deviation is unknown, use the t-distribution.

Other guidelines focus on sample size.

* If the sample size is large, use the normal distribution. (See the discussion above in the section on the Central Limit Theorem to understand what is meant by a "large" sample.)
* If the sample size is small, use the t-distribution.

In practice, researchers employ a mix of the above guidelines. On this site, we use the normal distribution when the population standard deviation is known and the sample size is large. We might use either distribution when standard deviation is unknown and the sample size is very large. We use the t-distribution when the sample size is small, unless the underlying distribution is not normal. The t distribution should not be used with small samples from populations that are not approximately normal.

## Test Your Understanding

In this section, we offer two examples that illustrate how sampling distributions are used to solve commom statistical problems. In each of these problems, the population sample size is known; and the sample size is large. So you should use the Normal Distribution Calculator, rather than the t-Distribution Calculator, to compute probabilities for these problems.

## Normal Distribution Calculator

The normal calculator solves common statistical problems, based on the normal distribution. The calculator computes cumulative probabilities, based on three simple inputs. Simple instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Normal Calculator](http://stattrek.com/Tables/Normal.aspx) |

**Example 1**  
  
Assume that a school district has 10,000 6th graders. In this district, the average weight of a 6th grader is 80 pounds, with a standard deviation of 20 pounds. Suppose you draw a random sample of 50 students. What is the probability that the average weight of a sampled student will be less than 75 pounds?

*Solution:* To solve this problem, we need to define the sampling distribution of the mean. Because our sample size is greater than 30, the Central Limit Theorem tells us that the sampling distribution will approximate a normal distribution.

To define our normal distribution, we need to know both the mean of the sampling distribution and the standard deviation. Finding the mean of the sampling distribution is easy, since it is equal to the mean of the population. Thus, the mean of the sampling distribution is equal to 80.

The standard deviation of the sampling distribution can be computed using the following formula.

σx = [ σ / sqrt(n) ] \* sqrt[ (N - n ) / (N - 1) ]   
σx = [ 20 / sqrt(50) ] \* sqrt[ (10,000 - 50 ) / (10,000 - 1) ] = (20/7.071) \* (0.995) = 2.81

Let's review what we know and what we want to know. We know that the sampling distribution of the mean is normally distributed with a mean of 80 and a standard deviation of 2.82. We want to know the probability that a sample mean is less than or equal to 75 pounds.

Because we know the population standard deviation and the sample size is large, we'll use the normal distribution to find probability. To solve the problem, we plug these inputs into the Normal Probability Calculator: mean = 80, standard deviation = 2.81, and normal random variable = 75. The Calculator tells us that the probability that the average weight of a sampled student is less than 75 pounds is equal to 0.038.

Note: Since the population size is more than 20 times greater than the sample size, we could have used the "approximate" formula σx = [ σ / sqrt(n) ] to compute the standard error. Had we done that, we would have found a standard error equal to [ 20 / sqrt(50) ] or 2.83.

**Example 2**  
  
Find the probability that of the next 120 births, no more than 40% will be boys. Assume equal probabilities for the births of boys and girls. Assume also that the number of births in the population (N) is very large, essentially infinite.

*Solution:* The Central Limit Theorem tells us that the proportion of boys in 120 births will be approximately normally distributed.

The mean of the sampling distribution will be equal to the mean of the population distribution. In the population, half of the births result in boys; and half, in girls. Therefore, the probability of boy births in the population is 0.50. Thus, the mean proportion in the sampling distribution should also be 0.50.

The standard deviation of the sampling distribution (i.e., the standard error) can be computed using the following formula.

σp = sqrt[ PQ/n ] \* sqrt[ (N - n ) / (N - 1) ]

Here, the finite population correction is equal to 1.0, since the population size (N) was assumed to be infinite. Therefore, standard error formula reduces to:

σp = sqrt[ PQ/n ]   
σp = sqrt[ (0.5)(0.5)/120 ] = sqrt[0.25/120 ] = 0.04564

Let's review what we know and what we want to know. We know that the sampling distribution of the proportion is normally distributed with a mean of 0.50 and a standard deviation of 0.04564. We want to know the probability that no more than 40% of the sampled births are boys.

Because we know the population standard deviation and the sample size is large, we'll use the normal distribution to find probability. To solve the problem, we plug these inputs into the Normal Probability Calculator: mean = .5, standard deviation = 0.04564, and the normal random variable = .4. The Calculator tells us that the probability that no more than 40% of the sampled births are boys is equal to 0.014.

Note: This problem can also be treated as a [binomial experiment](http://stattrek.com/Help/Glossary.aspx?Target=Binomial%20experiment). Elsewhere, we showed [how to analyze a binomial experiment](http://stattrek.com/Lesson2/Binomial.aspx?Tutorial=AP). The binomial experiment is actually the more exact analysis. It produces a probability of 0.018 (versus a probability of 0.14 that we found using the normal distribution). Without a computer, the binomial approach is computationally demanding. Therefore, many statistics texts emphasize the approach presented above, which uses the normal distribution to approximate the binomial.

Difference Between Proportions

Statistics problems often involve comparisons between two independent sample proportions. This lesson explains how to compute probabilities associated with differences between proportions.

Difference Between Proportions: Theory

Suppose we have two [populations](http://stattrek.com/Help/Glossary.aspx?Target=Population) with proportions equal to P1 and P2. Suppose further that we take all possible [samples](http://stattrek.com/Help/Glossary.aspx?Target=Sample) of size n1 and n2. And finally, suppose that the following assumptions are valid.

* The size of each population is large relative to the sample drawn from the population. That is, N1 is large relative to n1, and N2 is large relative to n2. (In this context, populations are considered to be large if they are at least 20 times bigger than their sample.)
* The samples from each population are big enough to justify using a normal distribution to model differences between proportions. The sample sizes will be big enough when the following conditions are met: n1P1 > 10, n1(1 -P1) > 10, n2P2 > 10, and n2(1 - P2) > 10. (This criterion requires that at least 40 observations be sampled from each population. When P1 or P1 is more extreme than 0.5, even more observations are required.)
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent); that is, observations in population 1 are not affected by observations in population 2, and vice versa.

Given these assumptions, we know the following.

* The set of differences between sample proportions will be normally distributed. We know this from the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem).
* The [expected value](http://stattrek.com/Help/Glossary.aspx?Target=Expected%20value) of the difference between all possible sample proportions is equal to the difference between population proportions. Thus, E(p1 - p2) = P1 - P2.
* The standard deviation of the difference between sample proportions (σd) is approximately equal to:

σd = sqrt{ [P1(1 - P1) / n1] + [P2(1 - P2) / n2] }

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

σ2d = σ2P1 - P2 = σ21 + σ22

If the populations N1 and N2 are both large relative to n1 and n2, respectively, then

σ21 = P1(1 - P1) / n1 And σ22 = P2(1 - P2) / n2

Therefore,

σ2d = [ P1(1 - P1) / n1 ] + [ P2(1 - P2) / n2 ]   
And   
σd = sqrt{ [ P1(1 - P1) / n1 ] + [ P2(1 - P2) / n2 ] }

Difference Between Proportions: Sample Problem

In this section, we work through a sample problem to show how to apply the theory presented above. In this example, we will use Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to compute probabilities. The calculator is free.

Normal Distribution Calculator

The normal calculator solves common statistical problems, based on the normal distribution. The calculator computes cumulative probabilities, based on three simple inputs. Simple instructions guide you quickly to an accurate solution. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the menu tab, at the top of every Stat Trek web page. Tap Menu - Statistical tables - Normal distribution. Or you can tap the "Normal Calculator" button below.

**Problem 1**

In one state, 52% of the voters are Republicans, and 48% are Democrats. In a second state, 47% of the voters are Republicans, and 53% are Democrats. Suppose 100 voters are surveyed from each state. Assume the survey uses simple random sampling.

What is the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state?

(A) 0.04   
(B) 0.05   
(C) 0.24   
(D) 0.71   
(E) 0.76

**Solution**

The correct answer is C. For this analysis, let P1 = the proportion of Republican voters in the first state, P2 = the proportion of Republican voters in the second state, p1 = the proportion of Republican voters in the sample from the first state, and p2 = the proportion of Republican voters in the sample from the second state. The number of voters sampled from the first state (n1) = 100, and the number of voters sampled from the second state (n2) = 100.

The solution involves four steps.

* Make sure the samples from each population are big enough to model differences with a normal distribution. Because n1P1 = 100 \* 0.52 = 52, n1(1 - P1) = 100 \* 0.48 = 48, n2P2 = 100 \* 0.47 = 47, and n2(1 - P2) = 100 \* 0.53 = 53 are each greater than 10, the sample size is large enough.
* Find the mean of the difference in sample proportions: E(p1 - p2) = P1 - P2 = 0.52 - 0.47 = 0.05.
* Find the standard deviation of the difference.

σd = sqrt{ [ P1(1 - P1) / n1 ] + [ P2(1 - P2) / n2 ] }   
σd = sqrt{ [ (0.52)(0.48) / 100 ] + [ (0.47)(0.53) / 100 ] }   
σd = sqrt (0.002496 + 0.002491) = sqrt(0.004987) = 0.0706

* Find the probability. This problem requires us to find the probability that p1 is less than p2. This is equivalent to finding the probability that p1 - p2 is less than zero. To find this probability, we need to transform the random variable (p1 - p2) into a [z-score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). That transformation appears below.

zp1 - p2 = (x - μp1 - p2) / σd = = (0 - 0.05)/0.0706 = -0.7082

Using Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the probability of a z-score being -0.7082 or less is 0.24.

Therefore, the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state is 0.24.

**Note:** Some analysts might have used the t-distribution to compute probabilities for this problem. We chose the normal distribution because the population variance was known and the sample size was large. In a previous lesson, we offered some guidelines for [choosing between the normal and the t-distribution.](http://stattrek.com/sampling/sampling-distribution.aspx#TvsNormal)

Difference Between Means

Statistics problems often involve comparisons between two independent sample means. This lesson explains how to compute probabilities associated with differences between means.

Difference Between Means: Theory

Suppose we have two [populations](http://stattrek.com/Help/Glossary.aspx?Target=Population) with means equal to μ1 and μ2. Suppose further that we take all possible [samples](http://stattrek.com/Help/Glossary.aspx?Target=Sample) of size n1 and n2. And finally, suppose that the following assumptions are valid.

* The size of each population is large relative to the sample drawn from the population. That is, N1 is large relative to n1, and N2 is large relative to n2. (In this context, populations are considered to be large if they are at least 10 times bigger than their sample.)
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent); that is, observations in population 1 are not affected by observations in population 2, and vice versa.
* The set of differences between sample means is normally distributed. This will be true if each population is normal or if the sample sizes are large. (Based on the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem), sample sizes of 40 would probably be large enough).

Given these assumptions, we know the following.

* The [expected value](http://stattrek.com/Help/Glossary.aspx?Target=Expected%20value) of the difference between all possible sample means is equal to the difference between population means. Thus, E(x1 - x2) = μd = μ1 - μ2.
* The standard deviation of the difference between sample means (σd) is approximately equal to:

σd = sqrt( σ12 / n1 + σ22 / n2 )

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

σ2d = σ2(x1 - x2) = σ2x1 + σ2x2

If the populations N1 and N2 are both large relative to n1 and n2, respectively, then

σ2x1 = σ21 / n1 And σ2x2 = σ22 / n2

Therefore,

σd2 = σ12 / n1 + σ22 / n2 And σd = sqrt( σ12 / n1 + σ22 / n2 )

Difference Between Means: Sample Problem

In this section, we work through a sample problem to show how to apply the theory presented above. In this example, we will use Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to compute probabilities.

Normal Distribution Calculator

The normal calculator solves common statistical problems, based on the normal distribution. The calculator computes cumulative probabilities, based on three simple inputs. Simple instructions guide you quickly to an accurate solution. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. Access this free calculator from the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Normal Calculator](http://stattrek.com/Tables/Normal.aspx) |

**Problem 1**

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.

In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have *at most* three more days of absences than the female sample?

(A) 0.025   
(B) 0.035   
(C) 0.045   
(D) 0.055   
(E) None of the above

**Solution**

The correct answer is B. The solution involves three or four steps, depending on whether you work directly with raw scores or z-scores. The "raw score" solution appears below:

* Find the mean difference (male absences minus female absences) in the population.

μd = μ1 - μ2 = 15 - 10 = 5

* Find the standard deviation of the difference.

σd = sqrt( σ12 / n1 + σ22 / n2 )   
σd = sqrt(72/100 + 62/50) = sqrt(49/100 + 36/50) = sqrt(0.49 + .72) = sqrt(1.21) = 1.1

* Find the probability. This problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3. To find this probability, we use Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx). Specifically, we enter the following inputs: 3, for the normal random variable; 5, for the mean; and 1.1, for the standard deviation. We find that the probability of the mean difference (male absences minus female absences) being 3 or less is about 0.035.

Thus, the probability that the difference between samples will be no more than 3 days is 0.035.

Alternatively, we could have worked with z-scores (which have a mean of 0 and a standard deviation of 1). Here's the z-score solution:

* Find the mean difference (male absences minus female absences) in the population.

μd = μ1 - μ2 = 15 - 10 = 5

* Find the standard deviation of the difference.

σd = sqrt( σ12 / n1 + σ22 / n2 )   
σd = sqrt(72/100 + 62/50) = sqrt(49/100 + 36/50) = sqrt(0.49 + .72) = sqrt(1.21) = 1.1

* Find the [z-score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score) that is produced when boys have three more days of absences than girls. When boys have three more days of absences, the number of male absences minus female absences is three. And the associated z-score is

z = (x - μ)/σ = (3 - 5)/1.1 = -2/1.1 = -1.818

* Find the probability. To find this probability, we use Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx). Specifically, we enter the following inputs: -1.818, for the normal random variable; 0, for the mean; and 1, for the standard deviation. We find that the probability of probability of a z-score being -1.818 or less is about 0.035.

Of course, the result is the same, whether you work with raw scores or z-scores.

**Note:** Some analysts might have used the t-distribution to compute probabilities for this problem. We chose the normal distribution because the population variance was known and the sample size was large. In a previous lesson, we offered some guidelines for [choosing between the normal and the t-distribution.](http://stattrek.com/sampling/sampling-distribution.aspx#TvsNormal)

# What is a Probability Distribution?

# A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

## Probability Distribution Prerequisites

To understand probability distributions, it is important to understand variables. random variables, and some notation.

* A **variable** is a symbol (*A*, *B*, *x*, *y*, etc.) that can take on any of a specified set of values.
* When the value of a variable is the outcome of a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment), that variable is a **random variable**.

Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,

* X represents the random variable X.
* P(X) represents the probability of X.
* P(X = x) refers to the probability that the random variable X is equal to a particular value, denoted by x. As an example, P(X = 1) refers to the probability that the random variable X is equal to 1.

## Probability Distributions

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

A **probability distribution** is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. Consider the coin flip experiment described above. The table below, which associates each outcome with its probability, is an example of a probability distribution.

|  |  |
| --- | --- |
| **Number of heads** | **Probability** |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

The above table represents the probability distribution of the random variable X.

## Cumulative Probability Distributions

A **cumulative probability** refers to the probability that the value of a random variable falls within a specified range.

Let us return to the coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.

P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75

Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation. In the table below, the cumulative probability refers to the probability than the random variable X is less than or equal to x.

|  |  |  |
| --- | --- | --- |
| **Number of heads: x** | **Probability: P(X = x)** | **Cumulative Probability: P(X < x)** |
| 0 | 0.25 | 0.25 |
| 1 | 0.50 | 0.75 |
| 2 | 0.25 | 1.00 |

Uniform Probability Distribution

The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the **uniform distribution**.

**Uniform Distribution.** Suppose the random variable X can assume k different values. Suppose also that the P(X = xk) is constant. Then,

P(X = xk) = 1/k

**Example 1**  
Suppose a die is tossed. What is the probability that the die will land on 5 ?

*Solution:* When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the P(X = 5) = 1/6.

**Example 2**  
Suppose we repeat the dice tossing experiment described in Example 1. This time, we ask what is the probability that the die will land on a number that is smaller than 5 ?

*Solution:* When a die is tossed, there are 6 possible outcomes represented by: S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is equally likely to occur. Thus, we have a uniform distribution.

This problem involves a cumulative probability. The probability that the die will land on a number smaller than 5 is equal to:

P( X < 5 ) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3

Probability Distributions: Discrete vs. Continuous

All probability distributions can be classified as **discrete probability distributions** or as **continuous probability distributions**, depending on whether they define probabilities associated with discrete variables or continuous variables.

Discrete vs. Continuous Variables

If a [variable](http://stattrek.com/Help/Glossary.aspx?Target=Variable) can take on any value between two specified values, it is called a **continuous variable**; otherwise, it is called a **discrete variable**.

Some examples will clarify the difference between discrete and continuous variables.

* Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
* Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Just like variables, [probability distributions](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) can be classified as discrete or continuous.

Discrete Probability Distributions

If a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random_variable) is a discrete variable, its [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) is called a **discrete probability distribution**.

An example will make this clear. Suppose you flip a coin two times. This simple [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

|  |  |
| --- | --- |
| **Number of heads** | **Probability** |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

The above table represents a *discrete* probability distribution because it relates each value of a discrete random variable with its probability of occurrence. In subsequent lessons, we will cover the following discrete probability distributions.

* [Binomial probability distribution](http://stattrek.com/Lesson2/Binomial.aspx)
* [Hypergeometric probability distribution](http://stattrek.com/Lesson2/Hypergeometric.aspx)
* [Multinomial probability distribution](http://stattrek.com/Lesson2/Multinomial.aspx)
* [Negative binomial distribution](http://stattrek.com/online-calculator/negative-binomial.aspx)
* [Poisson probability distribution](http://stattrek.com/Lesson2/Poisson.aspx)

**Note:** With a discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution can always be presented in tabular form.

Continuous Probability Distributions

If a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random_variable) is a continuous variable, its [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) is called a **continuous probability distribution**.

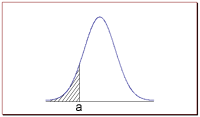
A continuous probability distribution differs from a discrete probability distribution in several ways.

* The probability that a continuous random variable will assume a particular value is zero.
* As a result, a continuous probability distribution cannot be expressed in tabular form.
* Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a **probability density function**. Sometimes, it is referred to as a **density function**, a **PDF**, or a **pdf**. For a continuous probability distribution, the density function has the following properties:

* Since the continuous random variable is defined over a continuous range of values (called the **domain** of the variable), the graph of the density function will also be continuous over that range.
* The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
* The probability that a random variable assumes a value between *a* and *b* is equal to the area under the density function bounded by *a* and *b*.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable *X* was less than or equal to *a*. The probability that *X* is less than or equal to *a* is equal to the area under the curve bounded by *a* and minus infinity - as indicated by the shaded area.



**Note:** The shaded area in the graph represents the probability that the random variable *X* is less than or equal to *a*. This is a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability). However, the probability that *X* is *exactly* equal to *a* would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as *a*) is always zero.

In subsequent lessons, we will cover the following continuous probability distributions.

* [Normal probability distribution](http://stattrek.com/Lesson2/Normal.aspx)
* [Student's t distribution](http://stattrek.com/Lesson3/TDistribution.aspx)
* [Chi-square distribution](http://stattrek.com/Lesson3/ChiSquare.aspx)
* [F distribution](http://stattrek.com/Lesson3/FDistribution.aspx)

Binomial Probability Distribution

To understand binomial distributions and binomial probability, it helps to understand binomial experiments and some associated notation; so we cover those topics first.

Binomial Experiment

A **binomial experiment** is a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:

* The experiment consists of *n* repeated trials.
* Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
* The probability of success, denoted by *P*, is the same on every trial.
* The trials are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent); that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

* The experiment consists of repeated trials. We flip a coin 2 times.
* Each trial can result in just two possible outcomes - heads or tails.
* The probability of success is constant - 0.5 on every trial.
* The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation

The following notation is helpful, when we talk about binomial probability.

* *x*: The number of successes that result from the binomial experiment.
* *n*: The number of trials in the binomial experiment.
* *P*: The probability of success on an individual trial.
* *Q*: The probability of failure on an individual trial. (This is equal to 1 - *P*.)
* *n!*: The [factorial](http://stattrek.com/statistics/dictionary.aspx?definition=factorial) of n (also known as n factorial).
* b(*x*; *n, P*): Binomial probability - the probability that an *n*-trial binomial experiment results in exactly *x* successes, when the probability of success on an individual trial is *P*.
* nCr: The number of [combinations](http://stattrek.com/Help/Glossary.aspx?Target=Combination) of *n* things, taken *r* at a time.

Binomial Distribution

A **binomial random variable** is the number of successes *x* in *n* repeated trials of a binomial experiment. The [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of a binomial random variable is called a **binomial distribution**.

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

|  |  |
| --- | --- |
| **Number of heads** | **Probability** |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

The binomial distribution has the following properties:

* The mean of the distribution (μx) is equal to *n* \* *P* .
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) (σ2x) is *n* \* *P* \* ( 1 - *P* ).
* The [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) (σx) is sqrt[ *n* \* *P* \* ( 1 - *P* ) ].

Binomial Formula and Binomial Probability

The **binomial probability** refers to the probability that a binomial experiment results in exactly *x* successes. For example, in the above table, we see that the binomial probability of getting exactly one head in two coin flips is 0.50.

Given *x*, *n*, and *P*, we can compute the binomial probability based on the binomial formula:

**Binomial Formula.** Suppose a binomial experiment consists of *n* trials and results in *x* successes. If the probability of success on an individual trial is *P*, then the binomial probability is:

b(*x*; *n, P*) = nCx \* Px \* (1 - P)n - x  
or   
b(*x*; *n, P*) = { n! / [ x! (n - x)! ] } \* Px \* (1 - P)n - x

**Example 1**  
  
Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

*Solution:* This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

b(2; 5, 0.167) = 5C2 \* (0.167)2 \* (0.833)3  
b(2; 5, 0.167) = 0.161

Cumulative Binomial Probability

A **cumulative binomial probability** refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

For example, we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

b(x < 45; 100, 0.5) =   
b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + ... + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5)

Binomial Calculator

As you may have noticed, the binomial formula requires many time-consuming computations. The Binomial Calculator can do this work for you - quickly, easily, and error-free. Use the Binomial Calculator to compute binomial probabilities and cumulative binomial probabilities. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Binomial Calculator](http://stattrek.com/Tables/Binomial.aspx) |

**Example 1**  
  
What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?

*Solution:* To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

b(x < 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + . . . + b(x = 45; 100, 0.5)   
b(x < 45; 100, 0.5) = 0.184

**Example 2**  
  
The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

*Solution:* To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

b(x < 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3)  
b(x < 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087   
b(x < 2; 5, 0.3) = 0.8369

**Example 3**  
  
What is the probability that the world series will last 4 games? 5 games? 6 games? 7 games? Assume that the teams are evenly matched.

*Solution:* This is a very tricky application of the binomial distribution. If you can follow the logic of this solution, you have a good understanding of the material covered in the tutorial, to this point.

In the world series, there are two baseball teams. The series ends when the winning team wins 4 games. Therefore, we define a success as a win by the team that ultimately becomes the world series champion.

For the purpose of this analysis, we assume that the teams are evenly matched. Therefore, the probability that a particular team wins a particular game is 0.5.

Let's look first at the simplest case. What is the probability that the series lasts only 4 games. This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

b(4; 4, 0.5) = 4C4 \* (0.5)4 \* (0.5)0 = 0.0625

Similarly, when we compute the probability of the American League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be 0.0625 + 0.0625 = 0.125; since the series would end if either the American or National League team won 4 games in a row.

Now let's tackle the question of finding probability that the world series ends in 5 games. The trick in finding this solution is to recognize that the series can only end in 5 games, if one team has won 3 out of the first 4 games. So let's first find the probability that the American League team wins exactly 3 of the first 4 games.

b(3; 4, 0.5) = 4C3 \* (0.5)3 \* (0.5)1 = 0.25

Okay, here comes some more tricky stuff, so listen up. Given that the American League team has won 3 of the first 4 games, the American League team has a 50/50 chance of winning the fifth game to end the series. Therefore, the probability of the American League team winning the series in 5 games is 0.25 \* 0.50 = 0.125. Since the National League team could also win the series in 5 games, the probability that the series ends in 5 games would be 0.125 + 0.125 = 0.25.

The rest of the problem would be solved in the same way. You should find that the probability of the series ending in 6 games is 0.3125; and the probability of the series ending in 7 games is also 0.3125.

Negative Binomial Distribution

In this lesson, we cover the negative binomial distribution and the geometric distribution. As we will see, the geometric distribution is a special case of the negative binomial distribution.

Negative Binomial Experiment

A **negative binomial experiment** is a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:

* The experiment consists of *x* repeated trials.
* Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
* The probability of success, denoted by *P*, is the same on every trial.
* The trials are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent); that is, the outcome on one trial does not affect the outcome on other trials.
* The experiment continues until *r* successes are observed, where *r* is specified in advance.

Consider the following statistical experiment. You flip a coin repeatedly and count the number of times the coin lands on heads. You continue flipping the coin until it has landed 5 times on heads. This is a negative binomial experiment because:

* The experiment consists of repeated trials. We flip a coin repeatedly until it has landed 5 times on heads.
* Each trial can result in just two possible outcomes - heads or tails.
* The probability of success is constant - 0.5 on every trial.
* The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.
* The experiment continues until a fixed number of successes have occurred; in this case, 5 heads.

Notation

The following notation is helpful, when we talk about negative binomial probability.

* *x*: The number of trials required to produce *r* successes in a negative binomial experiment.
* *r*: The number of successes in the negative binomial experiment.
* *P*: The probability of success on an individual trial.
* *Q*: The probability of failure on an individual trial. (This is equal to 1 - *P*.)
* b\*(*x*; *r, P*): Negative binomial probability - the probability that an *x*-trial negative binomial experiment results in the *rth* success on the *xth* trial, when the probability of success on an individual trial is *P*.
* nCr: The number of [combinations](http://stattrek.com/Help/Glossary.aspx?Target=Combination) of *n* things, taken *r* at a time.

Negative Binomial Distribution

A **negative binomial random variable** is the number *X* of repeated trials to produce *r* successes in a negative binomial experiment. The [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of a negative binomial random variable is called a **negative binomial distribution**. The negative binomial distribution is also known as the **Pascal distribution**.

Suppose we flip a coin repeatedly and count the number of heads (successes). If we continue flipping the coin until it has landed 2 times on heads, we are conducting a negative binomial experiment. The negative binomial random variable is the number of coin flips required to achieve 2 heads. In this example, the number of coin flips is a random variable that can take on any integer value between 2 and plus infinity. The negative binomial probability distribution for this example is presented below.

|  |  |
| --- | --- |
| **Number of coin flips** | **Probability** |
| 2 | 0.25 |
| 3 | 0.25 |
| 4 | 0.1875 |
| 5 | 0.125 |
| 6 | 0.078125 |
| 7 or more | 0.109375 |

Negative Binomial Probability

The **negative binomial probability** refers to the probability that a negative binomial experiment results in *r* - 1 successes after trial *x* - 1 and*r* successes after trial *x*. For example, in the above table, we see that the negative binomial probability of getting the second head on the sixth flip of the coin is 0.078125.

Given *x*, *r*, and *P*, we can compute the negative binomial probability based on the following formula:

**Negative Binomial Formula.** Suppose a negative binomial experiment consists of *x* trials and results in *r* successes. If the probability of success on an individual trial is *P*, then the negative binomial probability is:

b\*(*x*; *r, P*) = x-1Cr-1 \* Pr \* (1 - P)x - r

The Mean of the Negative Binomial Distribution

If we define the mean of the negative binomial distribution as the average number of trials required to produce r successes, then the mean is equal to:

μ = r / P

where μ is the mean number of trials, r is the number of successes, and P is the probability of a success on any given trial.

Alternative Views of the Negative Binomial Distribution

As if statistics weren't challenging enough, the above definition is not the only definition for the negative binomial distribution. Two common alternative definitions are:

* The negative binomial random variable is *R*, the number of successes before the binomial experiment results in *k* failures. The mean of *R* is:

μR = *kP*/*Q*

* The negative binomial random variable is *K*, the number of failures before the binomial experiment results in *r* successes. The mean of *K* is:

μK = *rQ*/*P*

The moral: If someone talks about a negative binomial distribution, find out how they are defining the negative binomial random variable.

On this website, when we refer to the negative binomial distribution, we are talking about the definition presented earlier. That is, we are defining the negative binomial random variable as *X*, the total number of trials required for the binomial experiment to produce *r* successes.

Geometric Distribution

The **geometric distribution** is a special case of the negative binomial distribution. It deals with the number of trials required for a single success. Thus, the geometric distribution is negative binomial distribution where the number of successes (*r*) is equal to 1.

An example of a geometric distribution would be tossing a coin until it lands on heads. We might ask: What is the probability that the first head occurs on the third flip? That probability is referred to as a **geometric probability** and is denoted by g(*x*; *P*). The formula for geometric probability is given below.

**Geometric Probability Formula.** Suppose a negative binomial experiment consists of *x* trials and results in one success. If the probability of success on an individual trial is *P*, then the geometric probability is:

g(*x*; *P*) = P \* Qx - 1

Sample Problems

The problems below show how to apply your new-found knowledge of the negative binomial distribution (see Example 1) and the geometric distribution (see Example 2).

Negative Binomial Calculator

As you may have noticed, the negative binomial formula requires some potentially time-consuming computations. The Negative Binomial Calculator can do this work for you - quickly, easily, and error-free. Use the Negative Binomial Calculator to compute negative binomial probabilities and geometric probabilities. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Negative Binomial Calculator](http://stattrek.com/Tables/NegBinomial.aspx) |

**Example 1**  
  
Bob is a high school basketball player. He is a 70% free throw shooter. That means his probability of making a free throw is 0.70. During the season, what is the probability that Bob makes his third free throw on his fifth shot?

*Solution:* This is an example of a negative binomial experiment. The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 3.

To solve this problem, we enter these values into the negative binomial formula.

b\*(*x*; *r, P*) = x-1Cr-1 \* Pr \* Qx - r  
b\*(*5*; *3, 0.7*) = 4C2 \* 0.73 \* 0.32  
b\*(*5*; *3, 0.7*) = 6 \* 0.343 \* 0.09 = 0.18522

Thus, the probability that Bob will make his third successful free throw on his fifth shot is 0.18522.

**Example 2**  
  
Let's reconsider the above problem from Example 1. This time, we'll ask a slightly different question: What is the probability that Bob makes his first free throw on his fifth shot?

*Solution:* This is an example of a geometric distribution, which is a special case of a negative binomial distribution. Therefore, this problem can be solved using the negative binomial formula or the geometric formula. We demonstrate each approach below, beginning with the negative binomial formula.

The probability of success (*P*) is 0.70, the number of trials (*x*) is 5, and the number of successes (*r*) is 1. We enter these values into the negative binomial formula.

b\*(*x*; *r, P*) = x-1Cr-1 \* Pr \* Qx - r  
b\*(*5*; *1, 0.7*) = 4C0 \* 0.71 \* 0.34  
b\*(*5*; *3, 0.7*) = 0.00567

Now, we demonstate a solution based on the geometric formula.

g(*x*; *P*) = P \* Qx - 1  
g(5; 0.7) = 0.7 \* 0.34 = 0.00567

Notice that each approach yields the same answer.

Hypergeometric Distribution

The probability distribution of a hypergeometric random variable is called a **hypergeometric distribution.** This lesson describes how hypergeometric random variables, hypergeometric experiments, hypergeometric probability, and the hypergeometric distribution are all related.

Notation

The following notation is helpful, when we talk about hypergeometric distributions and hypergeometric probability.

* *N*: The number of items in the [population](http://stattrek.com/Help/Glossary.aspx?Target=Population).
* *k*: The number of items in the population that are classified as successes.
* *n*: The number of items in the [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample).
* *x*: The number of items in the sample that are classified as successes.
* kCx: The number of [combinations](http://stattrek.com/Help/Glossary.aspx?Target=Combination) of *k* things, taken *x* at a time.
* h(*x*; *N*, *n*, *k*): **hypergeometric probability** - the probability that an *n*-trial hypergeometric experiment results in exactly*x* successes, when the population consists of *N* items, *k* of which are classified as successes.

Hypergeometric Experiments

A **hypergeometric experiment** is a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:

* A [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) of size *n* is randomly selected [without replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20without%20replacement) from a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) of *N* items.
* In the population, *k* items can be classified as successes, and *N - k* items can be classified as failures.

Consider the following statistical experiment. You have an urn of 10 marbles - 5 red and 5 green. You randomly select 2 marbles without replacement and count the number of red marbles you have selected. This would be a hypergeometric experiment.

Note that it would not be a [binomial experiment](http://stattrek.com/Help/Glossary.aspx?Target=Binomial_experiment). A binomial experiment requires that the probability of success be constant on every trial. With the above experiment, the probability of a success changes on every trial. In the beginning, the probability of selecting a red marble is 5/10. If you select a red marble on the first trial, the probability of selecting a red marble on the second trial is 4/9. And if you select a green marble on the first trial, the probability of selecting a red marble on the second trial is 5/9.

Note further that if you selected the marbles with replacement, the probability of success would not change. It would be 5/10 on every trial. Then, this would be a binomial experiment.

Hypergeometric Distribution

A **hypergeometric random variable** is the number of successes that result from a hypergeometric experiment. The [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of a hypergeometric random variable is called a **hypergeometric distribution**.

Given *x*, *N*, *n*, and *k*, we can compute the hypergeometric probability based on the following formula:

**Hypergeometric Formula.** Suppose a population consists of *N* items, *k* of which are successes. And a random sample drawn from that population consists of *n* items, *x* of which are successes. Then the hypergeometric probability is:

h(*x*; *N*, *n*, *k*) = [ kCx ] [ N-kCn-x ] / [ NCn ]

The hypergeometric distribution has the following properties:

* The mean of the distribution is equal to *n* \* *k* / *N* .
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) is *n* \* *k* \* ( *N* - *k* ) \* ( *N* - *n* ) / [ *N*2 \* ( *N* - 1 ) ] .

**Example 1**  
  
Suppose we randomly select 5 cards without replacement from an ordinary deck of playing cards. What is the probability of getting exactly 2 red cards (i.e., hearts or diamonds)?

*Solution:* This is a hypergeometric experiment in which we know the following:

* N = 52; since there are 52 cards in a deck.
* k = 26; since there are 26 red cards in a deck.
* n = 5; since we randomly select 5 cards from the deck.
* x = 2; since 2 of the cards we select are red.

We plug these values into the hypergeometric formula as follows:

h(*x*; *N*, *n*, *k*) = [ kCx ] [ N-kCn-x ] / [ NCn ]   
h(*2*; *52*, *5*, *26*) = [ 26C2 ] [ 26C3 ] / [ 52C5 ]   
h(*2*; *52*, *5*, *26*) = [ 325 ] [ 2600 ] / [ 2,598,960 ] = 0.32513

Thus, the probability of randomly selecting 2 red cards is 0.32513.

Hypergeometric Calculator

As you surely noticed, the hypergeometric formula requires many time-consuming computations. The Stat Trek Hypergeometric Calculator can do this work for you - quickly, easily, and error-free. Use the Hypergeometric Calculator to compute hypergeometric probabilities and cumulative hypergeometric probabilities. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Hypergeometric Calculator](http://stattrek.com/Tables/Hypergeometric.aspx) |

Cumulative Hypergeometric Probability

A **cumulative hypergeometric probability** refers to the probability that the hypergeometric random variable is greater than or equal to some specified lower limit and less than or equal to some specified upper limit.

For example, suppose we randomly select five cards from an ordinary deck of playing cards. We might be interested in the cumulative hypergeometric probability of obtaining 2 or fewer hearts. This would be the probability of obtaining 0 hearts plus the probability of obtaining 1 heart plus the probability of obtaining 2 hearts, as shown in the example below.

**Example 1**  
  
Suppose we select 5 cards from an ordinary deck of playing cards. What is the probability of obtaining 2 or fewer hearts?

*Solution:* This is a hypergeometric experiment in which we know the following:

* N = 52; since there are 52 cards in a deck.
* k = 13; since there are 13 hearts in a deck.
* n = 5; since we randomly select 5 cards from the deck.
* x = 0 to 2; since our selection includes 0, 1, or 2 hearts.

We plug these values into the hypergeometric formula as follows:

h(*x* < x; *N*, *n*, *k*) = h(*x* < 2; *52*, *5*, *13*)   
h(*x* < 2; *52*, *5*, *13*) = h(*x* = 0; *52*, *5*, *13*) + h(*x* = 1; *52*, *5*, *13*) + h(*x* = 2; *52*, *5*, *13*)   
h(*x* < 2; *52*, *5*, *13*) = [ (13C0) (39C5) / (52C5) ] + [ (13C1) (39C4) / (52C5) ] + [ (13C2) (39C3) / (52C5) ]   
h(*x* < 2; *52*, *5*, *13*) = [ (1)(575,757)/(2,598,960) ] + [ (13)(82,251)/(2,598,960) ] + [ (78)(9139)/(2,598,960) ]   
h(*x* < 2; *52*, *5*, *13*) = [ 0.2215 ] + [ 0.4114 ] + [ 0.2743 ]   
h(*x* < 2; *52*, *5*, *13*) = 0.9072

Thus, the probability of randomly selecting at most 2 hearts is 0.9072.

Multinomial Distribution

A **multinomial distribution** is the [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of the outcomes from a multinomial experiment.

Multinomial Experiment

A **multinomial experiment** is a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:

* The experiment consists of *n* repeated trials.
* Each trial has a discrete number of possible outcomes.
* On any given trial, the probability that a particular outcome will occur is constant.
* The trials are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent); that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You toss two dice three times, and record the outcome on each toss. This is a multinomial experiment because:

* The experiment consists of repeated trials. We toss the dice three times.
* Each trial can result in a discrete number of outcomes - 2 through 12.
* The probability of any outcome is constant; it does not change from one toss to the next.
* The trials are independent; that is, getting a particular outcome on one trial does not affect the outcome on other trials.

**Note:** A [binomial experiment](http://stattrek.com/Help/Glossary.aspx?Target=Binomial_experiment) is a special case of a multinomial experiment. Here is the main difference. With a binomial experiment, each trial can result in two - and only two - possible outcomes. With a multinomial experiment, each trial can have two *or more* possible outcomes.

Multinomial Distribution

A **multinomial distribution** is the [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of the outcomes from a multinomial experiment. The multinomial formula defines the probability of any outcome from a multinomial experiment.

**Multinomial Formula.** Suppose a multinomial experiment consists of *n* trials, and each trial can result in any of *k* possible outcomes: E1, E2, . . . , Ek. Suppose, further, that each possible outcome can occur with probabilities p1, p2, . . . , pk. Then, the probability (P) that E1 occurs n1 times, E2 occurs n2 times, . . . , and Ek occurs nk times is

P = [ n! / ( n1! \* n2! \* ... nk! ) ] \* ( p1n1 \* p2n2 \* . . . \* pknk )

where n = n1 + n2 + . . . + nk.

The examples below illustrate how to use the multinomial formula to compute the probability of an outcome from a multinomial experiment.

Multinomial Calculator

As you may have noticed, the multinomial formula requires many time-consuming computations. The Multinomial Calculator can do this work for you - quickly, easily, and error-free. Use the Multinomial Calculator to compute the probability of outcomes from multinomial experiments. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

Multinomial Probability: Sample Problems

**Example 1**  
  
Suppose a card is drawn randomly from an ordinary deck of playing cards, and then put back in the deck. This exercise is repeated five times. What is the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs?

*Solution:* To solve this problem, we apply the multinomial formula. We know the following:

* The experiment consists of 5 trials, so n = 5.
* The 5 trials produce 1 spade, 1 heart, 1 diamond, and 2 clubs; so n1 = 1, n2 = 1, n3 = 1, and n4 = 2.
* On any particular trial, the probability of drawing a spade, heart, diamond, or club is 0.25, 0.25, 0.25, and 0.25, respectively. Thus, p1 = 0.25, p2 = 0.25, p3 = 0.25, and p4 = 0.25.

We plug these inputs into the multinomial formula, as shown below:

P = [ n! / ( n1! \* n2! \* ... nk! ) ] \* ( p1n1 \* p2n2 \* . . . \* pknk )   
P = [ 5! / ( 1! \* 1! \* 1! \* 2! ) ] \* [ (0.25)1 \* (0.25)1 \* (0.25)1 \* (0.25)2 ]   
P = 0.05859

Thus, if we draw five cards [with replacement](http://stattrek.com/Help/Glossary.aspx?Target=sampling%20with%20replacement) from an ordinary deck of playing cards, the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs is 0.05859.

**Example 2**  
  
Suppose we have a bowl with 10 marbles - 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select 4 marbles from the bowl, [with replacement](http://stattrek.com/Help/Glossary.aspx?Target=sampling%20with%20replacement). What is the probability of selecting 2 green marbles and 2 blue marbles?

*Solution:* To solve this problem, we apply the multinomial formula. We know the following:

* The experiment consists of 4 trials, so n = 4.
* The 4 trials produce 0 red marbles, 2 green marbles, and 2 blue marbles; so nred = 0, ngreen = 2, and nblue = 2.
* On any particular trial, the probability of drawing a red, green, or blue marble is 0.2, 0.3, and 0.5, respectively. Thus, pred = 0.2, pgreen = 0.3, and pblue = 0.5

We plug these inputs into the multinomial formula, as shown below:

P = [ n! / ( n1! \* n2! \* ... nk! ) ] \* ( p1n1 \* p2n2 \* . . . \* pknk )   
P = [ 4! / ( 0! \* 2! \* 2! ) ] \* [ (0.2)0 \* (0.3)2 \* (0.5)2 ]   
P = 0.135

Thus, if we draw 4 marbles [with replacement](http://stattrek.com/Help/Glossary.aspx?Target=sampling%20with%20replacement) from the bowl, the probability of drawing 0 red marbles, 2 green marbles, and 2 blue marbles is 0.135.

Poisson Distribution

A Poisson distribution is the probability distribution that results from a Poisson experiment.

Attributes of a Poisson Experiment

A **Poisson experiment** is a [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) that has the following properties:

* The experiment results in outcomes that can be classified as successes or failures.
* The average number of successes (μ) that occurs in a specified region is known.
* The probability that a success will occur is proportional to the size of the region.
* The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

* *e*: A constant equal to approximately 2.71828. (Actually, *e* is the base of the natural logarithm system.)
* μ: The mean number of successes that occur in a specified region.
* *x*: The actual number of successes that occur in a specified region.
* P(*x*; μ): The **Poisson probability** that exactly *x* successes occur in a Poisson experiment, when the mean number of successes is μ.

Poisson Distribution

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The [probability distribution](http://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

**Poisson Formula.** Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ. Then, the Poisson probability is:

P(*x*; μ) = (e-μ) (μx) / x!

where *x* is the actual number of successes that result from the experiment, and *e* is approximately equal to 2.71828.

The Poisson distribution has the following properties:

* The mean of the distribution is equal to μ .
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) is also equal to μ .

**Example 1**  
  
The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

*Solution:* This is a Poisson experiment in which we know the following:

* μ = 2; since 2 homes are sold per day, on average.
* x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
* e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

P(*x*; μ) = (e-μ) (μx) / x!   
P(3; 2) = (2.71828-2) (23) / 3!   
P(3; 2) = (0.13534) (8) / 6   
P(3; 2) = 0.180

Thus, the probability of selling 3 homes tomorrow is 0.180 .

Poisson Calculator

Clearly, the Poisson formula requires many time-consuming computations. The Stat Trek Poisson Calculator can do this work for you - quickly, easily, and error-free. Use the Poisson Calculator to compute Poisson probabilities and cumulative Poisson probabilities. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

Cumulative Poisson Probability

A **cumulative Poisson probability** refers to the probability that the Poisson random variable is greater than some specified lower limit and less than some specified upper limit.

**Example 1**  
  
Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

*Solution:* This is a Poisson experiment in which we know the following:

* μ = 5; since 5 lions are seen per safari, on average.
* x = 0, 1, 2, or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
* e = 2.71828; since *e* is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5). To compute this sum, we use the Poisson formula:

P(x < 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)  
P(x < 3, 5) = [ (e-5)(50) / 0! ] + [ (e-5)(51) / 1! ] + [ (e-5)(52) / 2! ] + [ (e-5)(53) / 3! ]   
P(x < 3, 5) = [ (0.006738)(1) / 1 ] + [ (0.006738)(5) / 1 ] + [ (0.006738)(25) / 2 ] + [ (0.006738)(125) / 6 ]   
P(x < 3, 5) = [ 0.0067 ] + [ 0.03369 ] + [ 0.084224 ] + [ 0.140375 ]   
P(x < 3, 5) = 0.2650

Thus, the probability of seeing at no more than 3 lions is 0.2650.

What is the Normal Distribution?

The **normal distribution** refers to a family of [continuous probability distributions](http://stattrek.com/Help/Glossary.aspx?Target=Continuous%20probability%20distribution) described by the normal equation.

The Normal Equation

The normal distribution is defined by the following equation:

**Normal equation.** The value of the random variable *Y* is:

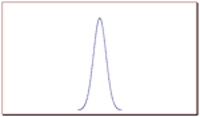
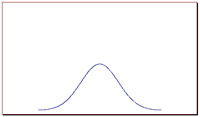
Y = { 1/[ σ \* sqrt(2π) ] } \* e-(x - μ)2/2σ2

where *X* is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and *e* is approximately 2.71828.

The random variable *X* in the normal equation is called the **normal random variable**. The normal equation is the [probability density function](http://stattrek.com/Help/Glossary.aspx?Target=Probability%20density%20function) for the normal distribution.

The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

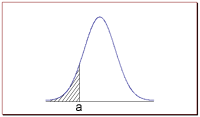


The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

* The total area under the normal curve is equal to 1.
* The probability that a normal random variable *X* equals any particular value is 0.
* The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity (as indicated by the *non-shaded* area in the figure below).
* The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity (as indicated by the *shaded* area in the figure below).



Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

* About 68% of the area under the curve falls within 1 standard deviation of the mean.
* About 95% of the area under the curve falls within 2 standard deviations of the mean.
* About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the **empirical rule** or the **68-95-99.7 rule**. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table. In the examples below, we illustrate the use of Stat Trek's [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), a free tool available on this site. In the next lesson, we demonstrate the use of normal distribution tables.

Normal Distribution Calculator

The normal calculator solves common statistical problems, based on the normal distribution. The calculator computes cumulative probabilities, based on three simple inputs. Simple instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

**Example 1**  
  
An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

*Solution:* Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

* The value of the normal random variable is 365 days.
* The mean is equal to 300 days.
* The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: P( X < 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

**Example 2**  
  
Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

*Solution:* Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

P( 90 < *X* < 110 ) = P( X < 110 ) - P( X < 90 )

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

* To compute P( X < 110 ), we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that P( X < 110 ) is 0.84.
* To compute P( X < 90 ), we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that P( X < 90 ) is 0.16.

We use these findings to compute our final answer as follows:

P( 90 < *X* < 110 ) = P( X < 110 ) - P( X < 90 )  
P( 90 < *X* < 110 ) = 0.84 - 0.16  
P( 90 < *X* < 110 ) = 0.68

Thus, about 68% of the test scores will fall between 90 and 110.

Standard Normal Distribution

The **standard normal distribution** is a special case of the [normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution). It is the distribution that occurs when a [normal random variable](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20random%20variable) has a mean of zero and a standard deviation of one.

Standard Score (aka, z Score)

The normal random variable of a standard normal distribution is called a **standard score** or a **z-score**. Every normal random variable *X* can be transformed into a *z* score via the following equation:

*z* = (*X* - μ) / σ

where *X* is a normal random variable, μ is the mean of *X*, and σ is the standard deviation of *X*.

Standard Normal Distribution Table

A **standard normal distribution table** shows a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) associated with a particular z-score. Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table.

For example, a section of the standard normal table is reproduced below. To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, P(Z < -1.31) = 0.0951.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| -1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0722 | 0.0708 | 0.0694 | 0.0681 |
| -1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| -1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |

Of course, you may not be interested in the probability that a standard normal random variable falls between minus infinity and a given value. You may want to know the probability that it lies between a given value and plus infinity. Or you may want to know the probability that a standard normal random variable lies between two given values. These probabilities are easy to compute from a normal distribution table. Here's how.

* Find P(Z > a). The probability that a standard normal random variable (z) is greater than a given value (a) is easy to find. The table shows the P(Z < a). The P(Z > a) = 1 - P(Z < a).   
    
  Suppose, for example, that we want to know the probability that a z-score will be greater than 3.00. From the table (see above), we find that P(Z < 3.00) = 0.9987. Therefore, P(Z > 3.00) = 1 - P(Z < 3.00) = 1 - 0.9987 = 0.0013.
* Find P(a < Z < b). The probability that a standard normal random variables lies between two values is also easy to find. The P(a < Z < b) = P(Z < b) - P(Z < a).   
    
  For example, suppose we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table (see above), we find that P(Z < -1.20) = 0.1151; and P(Z < -1.40) = 0.0808. Therefore, P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) = 0.1151 - 0.0808 = 0.0343.

In school or on the Advanced Placement Statistics Exam, you may be called upon to use or interpret standard normal distribution tables. Standard normal tables are commonly found in appendices of most statistics texts.

The Normal Distribution as a Model for Measurements

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

* Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier: *z* = (*X* - μ) / σ.
* Find probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek's free [normal distribution calculator](http://stattrek.com/Tables/Normal.aspx)), or handheld [graphing calculators](http://stattrek.com/Calc/graphing-calculators.aspx) to find probabilities associated with the z-scores.

The problem in the next section demonstrates the use of the normal distribution as a model for measurement.

Test Your Understanding

**Problem 1**

Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Molly? (Assume that test scores are normally distributed.)

(A) 0.10   
(B) 0.18   
(C) 0.50   
(D) 0.82   
(E) 0.90

**Solution**

The correct answer is B. As part of the solution to this problem, we assume that test scores are normally distributed. In this way, we use the [normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution) as a model for measurement. Given an assumption of normality, the solution involves three steps.

* First, we transform Molly's test score into a [z-score](http://stattrek.com/Help/Glossary.aspx?Target=z%20Score), using the z-score transformation equation.

*z* = (*X* - μ) / σ = (940 - 850) / 100 = 0.90

* Then, using an online calculator (e.g., Stat Trek's free [normal distribution calculator](http://stattrek.com/Tables/Normal.aspx)), a handheld [graphing calculator](http://stattrek.com/Calc/graphing-calculators.aspx), or the standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find P(Z < 0.90) = 0.8159.
* Therefore, the P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841.

Thus, we estimate that 18.41 percent of the students tested had a higher score than Molly.

**Student's t Distribution**

The **t distribution** (aka, **Student’s t-distribution**) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

**Why Use the t Distribution?**

According to the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem), the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of a statistic (like a sample mean) will follow a [normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution), as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a [z-score](http://stattrek.com/Help/Glossary.aspx?Target=Z-score), and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic** (also known as the **t score**), whose values are given by:

t = [ x - μ ] / [ s / sqrt( n ) ]

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the *t* statistic is called the **t distribution** or the **Student t distribution**.

The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

**Degrees of Freedom**

There are actually many different t distributions. The particular form of the t distribution is determined by its **degrees of freedom**. The degrees of freedom refers to the number of independent observations in a set of data.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the *t* statistic from samples of size 8 would be described by a t distribution having 8 - 1 or 7 degrees of freedom. Similarly, a t distribution having 15 degrees of freedom would be used with a sample of size 16.

For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.

**Properties of the t Distribution**

The t distribution has the following properties:

* The mean of the distribution is equal to 0 .
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) is equal to *v* / ( *v* - 2 ), where *v* is the degrees of freedom (see last section) and *v* > 2.
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the [standard normal distribution](http://stattrek.com/Help/Glossary.aspx?Target=Standard_normal_distribution).

**When to Use the t Distribution**

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The sampling distribution of a statistic should be bell-shaped if any of the following conditions apply.

* The population distribution is normal.
* The population distribution is [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is at least 30.
* The population distribution is moderately [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is at least 40.
* The sample size is greater than 40, without outliers.

The t distribution should *not* be used with small samples from populations that are not approximately normal.

**Probability and the Student t Distribution**

When a sample of size *n* is drawn from a population having a normal (or nearly normal) distribution, the sample mean can be transformed into a t statistic, using the equation presented at the beginning of this lesson. We repeat that equation below:

t = [ x - μ ] / [ s / sqrt( n ) ]

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, n is the sample size, and degrees of freedom are equal to n - 1.

The t statistic produced by this transformation can be associated with a unique [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability). This cumulative probability represents the likelihood of finding a sample mean less than or equal to x, given a random sample of size *n*.

The easiest way to find the probability associated with a particular t statistic is to use the [T Distribution Calculator](http://stattrek.com/Tables/T.aspx), a free tool provided by Stat Trek.

**Notation and t Statistics**

Statisticians use tα to represent the t statistic that has a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability) of (1 - α). For example, suppose we were interested in the t statistic having a cumulative probability of 0.95. In this example, α would be equal to (1 - 0.95) or 0.05. We would refer to the t statistic as t0.05

Of course, the value of t0.05 depends on the number of degrees of freedom. For example, with 2 degrees of freedom, that t0.05 is equal to 2.92; but with 20 degrees of freedom, that t0.05 is equal to 1.725.

Note: Because the t distribution is symmetric about a mean of zero, the following is true.

*t*α = -*t*1 - alpha       And       *t*1 - alpha = -*t*α

Thus, if t0.05 = 2.92, then t0.95 = -2.92.

**T Distribution Calculator**

The T Distribution Calculator solves common statistics problems, based on the t distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [T Distribution Calculator](http://stattrek.com/Tables/T.aspx) |

**Test Your Understanding**

**Problem 1**

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

**Note:** There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented below. Solution A is the traditional approach. It requires you to compute the t statistic, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability. Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t statistic "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

**Solution A**

The first thing we need to do is compute the t statistic, based on the following equation:

t = [ x - μ ] / [ s / sqrt( n ) ]   
t = ( 290 - 300 ) / [ 50 / sqrt( 15) ] = -10 / 12.909945 = - 0.7745966

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

Now, we are ready to use the [T Distribution Calculator](http://stattrek.com/Tables/T.aspx). Since we know the t statistic, we select "T score" from the Random Variable dropdown box. Then, we enter the following data:

* The degrees of freedom are equal to 15 - 1 = 14.
* The t statistic is equal to - 0.7745966.

The calculator displays the cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

**Solution B:**

This time, we will work directly with the raw data from the problem. We will not compute the t statistic; the [T Distribution Calculator](http://stattrek.com/Tables/T.aspx) will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

* The degrees of freedom are equal to 15 - 1 = 14.
* Assuming the CEO's claim is true, the population mean equals 300.
* The sample mean equals 290.
* The standard deviation of the sample is 50.

The calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

**Problem 2**

Suppose scores on an IQ test are normally distributed, with a population mean of 100. Suppose 20 people are randomly selected and tested. The standard deviation in the sample group is 15. What is the probability that the average test score in the sample group will be at most 110?

**Solution:**

To solve this problem, we will work directly with the raw data from the problem. We will not compute the t statistic; the [T Distribution Calculator](http://stattrek.com/Tables/T.aspx) will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

* The degrees of freedom are equal to 20 - 1 = 19.
* The population mean equals 100.
* The sample mean equals 110.
* The standard deviation of the sample is 15.

We enter these values into the [T Distribution Calculator](http://stattrek.com/Tables/T.aspx). The calculator displays the cumulative probability: 0.996. Hence, there is a 99.6% chance that the sample average will be no greater than 110.

**Chi-Square Distribution**

The distribution of the chi-square statistic is called the chi-square distribution. In this lesson, we learn to compute the chi-square statistic and find the probability associated with the statistic. Chi-square examples illustrate key points.

**The Chi-Square Statistic**

Suppose we conduct the following [statistical experiment](http://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment). We select a random sample of size *n* from a normal population, having a standard deviation equal to σ. We find that the standard deviation in our sample is equal to *s*. Given these data, we can define a [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic), called **chi-square**, using the following equation:

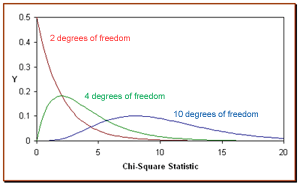
Χ2 = [ ( n - 1 ) \* s2 ] / σ2

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following [probability density function](http://stattrek.com/Help/Glossary.aspx?Target=Probability_density_function):

Y = Y0 \* ( Χ2 ) ( v/2 - 1 ) \* *e*-Χ2 / 2

where Y0 is a constant that depends on the number of degrees of freedom, Χ2 is the chi-square statistic, *v* = *n* - 1 is the number of [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees_of_freedom), and *e* is a constant equal to the base of the natural logarithm system (approximately 2.71828). Y0 is defined, so that the area under the chi-square curve is equal to one.

In the figure below, the red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is *n* - 1 = 3 - 1 = 2. Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4); and the blue curve, for samples of size 11 (degrees of freedom equal to 10).

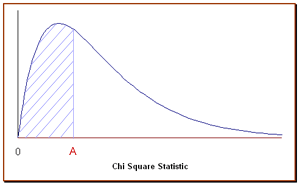


The chi-square distribution has the following properties:

* The mean of the distribution is equal to the number of degrees of freedom: μ = *v*.
* The variance is equal to two times the number of degrees of freedom: σ2 = 2 \* *v*
* When the degrees of freedom are greater than or equal to 2, the maximum value for Y occurs when Χ2 = *v* - 2.
* As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

**Cumulative Probability and the Chi-Square Distribution**

The chi-square distribution is constructed so that the total area under the curve is equal to 1. The area under the curve between 0 and a particular chi-square value is a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability) associated with that chi-square value. For example, in the figure below, the shaded area represents a cumulative probability associated with a chi-square statistic equal to *A*; that is, it is the probability that the value of a chi-square statistic will fall between 0 and *A*.



Fortunately, we don't have to compute the area under the curve to find the probability. The easiest way to find the cumulative probability associated with a particular chi-square statistic is to use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx), a free tool provided by Stat Trek.

**Chi-Square Distribution Calculator**

The Chi-Square Distribution Calculator solves common statistics problems, based on the chi-square distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) |

**Test Your Understanding: Chi-Square Examples**

**Problem 1**

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

**Solution**

We know the following:

* The standard deviation of the population is 4 minutes.
* The standard deviation of the sample is 6 minutes.
* The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

Χ2 = [ ( n - 1 ) \* s2 ] / σ2   
Χ2 = [ ( 7 - 1 ) \* 62 ] / 42 = 13.5

where Χ2 is the chi-square statistic, *n* is the sample size, *s* is the standard deviation of the sample, and σ is the standard deviation of the population.

**Problem 2**  
  
Let's revisit the problem presented above. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

**Solution**

We know the following:

* The sample size *n* is equal to 7.
* The degrees of freedom are equal to *n* - 1 = 7 - 1 = 6.
* The chi-square statistic is equal to 13.5 (see Example 1 above).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic falls between 0 and 13.5, we enter the degrees of freedom (6) and the chi-square statistic (13.5) into the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx). The calculator displays the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means (by the [subtraction rule](http://stattrek.com/Help/Glossary.aspx?Target=Subtraction_rule)) that the probability that the standard deviation would be *greater than* 6 minutes is 1 - 0.96 or .04.

**F Distribution**

The F distribution is the probability distribution associated with the f statistic. In this lesson, we show how to compute an f statistic and how to find probabilities associated with specific f statistic values.

**The f Statistic**

The ***f* statistic**, also known as an ***f* value**, is a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random_variable) that has an F distribution. (We discuss the F distribution in the next section.)

Here are the steps required to compute an ***f* statistic**:

* Select a random sample of size *n*1 from a normal population, having a standard deviation equal to σ1.
* Select an independent random sample of size *n*2 from a normal population, having a standard deviation equal to σ2.
* The *f* statistic is the ratio of *s*12/σ12 and *s*22/σ22.

The following equivalent equations are commonly used to compute an *f* statistic:

*f* = [ *s*12/σ12 ] / [ *s*22/σ22 ]   
*f* = [ *s*12 \* σ22 ] / [ *s*22 \* σ12 ]   
*f* = [ Χ21 / *v*1 ] / [ Χ22 / *v*2 ]   
*f* = [ Χ21 \* *v*2 ] / [ Χ22 \* *v*1 ]

where σ1 is the standard deviation of population 1, *s*1 is the standard deviation of the sample drawn from population 1, σ2 is the standard deviation of population 2, *s*2 is the standard deviation of the sample drawn from population 2, Χ21 is the [chi-square statistic](http://stattrek.com/Help/Glossary.aspx?Target=Chi_square_statistic) for the sample drawn from population 1, *v*1 is the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees_of_freedom) for Χ21, Χ22 is the chi-square statistic for the sample drawn from population 2, and *v*2 is the degrees of freedom for Χ22 . Note that degrees of freedom *v*1 = *n*1 - 1, and degrees of freedom *v*2 = *n*2 - 1 .

**The F Distribution**

The distribution of all possible values of the *f* statistic is called an **F distribution**, with *v*1 = *n*1 - 1 and *v*2 = *n*2 - 1 degrees of freedom.

The curve of the F distribution depends on the degrees of freedom, *v*1 and *v*2. When describing an F distribution, the number of degrees of freedom associated with the standard deviation in the numerator of the *f* statistic is always stated first. Thus, *f*(5, 9) would refer to an F distribution with *v*1 = 5 and *v*2 = 9 degrees of freedom; whereas *f*(9, 5) would refer to an F distribution with *v*1 = 9 and *v*2 = 5 degrees of freedom. Note that the curve represented by *f*(5, 9) would differ from the curve represented by *f*(9, 5).

The F distribution has the following properties:

* The mean of the distribution is equal to *v*2 / ( *v*2 - 2 ) for *v*2 > 2.
* The [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) is equal to [ 2 \* *v*22 \* ( *v*1 + *v*1 - 2 ) ] / [ *v*1 \* ( *v*2 - 2 )2 \* ( *v*2 - 4 ) ] for *v*2 > 4.

**Cumulative Probability and the F Distribution**

Every *f* statistic can be associated with a unique [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability). This cumulative probability represents the likelihood that the *f* statistic is less than or equal to a specified value.

Statisticians use *f*α to represent the value of an *f* statistic having a cumulative probability of (1 - α). For example, suppose we were interested in the *f* statistic having a cumulative probability of 0.95. We would refer to that *f* statistic as f0.05, since (1 - 0.95) = 0.05.

Of course, to find the value of *f*α, we would need to know the degrees of freedom, *v*1 and *v*2. Notationally, the degrees of freedom appear in parentheses as follows: *f*α(*v*1,*v*2). Thus, *f*0.05(5, 7) refers to value of the f statistic having a cumulative probability of 0.95, *v*1 = 5 degrees of freedom, and *v*2 = 7 degrees of freedom.

The easiest way to find the value of a particular *f* statistic is to use the [F Distribution Calculator](http://stattrek.com/Tables/F.aspx), a free tool provided by Stat Trek. For example, the value of *f*0.05(5, 7) is 3.97. The use of the F Distribution Calculator is illustrated in the examples below.

**F Distribution Calculator**

The F Distribution Calculator solves common statistics problems, based on the F distribution. The calculator computes cumulative probabilities, based on simple inputs. Clear instructions guide you to an accurate solution, quickly and easily. If anything is unclear, frequently-asked questions and sample problems provide straightforward explanations. The calculator is free. It can be found under the Stat Tables tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
|  |

**Sample Problems**

**Example 1**  
  
Suppose you randomly select 7 women from a population of women, and 12 men from a population of men. The table below shows the standard deviation in each sample and in each population.

|  |  |  |
| --- | --- | --- |
| **Population** | **Population standard deviation** | **Sample standard deviation** |
| Women | 30 | 35 |
| Men | 50 | 45 |

Compute the f statistic.

*Solution A:* The f statistic can be computed from the population and sample standard deviations, using the following equation:

*f* = [ *s*12/σ12 ] / [ *s*22/σ22 ]

where σ1 is the standard deviation of population 1, *s*1 is the standard deviation of the sample drawn from population 1, σ2 is the standard deviation of population 2, and *s*1 is the standard deviation of the sample drawn from population 2.

As you can see from the equation, there are actually two ways to compute an f statistic from these data. If the women's data appears in the numerator, we can calculate an f statistic as follows:

*f* = ( 352 / 302 ) / ( 452 / 502 ) = (1225 / 900) / (2025 / 2500) = 1.361 / 0.81 = 1.68

For this calculation, the numerator degrees of freedom *v*1 are 7 - 1 or 6; and the denominator degrees of freedom *v*2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate an f statistic as follows:

*f* = ( 452 / 502 ) / ( 352 / 302 ) = (2025 / 2500) / (1225 / 900) = 0.81 / 1.361 = 0.595

For this calculation, the numerator degrees of freedom *v*1 are 12 - 1 or 11; and the denominator degrees of freedom *v*2 are 7 - 1 or 6.

When you are trying to find the cumulative probability associated with an f statistic, you need to know *v*1 and *v*2. This point is illustrated in the next example.

**Example 2**  
  
Find the cumulative probability associated with each of the f statistics from Example 1, above.

*Solution:* To solve this problem, we need to find the degrees of freedom for each sample. Then, we will use the [F Distribution Calculator](http://stattrek.com/Tables/F.aspx) to find the probabilities.

* The degrees of freedom for the sample of women is equal to *n* - 1 = 7 - 1 = 6.
* The degrees of freedom for the sample of men is equal to *n* - 1 = 12 - 1 = 11.

Therefore, when the women's data appear in the numerator, the numerator degrees of freedom *v*1 is equal to 6; and the denominator degrees of freedom *v*2 is equal to 11. And, based on the computations shown in the previous example, the f statistic is equal to 1.68. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.78.

On the other hand, when the men's data appear in the numerator, the numerator degrees of freedom *v*1 is equal to 11; and the denominator degrees of freedom *v*2 is equal to 6. And, based on the computations shown in the previous example, the f statistic is equal to 0.595. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.22.

Estimation in Statistics

In statistics, **estimation** refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Point Estimate vs. Interval Estimate

Statisticians use sample [statistics](http://stattrek.com/Help/Glossary.aspx?Target=Statistic) to estimate population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter). For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

* **Point estimate**. A point estimate of a population parameter is a single value of a statistic. For example, the sample mean x is a point estimate of the population mean μ. Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.
* **Interval estimate**. An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, *a* < x < *b* is an interval estimate of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*.

Confidence Intervals

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular [sampling method](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20method). A confidence interval consists of three parts.

* A confidence level.
* A statistic.
* A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* + *margin of error*.

For example, suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the *sample statistic* + *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Test Your Understanding

**Problem 1**

Which of the following statements is true.

I. When the margin of error is small, the confidence level is high.   
II. When the margin of error is small, the confidence level is low.   
III. A confidence interval is a type of point estimate.   
IV. A population mean is an example of a point estimate.

(A) I only   
(B) II only   
(C) III only   
(D) IV only.   
(E) None of the above.

**Solution**

The correct answer is (E). The confidence level is not affected by the margin of error. When the margin of error is small, the confidence level can low or high or anything in between. A confidence interval is a type of interval estimate, not a type of point estimate. A *population* mean is not an example of a point estimate; a *sample* mean is an example of a point estimate.

What is the Standard Error?

The standard error is an estimate of the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) of a [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic). This lesson shows how to compute the standard error, based on sample data.

The standard error is important because it is used to compute other measures, like [confidence intervals](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) and [margins of error](http://stattrek.com/Help/Glossary.aspx?Target=margin-of-error).

Notation

The following notation is helpful, when we talk about the standard deviation and the standard error.

|  |  |  |  |
| --- | --- | --- | --- |
| Population parameter | | Sample statistic | |
|  | N: Number of observations in the population |  | n: Number of observations in the sample |
|  | Ni: Number of observations in population *i* |  | ni: Number of observations in sample *i* |
|  | P: Proportion of successes in population |  | p: Proportion of successes in sample |
|  | Pi: Proportion of successes in population *i* |  | pi: Proportion of successes in sample *i* |
|  | μ: Population mean |  | x: Sample estimate of population mean |
|  | μi: Mean of population i |  | xi: Sample estimate of μi |
|  | σ: Population standard deviation |  | s: Sample estimate of σ |
|  | σp: Standard deviation of p |  | SEp: Standard error of p |
|  | σx: Standard deviation of x |  | SEx: Standard error of x |

Standard Deviation of Sample Estimates

Statisticians use sample statistics to estimate population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter). Naturally, the value of a statistic may vary from one sample to the next.

The variability of a statistic is measured by its standard deviation. The table below shows formulas for computing the standard deviation of statistics from [simple random samples](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). These formulas are valid when the population size is much larger (at least 20 times larger) than the sample size.

|  |  |
| --- | --- |
| Statistic | Standard Deviation |
| Sample mean, x | σx = σ / sqrt( n ) |
| Sample proportion, p | σp = sqrt [ P(1 - P) / n ] |
| Difference between means, x1 - x2 | σx1-x2 = sqrt [ σ21 / n1 + σ22 / n2 ] |
| Difference between proportions, p1 - p2 | σp1-p2 = sqrt [ P1(1-P1) / n1 + P2(1-P2) / n2 ] |

**Note:** In order to compute the standard deviation of a sample statistic, you must know the value of one or more population parameters.

Standard Error of Sample Estimates

Sadly, the values of population parameters are often unknown, making it impossible to compute the standard deviation of a statistic. When this occurs, use the standard error.

The standard error is computed from known sample statistics. The table below shows how to compute the standard error for simple random samples, assuming the population size is at least 20 times larger than the sample size.

|  |  |
| --- | --- |
| Statistic | Standard Error |
| Sample mean, x | SEx = s / sqrt( n ) |
| Sample proportion, p | SEp = sqrt [ p(1 - p) / n ] |
| Difference between means, x1 - x2 | SEx1-x2 = sqrt [ s21 / n1 + s22 / n2 ] |
| Difference between proportions, p1 - p2 | SEp1-p2 = sqrt [ p1(1-p1) / n1 + p2(1-p2) / n2 ] |

The equations for the standard error are identical to the equations for the standard deviation, except for one thing - the standard error equations use statistics where the standard deviation equations use parameters. Specifically, the standard error equations use *p* in place of *P*, and *s* in place of σ.

Test Your Understanding

**Problem 1**

Which of the following statements is true.

I. The standard error is computed solely from sample attributes.   
II. The standard deviation is computed solely from sample attributes.   
III. The standard error is a measure of central tendency.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above.   
(E) None of the above.

**Solution**

The correct answer is (A). The standard error can be computed from a knowledge of sample attributes - sample size and sample statistics. The standard deviation *cannot* be computed solely from sample attributes; it requires a knowledge of one or more population parameters. The standard error is a measure of variability, *not* a measure of central tendency.

Margin of Error

In a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval), the range of values above and below the sample statistic is called the **margin of error**.

For example, suppose we wanted to know the percentage of adults that exercise daily. We could devise a [sample design](http://stattrek.com/Help/Glossary.aspx?Target=Sample%20design) to ensure that our sample estimate will not differ from the true population value by more than, say, 5 percent (the margin of error) 90 percent of the time (the [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20level)).

How to Compute the Margin of Error

The margin of error can be defined by either of the following equations.

Margin of error = Critical value x Standard deviation of the statistic   
  
Margin of error = Critical value x Standard error of the statistic

If you know the standard deviation of the statistic, use the first equation to compute the margin of error. Otherwise, use the second equation. Previously, we described [how to compute the standard deviation and standard error](http://stattrek.com/AP-Statistics-4/Standard-Error.aspx?Tutorial=stat).

How to Find the Critical Value

The **critical value** is a factor used to compute the margin of error. This section describes how to find the critical value, when the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the statistic is [normal](http://stattrek.com/Help/Glossary.aspx?Target=Normal) or nearly normal.

The [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem) states that the sampling distribution of a statistic will be nearly normal, if the sample size is large enough. As a rough guide, many statisticians say that a sample size of 30 is large enough when the population distribution is bell-shaped. But if the original population is badly skewed, has multiple peaks, and/or has outliers, researchers like the sample size to be even larger.

When the sampling distribution is nearly normal, the critical value can be expressed as a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) or as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). When the sample size is smaller, the critical value should only be expressed as a t statistic.

To find the critical value, follow these steps.

* Compute alpha (α): α = 1 - (confidence level / 100)
* Find the critical probability (p\*): p\* = 1 - α/2
* To express the critical value as a z score, find the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to the critical probability (p\*).
* To express the critical value as a t statistic, follow these steps.
  + Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF). When estimating a mean score or a proportion from a single sample, DF is equal to the sample size minus one. For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.
  + The critical t statistic (t\*) is the t statistic having degrees of freedom equal to DF and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to the critical probability (p\*).

T-Score vs. Z-Score

Should you express the critical value as a t statistic or as a z-score? One way to answer this question focuses on the population standard deviation.

* If the population standard deviation is known, use the z-score.
* If the population standard deviation is unknown, use the t statistic.

Another approach focuses on sample size.

* If the sample size is large, use the z-score. (The [central limit theorem](http://stattrek.com/statistics/dictionary.aspx?definition=central_limit_theorem) provides a useful basis for determining whether a sample is "large".)
* If the sample size is small, use the t statistic.

In practice, researchers employ a mix of the above guidelines. On this site, we use z-scores when the population standard deviation is known and the sample size is large. Otherwise, we use the t statistics, unless the sample size is small and the underlying distribution is not normal.

**Warning:** If the sample size is small and the population distribution is not normal, we cannot be confident that the sampling distribution of the statistic will be normal. In this situation, neither the t statistic nor the z-score should be used to compute critical values.

You can use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to find the critical z score, and the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to find the critical t statistic. You can also use a [graphing calculator](http://stattrek.com/AP/Calculator.aspx) or standard statistical tables (found in the appendix of most introductory statistics texts).

Test Your Understanding

**Problem 1**

Nine hundred (900) high school freshmen were randomly selected for a national survey. Among survey participants, the mean grade-point average (GPA) was 2.7, and the standard deviation was 0.4. What is the margin of error, assuming a 95% confidence level?

(A) 0.013   
(B) 0.025   
(C) 0.500   
(D) 1.960   
(E) None of the above.

**Solution**

The correct answer is (B). To compute the margin of error, we need to find the critical value and the standard error of the mean. To find the critical value, we take the following steps.

* Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 0.95 = 0.05
* Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
* Find the degrees of freedom (df): df = n - 1 = 900 -1 = 899
* Find the critical value.
* Find the critical value. Since we don't know the population standard deviation, we'll express the critical value as a t statistic. For this problem, it will be the t statistic having 899 degrees of freedom and a cumulative probability equal to 0.975. Using the [t Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx), we find that the critical value is 1.96.

Next, we find the standard error of the mean, using the following equation:

SEx = s / sqrt( n ) = 0.4 / sqrt( 900 ) = 0.4 / 30 = 0.013

And finally, we compute the margin of error (ME).

ME = Critical value x Standard error = 1.96 \* 0.013 = 0.025

This means we can be 95% confident that the mean grade point average in the population is 2.7 plus or minus 0.025, since the margin of error is 0.025.

**Note:** The larger the sample size, the more closely the t distribution looks like the normal distribution. For this problem, since the sample size is very large, we would have found the same result with a z-score as we found with a t statistic. That is, the critical value would still have been 1.96. The choice of t statistic versus z-score does not make much practical difference when the sample size is very large.

What is a Confidence Interval?

Statisticians use a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) to describe the amount of uncertainty associated with a sample estimate of a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter).

How to Interpret Confidence Intervals

Suppose that a 90% confidence interval states that the population mean is greater than 100 and less than 200. How would you interpret this statement?

Some people think this means there is a 90% chance that the population mean falls between 100 and 200. This is incorrect. Like any population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter), the population mean is a constant, not a [random variable](http://stattrek.com/Help/Glossary.aspx?Target=Random%20variable). It does not change. The probability that a constant falls within any given range is always 0.00 or 1.00.

The [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=confidence_level) describes the uncertainty associated with a *sampling method*. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; A 95% confidence level means that 95% of the intervals would include the parameter; and so on.

Confidence Interval Data Requirements

To express a confidence interval, you need three pieces of information.

* [Confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20level)
* [Statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic)
* [Margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin%20of%20error)

Given these inputs, the range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty associated with the confidence interval is specified by the confidence level.

Often, the margin of error is not given; you must calculate it. Previously, we described [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-of-Error.aspx).

How to Construct a Confidence Interval

There are four steps to constructing a confidence interval.

* Identify a sample statistic. Choose the statistic (e.g, sample mean, sample proportion) that you will use to estimate a population parameter.
* Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.

Margin of error = Critical value \* Standard deviation of statistic   
Margin of error = Critical value \* Standard error of statistic

For guidance, see [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx).

* Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.

Confidence interval = sample statistic + Margin of error

The sample problem in the next section applies the above four steps to construct a 95% confidence interval for a mean score. The next few lessons discuss this topic in greater detail.

Sample Planning Wizard

As you may have guessed, the four steps required to specify a confidence interval can involve many time-consuming computations. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and error-free. In addition to constructing a confidence interval, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to construct a confidence interval, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>**[Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Sample Planning Wizard](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=False) |  | [Register Now](http://stattrek.com/LogIn/ChooseLicense.aspx) |  | [View Demo](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=True) |

Test Your Understanding

**Problem 1**

Suppose we want to estimate the average weight of an adult male in Dekalb County, Georgia. We draw a random sample of 1,000 men from a population of 1,000,000 men and weigh them. We find that the average man in our sample weighs 180 pounds, and the standard deviation of the sample is 30 pounds. What is the 95% confidence interval.

(A) 180 + 1.86   
(B) 180 + 3.0   
(C) 180 + 5.88   
(D) 180 + 30   
(E) None of the above

**Solution**

The correct answer is (A). To specify the confidence interval, we work through the four steps below.

* Identify a sample statistic. Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (180) as the sample statistic.
* Select a confidence level. In this case, the confidence level is defined for us in the problem. We are working with a 95% confidence level.
* Find the margin of error. Previously, we described [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx). The key steps are shown below.
  + Find standard error. The standard error (SE) of the mean is:

SE = s / sqrt( n ) = 30 / sqrt(1000) = 30/31.62 = 0.95

* + Find critical value. The critical value is a factor used to compute the margin of error. To express the critical value as a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) (t\*), follow these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 0.05
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 1000 - 1 = 999
    - The critical value is the t statistic having 999 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.975. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 1.96.

**Note:** We might also have expressed the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). Because the sample size is large, a z score analysis produces the same result - a critical value equal to 1.96.

* + Compute margin of error (ME): ME = critical value \* standard error = 1.96 \* 0.95 = 1.86
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level. Therefore, this 95% confidence interval is 180 + 1.86.

Confidence Interval: Proportion (Large Sample)

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for a sample proportion, *p*, when the sample size is large.

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The sample is sufficiently large. As a rule of thumb, a sample is considered "sufficiently large" if it includes at least 10 successes and 10 failures.

Note the implications of the second condition. If the population proportion were close to 0.5, the sample size required to produce at least 10 successes and at least 10 failures would probably be close to 20. But if the population proportion were extreme (i.e., close to 0 or 1), a much larger sample would probably be needed to produce at least 10 successes and 10 failures.

For example, imagine that the probability of success were 0.1, and the sample were selected using simple random sampling. In this situation, a sample size close to 100 might be needed to get 10 successes.

The Variability of the Sample Proportion

To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for a sample proportion, we need to know the variability of the sample proportion. This means we need to know how to compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) and/or the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).

* Suppose *k* possible samples of size *n* can be selected from the population. The standard deviation of the sampling distribution is the "average" deviation between the *k* sample proportions and the true population proportion, P. The standard deviation of the sample proportion σp is:

σp = sqrt[ P \* ( 1 - P ) / n ] \* sqrt[ ( N - n ) / ( N - 1 ) ]

where P is the population proportion, n is the sample size, and N is the population size. When the population size is much larger (at least 20 times larger) than the sample size, the standard deviation can be approximated by:

σp = sqrt[ P \* ( 1 - P ) / n ]

* When the true population proportion *P* is not known, the standard deviation of the sampling distribution cannot be calculated. Under these circumstances, use the standard error. The standard error (SE) can be calculated from the equation below.

SEp = sqrt[ p \* ( 1 - p ) / n ] \* sqrt[ ( N - n ) / ( N - 1 ) ]

where p is the sample proportion, n is the sample size, and N is the population size. When the population size at least 20 times larger than the sample size, the standard error can be approximated by:

SEp = sqrt[ p \* ( 1 - p ) / n ]

Alert

The Advanced Placement Statistics Examination only covers the "approximate" formulas for the standard deviation and standard error. However, students are expected to be aware of the limitations of these formulas; namely, the approximate formulas should only be used when the population size is at least 20 times larger than the sample size.

How to Find the Confidence Interval for a Proportion

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). For convenience, we repeat the key steps below.

* Identify a sample statistic. Use the sample proportion to estimate the population proportion.
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx).
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

In the next section, we work through a problem that shows how to use this approach to construct a confidence interval for a proportion.

Sample Planning Wizard

As you may have noticed, the steps required to estimate a population proportion are not trivial. They can be time-consuming and complex. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and error-free. In addition to constructing a confidence interval, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to construct a confidence interval, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Register Now](http://stattrek.com/LogIn/ChooseLicense.aspx) |  | [View Demo](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=True) |  | [View Wizard](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=False) |

Test Your Understanding

**Problem 1**

A major metropolitan newspaper selected a simple random sample of 1,600 readers from their list of 100,000 subscribers. They asked whether the paper should increase its coverage of local news. Forty percent of the sample wanted more local news. What is the 99% confidence interval for the proportion of readers who would like more coverage of local news?

(A) 0.30 to 0.50   
(B) 0.32 to 0.48   
(C) 0.35 to 0.45   
(D) 0.37 to 0.43   
(E) 0.39 to 0.41

**Solution**

The answer is (D). The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* The sample should include at least 10 successes and 10 failures. Suppose we classify a "more local news" response as a success, and any other response as a failure. Then, we have 0.40 \* 1600 = 640 successes, and 0.60 \* 1600 = 960 failures - plenty of successes and failures.
* If the population size is much larger than the sample size, we can use an "approximate" formula for the standard deviation or the standard error. This condition is satisfied, so we will use one of the simpler "approximate" formulas.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate a population proportion, we choose the sample proportion (0.40) as the sample statistic.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 99% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard deviation or standard error. Since we do not know the population proportion, we cannot compute the standard deviation; instead, we compute the standard error. And since the population is more than 20 times larger than the sample, we can use the following formula to compute the standard error (SE) of the proportion:

SE = sqrt [ p(1 - p) / n ] = sqrt [ (0.4)\*(0.6) / 1600 ] = sqrt [ 0.24/1600 ] = 0.012

* + Find critical value. The critical value is a factor used to compute the margin of error. Because the sampling distribution is approximately normal and the sample size is large, we can express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score) by following these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - (99/100) = 0.01
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.01/2 = 0.995
    - Find the degrees of freedom (df): df = n - 1 = 1600 -1 = 1599
    - Find the critical value. Since we don't know the population standard deviation, we'll express the critical value as a t statistic. For this problem, it will be the t statistic having 1599 degrees of freedom and a cumulative probability equal to 0.995. Using the [t Distribution Calculator](http://stattrek.com/online-calculator/t-distribution.aspx), we find that the critical value is 2.58.
  + Compute margin of error (ME): ME = critical value \* standard error = 2.58 \* 0.012 = 0.03
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 99% confidence interval is 0.37 to 0.43. That is, the 99% confidence interval is the range defined by 0.4 + 0.03.

Estimating a Proportion, Given a Small Sample

In this lesson, we explain how to estimate a confidence interval for a proportion, when the sample size is small.

Confidence Interval: Proportion (Small Sample)

In the [previous lesson](http://stattrek.com/Lesson4/Proportion.aspx?Tutorial=stat), we showed how to estimate a confidence interval for a proportion when a simple random sample includes at least 10 successes and 10 failures.

When the sample does not include at least 10 successes and 10 failures, the sample size will often be too small to justify the estimation approach presented in the previous lesson. This lesson describes how to construct a confidence interval for a proportion when the sample has fewer than 10 successes and/or fewer than 10 failures. The key steps are:

* Use the sample proportion as a [point estimate](http://stattrek.com/Help/Glossary.aspx?Target=Point_estimate) of the population proportion.
* Determine whether the sample proportion is the outcome of a [binomial experiment](http://stattrek.com/Help/Glossary.aspx?Target=Binomial_experiment) or a [hypergeometric experiment](http://stattrek.com/Help/Glossary.aspx?Target=Hypergeometric_experiment).
* Based on the above two bullet points, define the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the proportion.
* Use the sampling distribution to develop a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval).

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The sample includes at least 1 success and 1 failure.

The following examples illustrate how this works. The first example involves a binomial experiment; and the second example, a hypergeometric experiment.

Example 1: Find Confidence Interval When Sampling With Replacement

Suppose an urn contains 30 marbles. Some marbles are red, and the rest are green. Five marbles are randomly selected, [with replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_with_replacement), from the urn. Two of the selected marbles are red, and three are green. Construct an 80% confidence interval for the proportion of red marbles in the urn.

*Solution:* To solve this problem, we need to define the sampling distribution of the proportion.

* First, we assume that the population proportion is equal to the sample proportion. Thus, since 2 of the 5 marbles were red, we assume the proportion of red marbles is equal to 0.4.
* Second, since we sampled with replacement, the sample proportion can be considered an outcome of a binomial experiment.
* Assuming that the population proportion is 0.4 and the sample proportion is the outcome of a binomial experiment, the sampling distribution of the proportion can be determined. It appears in the table below. (Previously, we showed [how to compute binomial probabilities](http://stattrek.com/Lesson2/Binomial.aspx?Tutorial=stat) that form the body of the table.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of red marbles in sample** | **Sample proportion** | **Probability** | **Cumulative probability** |
| 0 | 0.0 | 0.07776 | 0.07776 |
| 1 | 0.2 | 0.2592 | 0.3396 |
| 2 | 0.4 | 0.3456 | 0.68256 |
| 3 | 0.6 | 0.2304 | 0.91296 |
| 4 | 0.8 | 0.0768 | 0.98976 |
| 5 | 1.0 | 0.01024 | 1.00 |

We see that the probability of getting 0 red marbles in the sample is 0.07776; the probability of getting 1 red marble is 0.2592; etc. Given the entries in the above table, it is not possible to create an 80% confidence interval *exactly*. However, we can come close. When the true population proportion is 0.4, the probability the probability that a sample proportion falls between 0.2 and 0.6 is equal to 0.2592 + 0.3456 + 0.2304 or 0.8352. Thus, based on this sample, we can say that an 83.52% confidence interval is described by the range from 0.2 to 0.6.

Sample Planning Wizard

As you may have noticed, the steps required to construct a sampling distribution based on a binomial experiment are not trivial. They can be time-consuming and complex. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and error-free. In addition to constructing a confidence interval, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to construct a confidence interval, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Register Now](http://stattrek.com/LogIn/ChooseLicense.aspx) |  | [View Demo](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=True) |  | [View Wizard](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=False) |

Example 2: Find Confidence Interval When Sampling Without Replacement

Let's take another look at the problem from Example 1. This time, however, we will assume that the marbles are sampled [without replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_without_replacement). Suppose an urn contains 30 marbles. Some marbles are red, and the rest are green. Five marbles are randomly selected, without replacement, from the urn. Two of the selected marbles are red, and three are green. Construct an 80% confidence interval for the proportion of red marbles in the urn.

*Solution:* To solve this problem, we need to define the sampling distribution of the proportion.

* First, we assume that the population proportion is equal to the sample proportion. Thus, since 2 of the 5 marbles were red, we assume the proportion of red marbles is equal to 0.4.
* Second, since we sampled without replacement, the sample proportion can be considered an outcome of a hypergeometric experiment.
* Assuming that the population proportion is 0.4 and the sample proportion is the outcome of a hypergeometric experiment, the sampling distribution of the proportion can be determined. It appears in the table below. (Previously, we showed [how to compute hypergeometric probabilities](http://stattrek.com/Lesson2/Hypergeometric.aspx?Tutorial=stat) that form the body of the table.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of red marbles in sample** | **Sample proportion** | **Probability** | **Cumulative probability** |
| 0 | 0.0 | 0.0601 | 0.0601 |
| 1 | 0.2 | 0.2577 | 0.3178 |
| 2 | 0.4 | 0.3779 | 0.6957 |
| 3 | 0.6 | 0.2362 | 0.9319 |
| 4 | 0.8 | 0.0625 | 0.9944 |
| 5 | 1.0 | 0.0056 | 1.0000 |

We see that the probability of getting 0 red marbles in the sample is 0.0601; the probability of getting 1 red marble is 0.2577; etc. Given the entries in the above table, it is not possible to create an 80% confidence interval *exactly*. However, we can come close. When the true population proportion is 0.4, the probability the probability that a sample proportion falls between 0.2 and 0.6 is equal to 0.2577 + 0.3779 + 0.2362 or 0.8718. Thus, based on this sample, we can say that an 87.18% confidence interval is described by the range from 0.2 to 0.6.

It is informative to compare the findings from Examples 1 and 2. In both problems, the [interval estimate](http://stattrek.com/Help/Glossary.aspx?Target=Interval_estimate) ranged from 0.2 to 0.6. However, the [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level) was greater for Example 2 (which sampled without replacement) than for Example 1 (which sampled with replacement). This illustrates the fact that precision is greater when sampling without replacement than when sampling with replacement.

Confidence Interval: Difference Between Proportions

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for the difference between two sample proportions, *p*1 - *p*2.

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* Both samples are [simple random samples](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).
* Each sample includes at least 10 successes and 10 failures.

The Variability of the Difference Between Proportions

To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for the difference between two sample proportions, we need to know about the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the difference. Specifically, we need to know how to compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation) or [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error) of the sampling distribution.

* The standard deviation of the sampling distribution is the "average" deviation between all possible sample differences (*p*1 - *p*2) and the true population difference, (*P*1 - *P*2). The standard deviation of the difference between sample proportions σp1 - p2 is:

σp1 - p2 =   
sqrt{ [P1 \* (1 - P1) / n1] \* [(N1 - n1) / (N1 - 1)] + [P2 \* (1 - P2) / n2] \* [(N2 - n2) / (N2 - 1)] }

where P1 is the population proportion for sample 1, P2 is the population proportion for sample 2, n1 is the sample size from population 1, n2 is the sample size from population 2, N1 is the number of observations in population 1, and N2 is the number of observations in population 2. When each sample is small (less than 5% of its population), the standard deviation can be approximated by:

σp1 - p2 = sqrt{ [P1 \* (1 - P1) / n1] + [P2 \* (1 - P2) / n2] }

* When the population parameters (P1 and P2) are not known, the standard deviation of the sampling distribution cannot be calculated. Under these circumstances, use the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). The standard error (SE) can be calculated from the equation below.

SEp1 - p2 =   
sqrt{ [p1 \* (1 - p1) / n1] \* [(N1 - n1) / (N1 - 1)] + [p2 \* (1 - p2) / n2] \* [(N2 - n2) / (N2 - 1)] }

where p1 is the sample proportion for sample 1, and where p2 is the sample proportion for sample 2. When each sample is small (less than 5% of its population), the standard deviation can be approximated by:

SEp1 - p2 = sqrt{ [p1 \* (1 - p1) / n1] + [p2 \* (1 - p2) / n2] }

**Note:** The Advanced Placement Statistics Examination only covers the "approximate" formulas for the standard deviation and standard error. However, students are expected to be aware of the limitations of these formulas; namely, that they should only be used when each population is at least 20 times larger than its respective sample.

How to Find the Confidence Interval for a Proportion

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). For convenience, we repeat the key steps below.

* Identify a sample statistic. Use the sample proportions (p1 - p2) to estimate the difference between population proportions (P1 - P2).
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx).
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

In the next section, we work through a problem that shows how to use this approach to construct a confidence interval for the difference between proportions.

Test Your Understanding

**Problem 1**

Suppose the Cartoon Network conducts a nation-wide survey to assess viewer attitudes toward Superman. Using a simple random sample, they select 400 boys and 300 girls to participate in the study. Forty percent of the boys say that Superman is their favorite character, compared to thirty percent of the girls. What is the 90% confidence interval for the true difference in attitudes toward Superman?

(A) 0 to 20 percent more boys prefer Superman   
(B) 2 to 18 percent more boys prefer Superman   
(C) 4 to 16 percent more boys prefer Superman   
(D) 6 to 14 percent more boys prefer Superman   
(E) None of the above

**Solution**

The correct answer is (C). The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* Both samples should be independent. This condition is satisfied since neither sample was affected by responses of the other sample.
* The sample should include at least 10 successes and 10 failures. Suppose we classify choosing Superman as a success, and any other response as a failure. Then, we have plenty of successes and failures in both samples.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) should be approximately normally distributed. Because each sample size is large, we know from the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem) that the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the difference between sample proportions will be [normal](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution) or nearly normal; so this condition is satisfied.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate the difference between population proportions, we choose the difference between sample proportions as the sample statistic. Thus, the sample statistic is pboy - pgirl = 0.40 - 0.30 = 0.10.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard deviation or standard error. Since we do not know the population proportions, we cannot compute the standard deviation; instead, we compute the standard error. And since each population is more than 20 times larger than its sample, we can use the following formula to compute the standard error (SE) of the difference between proportions:

SE = sqrt{ [p1 \* (1 - p1) / n1] + [p2 \* (1 - p2) / n2] }   
SE = sqrt{ [0.40 \* 0.60 / 400] + [0.30 \* 0.70 / 300] }   
SE = sqrt[ (0.24 / 400) + (0.21 / 300) ] = sqrt(0.0006 + 0.0007) = sqrt(0.0013) = 0.036

* + Find critical value. The critical value is a factor used to compute the margin of error. Because the sampling distribution is approximately normal and the sample sizes are large, we can express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score) by following these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - (90/100) = 0.10
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.10/2 = 0.95
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.95. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 1.645.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.645 \* 0.036 = 0.06
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 90% confidence interval is 0.04 to 0.16. That is, we are 90% confident that the true difference between population proportion is in the range defined by 0.10 + 0.06. Since both ends of the confidence interval are positive, we can conclude that more boys than girls choose Superman as their favorite cartoon character.

Confidence Interval: Sample Mean

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) around a sample mean, x.

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed when the sample size is greater than or equal to 30.

The Variability of the Sample Mean

To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for a sample mean, we need to know the variability of the sample mean. This means we need to know how to compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) or the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).

* Suppose *k* possible samples of size *n* can be selected from a population of size *N*. The standard deviation of the sampling distribution is the "average" deviation between the *k* sample means and the true population mean, μ. The standard deviation of the sample mean σx is:

σx = σ \* sqrt{ ( 1/n ) \* ( 1 - n/N ) \* [ N / ( N - 1 ) ] }

where σ is the standard deviation of the population, *N* is the population size, and *n* is the sample size. When the population size is much larger (at least 20 times larger) than the sample size, the standard deviation can be approximated by:

σx = σ / sqrt( n )

* When the standard deviation of the population σ is unknown, the standard deviation of the sampling distribution cannot be calculated. Under these circumstances, use the standard error. The standard error (SE) can be calculated from the equation below.

SEx = s \* sqrt{ ( 1/n ) \* ( 1 - n/N ) \* [ N / ( N - 1 ) ] }

where *s* is the standard deviation of the sample, N is the population size, and *n* is the sample size. When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by:

SEx = s / sqrt( n )

Note: In real-world analyses, the standard deviation of the population is seldom known. Therefore, the standard error is used more often than the standard deviation.

Alert

The Advanced Placement Statistics Examination only covers the "approximate" formulas for the standard deviation and standard error. However, students are expected to be aware of the limitations of these formulas; namely, the approximate formulas should only be used when the population size is at least 20 times larger than the sample size.

How to Find the Confidence Interval for a Mean

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). For convenience, we repeat the key steps below.

* Identify a sample statistic. Use the sample mean to estimate the population mean.
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx).
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

In the next section, we work through a problem that shows how to use this approach to construct a confidence interval to estimate a population mean.

Sample Planning Wizard

As you may have noticed, the steps required to construct a confidence interval for a mean score require many time-consuming computations. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and error-free. In addition to constructing a confidence interval, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to construct a confidence interval, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Register Now](http://stattrek.com/LogIn/ChooseLicense.aspx) |  | [View Demo](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=True) |  | [View Wizard](http://stattrek.com/SPWizard/SamplePlanningWizard.aspx?ShowDemo=False) |

Test Your Understanding

**Problem 1**

Suppose a simple random sample of 150 students is drawn from a population of 3000 college students. Among sampled students, the average IQ score is 115 with a standard deviation of 10. What is the 99% [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval) for the students' IQ score?

(A) 115 + 0.01   
(B) 115 + 0.82   
(C) 115 + 2.1   
(D) 115 + 2.6   
(E) None of the above

**Solution**

The correct answer is (C). The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) should be approximately normally distributed. Because the sample size is large, we know from the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem) that the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the mean will be [normal](http://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution) or nearly normal; so this condition is satisfied.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate a population mean, we choose the sample mean (115) as the sample statistic.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 99% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard deviation or standard error. Since we do not know the standard deviation of the population, we cannot compute the standard deviation of the sample mean; instead, we compute the standard error (SE). Because the sample size is much smaller than the population size, we can use the "approximate" formula for the standard error.

SE = s / sqrt( n ) = 10 / sqrt(150) = 10 / 12.25 = 0.82

* + Find critical value. The critical value is a factor used to compute the margin of error. For this example, we'll express the critical value as a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 99/100 = 0.01
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.01/2 = 0.995
    - Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 150 - 1 = 149
    - The critical value is the t statistic having 149 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.995. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.61.

Note: We might also have expressed the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). Because the sample size is fairly large, a z score analysis produces a similar result - a critical value equal to 2.58.

* + Compute margin of error (ME): ME = critical value \* standard error = 2.61 \* 0.82 = 2.1
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 99% confidence interval is 112.9 to 117.1. That is, we are 99% confident that the true population mean is in the range defined by 115 + 2.1.

Confidence Interval: Difference Between Means

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for the difference between two means.

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* Both samples are [simple random samples](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).
* Each [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) is at least 20 times larger than its respective [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample).
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the difference between means is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed when the sample size is greater than or equal to 30.

The Variability of the Difference Between Sample Means

To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval), we need to know the variability of the difference between sample means. This means we need to know how to compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the difference.

* If the population standard deviations are known, the standard deviation of the sampling distribution is:

σx1-x2 = sqrt [ σ21 / n1 + σ22 / n2 ]

where σ1 is the standard deviation of the population 1, σ2 is the standard deviation of the population 2, and n1 is the size of sample 1, and n2 is the size of sample 2.

* When the standard deviation of either population is unknown and the sample sizes (n1 and n2) are large, the standard deviation of the sampling distribution can be estimated by the standard error, using the equation below.

SEx1-x2 = sqrt [ s21 / n1 + s22 / n2 ]

where SE is the standard error, s1 is the standard deviation of the sample 1, s2 is the standard deviation of the sample 2, and n1 is the size of sample 1, and n2 is the size of sample 2.

**Note:** In real-world analyses, the standard deviation of the population is seldom known. Therefore, SEx1-x2 is used more often than σx1-x2.

Alert

Some texts present additional options for calculating standard deviations. These formulas, which should only be used under special circumstances, are described below.

* Standard deviation. Use this formula when the population standard deviations are known and are equal.   
  σx1 - x2 = σd = σ \* sqrt[ (1 / n1) + (1 / n2)] where σ = σ1 = σ2
* Pooled standard deviation. Use this formula when the population standard deviations are unknown, but assumed to be equal; and the samples sizes (n1) and (n2) are small (under 30).   
  SDpooled = sqrt{ [ (n1 -1) \* s12) + (n2 -1) \* s22) ] / (n1 + n2 - 2) } where σ1 = σ2

Remember, these two formulas should be used only when the various required underlying assumptions are justified.

How to Find the Confidence Interval for the Difference Between Means

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). For convenience, we repeat the key steps below.

* Identify a sample statistic. Use the difference between sample means to estimate the difference between population means.
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx), based on the [critical value](http://stattrek.com/Help/Glossary.aspx?Target=Critical%20value) and standard deviation.

When the sample size is large, you can use a t statistic or a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score) for the critical value. Since it does not require computing degrees of freedom, the z score is a little easier. When the sample sizes are small (less than 40), use a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) for the critical value.

If you use a t statistic, you will need to compute [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF). Here's how.

* + The following formula is appropriate whenever a t statistic is used to analyze the difference between means.

DF = (s12/n1 + s22/n2)2 / { [ (s12 / n1)2 / (n1 - 1) ] + [ (s22 / n2)2 / (n2 - 1) ] }

* + If you are working with a pooled standard deviation (see above), DF = n1 + n2 - 2.

The next section presents sample problems that illustrate how to use z scores and t statistics as critical values.

* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Test Your Understanding

**Problem 1: Small Samples**

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.

What is the 90% [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval) for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools? (Hint: Since the sample sizes are small, use a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) as the [critical value](http://stattrek.com/Help/Glossary.aspx?Target=critical%20value).)

(A) 50 + 1.70   
(B) 50 + 28.49   
(C) 50 + 32.74   
(D) 50 + 55.66   
(E) None of the above

**Solution**

The correct answer is (D). The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* The samples must be [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent). Since responses from one sample did not affect responses from the other sample, the samples are independent.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) should be approximately normally distributed. The problem states that test scores in each population are normally distributed, so the difference between test scores will also be normally distributed.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate the difference between population means, we choose the difference between sample means as the sample statistic. Thus, x1 - x2 = 1000 - 950 = 50.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard error. Using the sample standard deviations, we compute the standard error (SE), which is an estimate of the standard deviation of the difference between sample means.

SE = sqrt [ s21 / n1 + s22 / n2 ]   
SE = sqrt [(100)2 / 15 + (90)2 / 20]   
SE = sqrt (10,000/15 + 8100/20) = sqrt(666.67 + 405) = 32.74

* + Find critical value. The critical value is a factor used to compute the margin of error. Because the sample sizes are small, we express the critical value as a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) rather than a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 90/100 = 0.10
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.10/2 = 0.95
    - Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df):

DF = (s12/n1 + s22/n2)2 / { [ (s12 / n1)2 / (n1 - 1) ] + [ (s22 / n2)2 / (n2 - 1) ] }   
DF = (1002/15 + 902/20)2 / { [ (1002 /15)2 / 14 ] + [ (902 /20)2 / 19 ] }   
DF = (666.67 + 405}2 / (31746.03 + 8632.89) = 1150614.5 / 40378.92 = 28.495

Rounding off to the nearest whole number, we conclude that there are 28 degrees of freedom.

* + - The critical value is the t statistic having 28 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.95. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 1.7.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.7 \* 32.74 = 55.66
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 90% confidence interval is 50 + 55.66; that is, -5.66 to 105.66. Here's how to interpret this confidence interval. Suppose we repeated this study with different random samples for school A and school B. Based on the confidence interval, we would expect the observed difference in sample means to be between -5.66 and 105.66 90% of the time.

**Problem 2: Large Samples**

The local baseball team conducts a study to find the amount spent on refreshments at the ball park. Over the course of the season they gather simple random samples of 500 men and 1000 women. For men, the average expenditure was $20, with a standard deviation of $3. For women, it was $15, with a standard deviation of $2.

What is the 99% [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval) for the spending difference between men and women? Assume that the two populations are independent and normally distributed.

(A) $5 + $0.15   
(B) $5 + $0.38   
(C) $5 + $1.15   
(D) $5 + $1.38   
(E) None of the above

**Solution**

The correct answer is (B). The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* The samples must be [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent). Again, the problem statement satisfies this condition.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) should be approximately normally distributed. The problem states that test scores in each population are normally distributed, so the difference between test scores will also be normally distributed.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate the difference between population means, we choose the difference between sample means as the sample statistic. Thus, x1 - x2 = $20 - $15 = $5.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 99% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard error. The standard error is an estimate of the standard deviation of the difference between population means. We use the sample standard deviations to estimate the standard error (SE).

SE = sqrt [ s21 / n1 + s22 / n2 ]   
SE = sqrt [(3)2 / 500 + (2)2 / 1000] = sqrt (9/500 + 4/1000) = sqrt(0.018 + 0.004) = 0.148

* + Find critical value. The critical value is a factor used to compute the margin of error. Because the sample sizes are large enough, we express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 99/100 = 0.01
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.01/2 = 0.995
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.995. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 2.58.
  + Compute margin of error (ME): ME = critical value \* standard error = 2.58 \* 0.148 = 0.38
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 99% confidence interval is $5 + $0.38; that is, $4.62 to $5.38.

Mean Difference Between Matched Pairs

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) to estimate the mean difference between matched [data pairs](http://stattrek.com/Help/Glossary.aspx?Target=Paired%20data).

Estimation Requirements

The approach described in this lesson is valid whenever the following conditions are met:

* The data set is a [simple random sample](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling) of observations from the population of interest.
* Each element of the population includes measurements on two paired variables (e.g., *x* and *y*) such that the paired difference between *x* and *y* is: *d* = *x* - *y*.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the mean difference between data pairs (*d*) is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed if the sample is described by at least one of the following statements.

* The population distribution of paired differences (i.e., the variable *d*) is normal.
* The sample distribution of paired differences is [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is 15 or less.
* The sample distribution is moderately [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is between 16 and 40.
* The sample size is greater than 40, without outliers.

The Variability of the Mean Difference Between Matched Pairs

Suppose *d* is the mean difference between sample data pairs. To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for *d*, we need to know how to compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) and/or the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) for *d*.

* The standard deviation of the mean difference σd is:

σd = σd \* sqrt{ ( 1/n ) \* ( 1 - n/N ) \* [ N / ( N - 1 ) ] }

where σd is the standard deviation of the population difference, *N* is the population size, and *n* is the sample size. When the population size is much larger (at least 10 times larger) than the sample size, the standard deviation can be approximated by:

σd = σd / sqrt( n )

* When the standard deviation of the population σd is unknown, the standard deviation of the sampling distribution cannot be calculated. Under these circumstances, use the standard error. The standard error (SE) can be calculated from the equation below.

SEd = sd \* sqrt{ ( 1/n ) \* ( 1 - n/N ) \* [ N / ( N - 1 ) ] }

where *s*d is the standard deviation of the sample difference, *N* is the population size, and *n* is the sample size. When the population size is much larger (at least 10 times larger) than the sample size, the standard error can be approximated by:

SEd = sd / sqrt( n )

Note: In real-world analyses, the standard deviation of the population is seldom known. Therefore, the standard error is used more often than the standard deviation.

Alert

The Advanced Placement Statistics Examination only covers the "approximate" formulas for the standard deviation and standard error. However, students are expected to be aware of the limitations of these formulas; namely, the approximate formulas should only be used when the population size is at least 10 times larger than the sample size.

How to Find the Confidence Interval for Mean Difference With Paired Data

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). For convenience, we repeat the key steps below.

* Identify a sample statistic. Use the mean difference between sample data pairs (d to estimate the mean difference between population data pairs μd.
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx), based on the [critical value](http://stattrek.com/Help/Glossary.aspx?Target=Critical%20value) and standard deviation.

When the sample size is large, you can use a t statistic or a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score) for the critical value. Since it does not require computing degrees of freedom, the z score is a little easier. When the sample sizes are small (less than 40), use a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) for the critical value. (For additional explanation, see [choosing between a t statistic and a z-score.](http://stattrek.com/estimation/margin-of-error.aspx#TvsZ).)

If you use a t statistic, you will need to compute [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF). In this case, the degrees of freedom is equal to the sample size minus one: DF = n - 1.

* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Test Your Understanding

**Problem**

Twenty-two students were randomly selected from a population of 1000 students. The sampling method was simple random sampling. All of the students were given a standardized English test and a standardized math test. Test results are summarized below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Student | English | Math | Difference, d | (d - d)2 | | 1 | 95 | 90 | 5 | 16 | | 2 | 89 | 85 | 4 | 9 | | 3 | 76 | 73 | 3 | 4 | | 4 | 92 | 90 | 2 | 1 | | 5 | 91 | 90 | 1 | 0 | | 6 | 53 | 53 | 0 | 1 | | 7 | 67 | 68 | -1 | 4 | | 8 | 88 | 90 | -2 | 9 | | 9 | 75 | 78 | -3 | 16 | | 10 | 85 | 89 | -4 | 25 | | 11 | 90 | 95 | -5 | 36 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Student | English | Math | Difference, d | (d - d)2 | | 12 | 85 | 83 | 2 | 1 | | 13 | 87 | 83 | 4 | 9 | | 14 | 85 | 83 | 2 | 1 | | 15 | 85 | 82 | 3 | 4 | | 16 | 68 | 65 | 3 | 4 | | 17 | 81 | 79 | 2 | 1 | | 18 | 84 | 83 | 1 | 0 | | 19 | 71 | 60 | 11 | 100 | | 20 | 46 | 47 | -1 | 4 | | 21 | 75 | 77 | -2 | 9 | | 22 | 80 | 83 | -3 | 16 | |

Σ(d - d)2 = 270   
d = 1

Find the 90% confidence interval for the mean difference between student scores on the math and English tests. Assume that the mean differences are approximately normally distributed.

**Solution**

The approach that we used to solve this problem is valid when the following conditions are met.

* The sampling method must be [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling). This condition is satisfied; the problem statement says that we used simple random sampling.
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) should be approximately normally distributed. The problem statement says that the differences were normally distributed; so this condition is satisfied.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate a population mean difference in math and English test scores, we use the sample mean difference (d = 1) as the sample statistic.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard deviation or standard error. Since we do not know the standard deviation of the population, we cannot compute the standard deviation of the sample mean; instead, we compute the standard error (SE). Since the sample size is much smaller than the population size, we can use the approximation equation for the standard error.

sd = sqrt [ (Σ(di - d)2 / (n - 1) ] = sqrt[ 270/(22-1) ] = sqrt(12.857) = 3.586   
  
SE = sd / sqrt( n ) = 3.586 / [ sqrt(22) ] = 3.586/4.69 = 0.765

* + Find critical value. The critical value is a factor used to compute the margin of error. Because the sample size is small, we express the critical value as a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) rather than a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). (See [how to choose between a t statistic and a z-score](http://stattrek.com/m/estimation/margin-of-error.aspx#TvsZ).) To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 90/100 = 0.10
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.10/2 = 0.95
    - Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = n - 1 = 22 - 1 = 21
    - The critical value is the t statistic having 21 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.95. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 1.72.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.72 \* 0.765 = 1.3
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 90% confidence interval is -0.3 to 2.3 or 1 + 1.3.

# What is Hypothesis Testing?

A **statistical hypothesis** is an assumption about a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter). This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

Statistical Hypotheses

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

* **Null hypothesis**. The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.
* **Alternative hypothesis**. The alternative hypothesis, denoted by H1 or Ha, is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

H0: P = 0.5   
Ha: P ≠ 0.5

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

Can We Accept the Null Hypothesis?

Some researchers say that a hypothesis test can have one of two outcomes: you accept the null hypothesis or you reject the null hypothesis. Many statisticians, however, take issue with the notion of "accepting the null hypothesis." Instead, they say: you reject the null hypothesis or you fail to reject the null hypothesis.

Why the distinction between "acceptance" and "failure to reject?" Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis.

Hypothesis Tests

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called **hypothesis testing**, consists of four steps.

* State the hypotheses. This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
* Formulate an analysis plan. The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
* Analyze sample data. Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.
* Interpret results. Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Decision Errors

Two types of errors can result from a hypothesis test.

* **Type I error**. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by α.
* **Type II error**. A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β. The probability of *not* committing a Type II error is called the **Power** of the test.

Decision Rules

The analysis plan includes decision rules for rejecting the null hypothesis. In practice, statisticians describe these decision rules in two ways - with reference to a P-value or with reference to a region of acceptance.

* P-value. The strength of evidence in support of a null hypothesis is measured by the **P-value**. Suppose the test statistic is equal to *S*. The P-value is the probability of observing a test statistic as extreme as *S*, assuming the null hypotheis is true. If the P-value is less than the significance level, we reject the null hypothesis.
* Region of acceptance. The **region of acceptance** is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level.

The set of values outside the region of acceptance is called the **region of rejection**. If the test statistic falls within the region of rejection, the null hypothesis is rejected. In such cases, we say that the hypothesis has been rejected at the α level of significance.

These approaches are equivalent. Some statistics texts use the P-value approach; others use the region of acceptance approach. In subsequent lessons, this tutorial will present examples that illustrate each approach.

One-Tailed and Two-Tailed Tests

A test of a statistical hypothesis, where the region of rejection is on only one side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution), is called a **one-tailed test**. For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a **two-tailed test**. For example, suppose the null hypothesis states that the mean is equal to 10. The alternative hypothesis would be that the mean is less than 10 or greater than 10. The region of rejection would consist of a range of numbers located on both sides of sampling distribution; that is, the region of rejection would consist partly of numbers that were less than 10 and partly of numbers that were greater than 10.

How to Test Hypotheses

This lesson describes a general procedure that can be used to test statistical hypotheses.

How to Conduct Hypothesis Tests

All hypothesis tests are conducted the same way. The researcher states a hypothesis to be tested, formulates an analysis plan, analyzes sample data according to the plan, and accepts or rejects the null hypothesis, based on results of the analysis.

* **State the hypotheses.** Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.
* **Formulate an analysis plan.** The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.
  + Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
  + Test method. Typically, the test method involves a test statistic and a [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution). Computed from sample data, the test statistic might be a mean score, proportion, difference between means, difference between proportions, z-score, t statistic, chi-square, etc. Given a test statistic and its sampling distribution, a researcher can assess probabilities associated with the test statistic. If the test statistic probability is less than the significance level, the null hypothesis is rejected.
* **Analyze sample data.** Using sample data, perform computations called for in the analysis plan.
  + Test statistic. When the null hypothesis involves a mean or proportion, use either of the following equations to compute the test statistic.

Test statistic = (Statistic - Parameter) / (Standard deviation of statistic)   
Test statistic = (Statistic - Parameter) / (Standard error of statistic)

where *Parameter* is the value appearing in the null hypothesis, and *Statistic* is the [point estimate](http://stattrek.com/Help/Glossary.aspx?Target=Point%20estimate) of *Parameter*. As part of the analysis, you may need to compute the standard deviation or standard error of the statistic. Previously, we presented common [formulas for the standard deviation and standard error](http://stattrek.com/AP-Statistics-4/Standard-Error.aspx).  
  
When the parameter in the null hypothesis involves categorical data, you may use a chi-square statistic as the test statistic. Instructions for computing a chi-square test statistic are presented in the lesson on the [chi-square goodness of fit test](http://stattrek.com/AP-Statistics-4/Goodness-Of-Fit.aspx).

* + P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic, assuming the null hypotheis is true.
* **Interpret the results.** If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Applications of the General Hypothesis Testing Procedure

The next few lessons show how to apply the general hypothesis testing procedure to different kinds of statistical problems.

* [Proportions](http://stattrek.com/Lesson5/Proportion.aspx?Tutorial=stat)
* [Difference between proportions](http://stattrek.com/AP-Statistics-4/Test-Difference-Proportion.aspx?Tutorial=stat)
* [Proportions from small samples](http://stattrek.com/Lesson5/ProportionSmall.aspx?Tutorial=stat)
* [Regression slope](http://stattrek.com/AP-Statistics-4/Test-Slope.aspx?Tutorial=stat)
* [Means](http://stattrek.com/Lesson5/Mean.aspx?Tutorial=stat)
* [Difference between means](http://stattrek.com/AP-Statistics-4/Unpaired-Means.aspx?Tutorial=stat)
* [Difference between matched pairs](http://stattrek.com/AP-Statistics-4/Paired-Means.aspx?Tutorial=stat)
* [Goodness of fit](http://stattrek.com/AP-Statistics-4/Goodness-of-Fit.aspx?Tutorial=stat)
* [Homogeneity](http://stattrek.com/AP-Statistics-4/Homogeneity.aspx?Tutorial=stat)
* [Independence](http://stattrek.com/AP-Statistics-4/Independence.aspx?Tutorial=stat)

At this point, don't worry if the general procedure for testing hypotheses seems a little bit unclear. The procedure will be clearer after you read through a few of the examples presented in subsequent lessons.

Test Your Understanding

**Problem 1**

In hypothesis testing, which of the following statements is always true?

I. The P-value is greater than the significance level.   
II. The P-value is computed from the significance level.   
III. The P-value is the parameter in the null hypothesis.   
IV. The P-value is a test statistic.   
V. The P-value is a probability.

(A) I only   
(B) II only   
(C) III only   
(D) IV only   
(E) V only

**Solution**

The correct answer is (E). The P-value is the probability of observing a sample statistic as extreme as the test statistic. It can be greater than the significance level, but it can also be smaller than the significance level. It is not computed from the significance level, it is not the parameter in the null hypothesis, and it is not a test statistic.

Hypothesis Test for a Mean

This lesson explains how to conduct a hypothesis test of a mean, when the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The sampling distribution is normal or nearly normal.

Generally, the sampling distribution will be approximately normally distributed if any of the following conditions apply.

* The population distribution is normal.
* The population distribution is [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is 15 or less.
* The population distribution is moderately [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is between 16 and 40.
* The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of hypotheses. Each makes a statement about how the population mean μ is related to a specified value *M*. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | μ = M | μ ≠ M | 2 |
| 2 | μ > M | μ < M | 1 |
| 3 | μ < M | μ > M | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [one-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20t-test) to determine whether the hypothesized mean differs significantly from the observed sample mean.

Analyze Sample Data

Using sample data, conduct a one-sample t-test. This involves finding the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

* Standard error. Compute the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=standard%20error) (SE) of the sampling distribution.

SE = s \* sqrt{ ( 1/n ) \* [ ( N - n ) / ( N - 1 ) ] }

where *s* is the standard deviation of the sample, N is the population size, and *n* is the sample size. When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by:

SE = s / sqrt( n )

* Degrees of freedom. The degrees of freedom (DF) is equal to the sample size (n) minus one. Thus, DF = n - 1.
* Test statistic. The test statistic is a t statistic (t) defined by the following equation.

t = (x - μ) / SE

where x is the sample mean, μ is the hypothesized population mean in the null hypothesis, and SE is the standard error.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to assess the probability associated with the t statistic, given the degrees of freedom computed above. (See sample problems at the end of this lesson for examples of how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

In this section, two sample problems illustrate how to conduct a hypothesis test of a mean score. The first problem involves a two-tailed test; the second problem, a one-tailed test.

Sample Planning Wizard

As you probably noticed, the process of testing a hypothesis about a mean score can be complex and time-consuming. Stat Trek's Sample Planning Wizard can do the same job quickly, easily, and error-free. In addition to conducting the hypothesis test, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to test a hypothesis, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Problem 1: Two-Tailed Test**

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. From his stock of 2000 engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μ = 300   
Alternative hypothesis: μ ≠ 300

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample mean is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method is a [one-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20t-test).
* **Analyze sample data**. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

SE = s / sqrt(n) = 20 / sqrt(50) = 20/7.07 = 2.83   
DF = n - 1 = 50 - 1 = 49   
t = (x - μ) / SE = (295 - 300)/2.83 = -1.77

where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that the t statistic having 49 degrees of freedom is less than -1.77 or greater than 1.77.

We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < -1.77) = 0.04, and P(t > 1.77) = 0.04. Thus, the P-value = 0.04 + 0.04 = 0.08.

* **Interpret results**. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was normally distributed, and the sample size was small relative to the population size (less than 5%).

**Problem 2: One-Tailed Test**  
  
Bon Air Elementary School has 1000 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01. (Assume that test scores in the population of engines are normally distributed.)

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μ >= 110   
Alternative hypothesis: μ < 110

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the sample mean is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.01. The test method is a [one-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20t-test).
* **Analyze sample data**. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

SE = s / sqrt(n) = 10 / sqrt(20) = 10/4.472 = 2.236   
DF = n - 1 = 20 - 1 = 19   
t = (x - μ) / SE = (108 - 110)/2.236 = -0.894

where s is the standard deviation of the sample, x is the sample mean, μ is the hypothesized population mean, and n is the sample size.

Here is the logic of the analysis: Given the alternative hypothesis (μ < 110), we want to know whether the observed sample mean is small enough to cause us to reject the null hypothesis.

The observed sample mean produced a t statistic test statistic of -0.894. We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < -0.894) = 0.19. This means we would expect to find a sample mean of 108 or smaller in 19 percent of our samples, if the true population IQ were 110. Thus the P-value in this analysis is 0.19.

* **Interpret results**. Since the P-value (0.19) is greater than the significance level (0.01), we cannot reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was normally distributed, and the sample size was small relative to the population size (less than 5%).

Hypothesis Test: Difference Between Means

This lesson explains how to conduct a hypothesis test for the difference between two means. The test procedure, called the **two-sample t-test**, is appropriate when the following conditions are met:

* The sampling method for each sample is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).
* Each [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) is at least 20 times larger than its respective [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample).
* The sampling distribution is approximately normal, which is generally the case if any of the following conditions apply.
  + The population distribution is normal.
  + The population data are [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is 15 or less.
  + The population data are slightly [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is 16 to 40.
  + The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of null and alternative hypotheses. Each makes a statement about the difference *d* between the mean of one population μ1 and the mean of another population μ2. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | μ1 - μ2 = d | μ1 - μ2 ≠ d | 2 |
| 2 | μ1 - μ2 > d | μ1 - μ2 < d | 1 |
| 3 | μ1 - μ2 < d | μ1 - μ2 > d | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population means (i.e., d = 0), the null and alternative hypothesis are often stated in the following form.

H0: μ1 = μ2  
Ha: μ1 ≠ μ2

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [two-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=Two-sample%20t-test) to determine whether the difference between means found in the sample is significantly different from the hypothesized difference between means.

Analyze Sample Data

Using sample data, find the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

* Standard error. Compute the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=standard%20error) (SE) of the sampling distribution.

SE = sqrt[ (s12/n1) + (s22/n2) ]

where s1 is the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) of sample 1, s2 is the standard deviation of sample 2, n1 is the size of sample 1, and n2 is the size of sample 2.

* Degrees of freedom. The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is:

DF = (s12/n1 + s22/n2)2 / { [ (s12 / n1)2 / (n1 - 1) ] + [ (s22 / n2)2 / (n2 - 1) ] }

If DF does not compute to an integer, round it off to the nearest whole number. Some texts suggest that the degrees of freedom can be approximated by the smaller of n1 - 1 and n2 - 1; but the above formula gives better results.

* Test statistic. The test statistic is a t statistic (t) defined by the following equation.

t = [ (x1 - x2) - d ] / SE

where x1 is the mean of sample 1, x2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to assess the probability associated with the t statistic, having the degrees of freedom computed above. (See sample problems at the end of this lesson for examples of how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

In this section, two sample problems illustrate how to conduct a hypothesis test of a difference between mean scores. The first problem involves a two-tailed test; the second problem, a one-tailed test.

**Problem 1: Two-Tailed Test**

Within a school district, students were randomly assigned to one of two Math teachers - Mrs. Smith and Mrs. Jones. After the assignment, Mrs. Smith had 30 students, and Mrs. Jones had 25 students.

At the end of the year, each class took the same standardized test. Mrs. Smith's students had an average test score of 78, with a standard deviation of 10; and Mrs. Jones' students had an average test score of 85, with a standard deviation of 15.

Test the hypothesis that Mrs. Smith and Mrs. Jones are equally effective teachers. Use a 0.10 level of significance. (Assume that student performance is approximately normal.)

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μ1 - μ2 = 0   
Alternative hypothesis: μ1 - μ2 ≠ 0

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.10. Using sample data, we will conduct a [two-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=Two-sample%20t-test) of the null hypothesis.
* **Analyze sample data**. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

SE = sqrt[(s12/n1) + (s22/n2)]   
SE = sqrt[(102/30) + (152/25] = sqrt(3.33 + 9) = sqrt(12.33) = 3.51   
  
DF = (s12/n1 + s22/n2)2 / { [ (s12 / n1)2 / (n1 - 1) ] + [ (s22 / n2)2 / (n2 - 1) ] }   
DF = (102/30 + 152/25)2 / { [ (102 / 30)2 / (29) ] + [ (152 / 25)2 / (24) ] }   
DF = (3.33 + 9)2 / { [ (3.33)2 / (29) ] + [ (9)2 / (24) ] } = 152.03 / (0.382 + 3.375) = 152.03/3.757 = 40.47   
  
t = [ (x1 - x2) - d ] / SE = [ (78 - 85) - 0 ] / 3.51 = -7/3.51 = -1.99

where s1 is the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) of sample 1, s2 is the standard deviation of sample 2, n1 is the size of sample 1, n2 is the size of sample 2, x1 is the mean of sample 1, x2 is the mean of sample 2, d is the hypothesized difference between the population means, and SE is the standard error.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that a t statistic having 40 degrees of freedom is more extreme than -1.99; that is, less than -1.99 or greater than 1.99.

We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < -1.99) = 0.027, and P(t > 1.99) = 0.027. Thus, the P-value = 0.027 + 0.027 = 0.054.

* **Interpret results**. Since the P-value (0.054) is less than the significance level (0.10), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the samples were drawn from a normal population.

**Problem 2: One-Tailed Test**

The Acme Company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for *at least* 7 minutes longer than the old battery.

To test the claim, the company selects a simple random sample of 100 new batteries and 100 old batteries. The old batteries run continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries, 200 minutes with a standard deviation of 40 minutes.

Test the engineer's claim that the new batteries run at least 7 minutes longer than the old. Use a 0.05 level of significance. (Assume that there are no outliers in either sample.)

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μ1 - μ2 >= 7   
Alternative hypothesis: μ1 - μ2 < 7

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the mean difference between sample means is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [two-sample t-test](http://stattrek.com/Help/Glossary.aspx?Target=Two-sample%20t-test) of the null hypothesis.
* **Analyze sample data**. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

SE = sqrt[(s12/n1) + (s22/n2)]   
SE = sqrt[(402/100) + (202/100] = sqrt(16 + 4) = 4.472   
  
DF = (s12/n1 + s22/n2)2 / { [ (s12 / n1)2 / (n1 - 1) ] + [ (s22 / n2)2 / (n2 - 1) ] }   
DF = (402/100 + 202/100)2 / { [ (402 / 100)2 / (99) ] + [ (202 / 100)2 / (99) ] }   
DF = (20)2 / { [ (16)2 / (99) ] + [ (2)2 / (99) ] } = 400 / (2.586 + 0.162) = 145.56   
  
t = [ (x1 - x2) - d ] / SE = [(200 - 190) - 7] / 4.472 = 3/4.472 = 0.67

where s1 is the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) of sample 1, s2 is the standard deviation of sample 2, n1 is the size of sample 1, n2 is the size of sample 2, x1 is the mean of sample 1, x2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

Here is the logic of the analysis: Given the alternative hypothesis (μ1 - μ2 < 7), we want to know whether the observed difference in sample means is small enough (i.e., sufficiently less than 7) to cause us to reject the null hypothesis.

The observed difference in sample means (10) produced a t statistic of 0.67. We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < 0.67) = 0.75.

This means we would expect to find an observed difference in sample means of 10 or less in 75% of our samples, if the true difference were actually 7. Therefore, the P-value in this analysis is 0.75.

* **Interpret results**. Since the P-value (0.75) is greater than the significance level (0.05), we cannot reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the sample size was large without outliers.

Hypothesis Test: Difference Between Paired Means

This lesson explains how to conduct a hypothesis test for the difference between [paired means](http://stattrek.com/Help/Glossary.aspx?Target=Paired%20data). The test procedure, called the **matched-pairs t-test**, is appropriate when the following conditions are met:

* The sampling method for each sample is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The test is conducted on paired data. (As a result, the data sets are *not* [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).)
* The sampling distribution is approximately normal, which is generally true if any of the following conditions apply.
  + The population distribution is normal.
  + The population data are [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is 15 or less.
  + The population data are slightly [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is 16 to 40.
  + The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The hypotheses concern a new variable d, which is based on the difference between paired values from two data sets.

d = x1 - x2

where x1 is the value of variable x in the first data set, and x2 is the value of the variable from the second data set that is paired with x1.

The table below shows three sets of null and alternative hypotheses. Each makes a statement about how the true difference in population values μd is related to some hypothesized value D. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | μd= D | μd ≠ D | 2 |
| 2 | μd > D | μd < D | 1 |
| 3 | μd < D | μd > D | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [matched-pairs t-test](http://stattrek.com/Help/Glossary.aspx?Target=Matched-pairs%20t-test) to determine whether the difference between sample means for paired data is significantly different from the hypothesized difference between population means.

Analyze Sample Data

Using sample data, find the standard deviation, standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

* Standard deviation. Compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) (sd) of the differences computed from n matched pairs.

sd = sqrt [ (Σ(di - d)2 / (n - 1) ]

where di is the difference for pair *i*, d is the sample mean of the differences, and n is the number of paired values.

* Standard error. Compute the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=standard%20error) (SE) of the sampling distribution of d.

SE = sd \* sqrt{ ( 1/n ) \* [ (N - n) / ( N - 1 ) ] }

where *s*d is the standard deviation of the sample difference, *N* is the number of matched pairs in the population, and *n* is the number of matched pairs in the sample. When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by:

SE = sd / sqrt( n )

* Degrees of freedom. The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is: DF = n - 1 .
* Test statistic. The test statistic is a t statistic (t) defined by the following equation.

t = [ (x1 - x2) - D ] / SE = (d - D) / SE

where x1 is the mean of sample 1, x2 is the mean of sample 2, d is the mean difference between paired values in the sample, D is the hypothesized difference between population means, and SE is the standard error.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to assess the probability associated with the t statistic, having the degrees of freedom computed above. (See the sample problem at the end of this lesson for guidance on how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

Forty-four sixth graders were randomly selected from a school district. Then, they were divided into 22 matched pairs, each pair having equal IQ's. One member of each pair was randomly selected to receive special training. Then, all of the students were given an IQ test. Test results are summarized below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Pair | Training | No training | Difference, d | (d - d)2 | | 1 | 95 | 90 | 5 | 16 | | 2 | 89 | 85 | 4 | 9 | | 3 | 76 | 73 | 3 | 4 | | 4 | 92 | 90 | 2 | 1 | | 5 | 91 | 90 | 1 | 0 | | 6 | 53 | 53 | 0 | 1 | | 7 | 67 | 68 | -1 | 4 | | 8 | 88 | 90 | -2 | 9 | | 9 | 75 | 78 | -3 | 16 | | 10 | 85 | 89 | -4 | 25 | | 11 | 90 | 95 | -5 | 36 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Pair | Training | No training | Difference, d | (d - d)2 | | 12 | 85 | 83 | 2 | 1 | | 13 | 87 | 83 | 4 | 9 | | 14 | 85 | 83 | 2 | 1 | | 15 | 85 | 82 | 3 | 4 | | 16 | 68 | 65 | 3 | 4 | | 17 | 81 | 79 | 2 | 1 | | 18 | 84 | 83 | 1 | 0 | | 19 | 71 | 60 | 11 | 100 | | 20 | 46 | 47 | -1 | 4 | | 21 | 75 | 77 | -2 | 9 | | 22 | 80 | 83 | -3 | 16 | |

Σ(d - d)2 = 270   
d = 1

Do these results provide evidence that the special training helped or hurt student performance? Use an 0.05 level of significance. Assume that the mean differences are approximately normally distributed.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μd = 0   
Alternative hypothesis: μd ≠ 0

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [matched-pairs t-test](http://stattrek.com/Help/Glossary.aspx?Target=Matched-pairs%20t-test) of the null hypothesis.
* **Analyze sample data**. Using sample data, we compute the standard deviation of the differences (s), the standard error (SE) of the mean difference, the degrees of freedom (DF), and the t statistic test statistic (t).

s = sqrt [ (Σ(di - d)2 / (n - 1) ] = sqrt[ 270/(22-1) ] = sqrt(12.857) = 3.586   
  
SE = s / sqrt(n) = 3.586 / [ sqrt(22) ] = 3.586/4.69 = 0.765   
  
DF = n - 1 = 22 -1 = 21   
  
t = [ (x1 - x2) - D ] / SE = (d - D)/ SE = (1 - 0)/0.765 = 1.307

where di is the observed difference for pair *i*, d is mean difference between sample pairs, D is the hypothesized mean difference between population pairs, and n is the number of pairs.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that a t statistic having 21 degrees of freedom is more extreme than 1.307; that is, less than -1.307 or greater than 1.307.

We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < -1.307) = 0.103, and P(t > 1.307) = 0.103. Thus, the P-value = 0.103 + 0.103 = 0.206.

* **Interpret results**. Since the P-value (0.206) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples consisted of paired data, and the mean differences were normally distributed. In addition, we used the approximation formula to compute the standard error, since the sample size was small relative to the population size.

Hypothesis Test: Difference Between Paired Means

This lesson explains how to conduct a hypothesis test for the difference between [paired means](http://stattrek.com/Help/Glossary.aspx?Target=Paired%20data). The test procedure, called the **matched-pairs t-test**, is appropriate when the following conditions are met:

* The sampling method for each sample is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The test is conducted on paired data. (As a result, the data sets are *not* [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).)
* The sampling distribution is approximately normal, which is generally true if any of the following conditions apply.
  + The population distribution is normal.
  + The population data are [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry), [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution), without [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier), and the sample size is 15 or less.
  + The population data are slightly [skewed](http://stattrek.com/Help/Glossary.aspx?Target=Skewness), unimodal, without outliers, and the sample size is 16 to 40.
  + The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The hypotheses concern a new variable d, which is based on the difference between paired values from two data sets.

d = x1 - x2

where x1 is the value of variable x in the first data set, and x2 is the value of the variable from the second data set that is paired with x1.

The table below shows three sets of null and alternative hypotheses. Each makes a statement about how the true difference in population values μd is related to some hypothesized value D. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | μd= D | μd ≠ D | 2 |
| 2 | μd > D | μd < D | 1 |
| 3 | μd < D | μd > D | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [matched-pairs t-test](http://stattrek.com/Help/Glossary.aspx?Target=Matched-pairs%20t-test) to determine whether the difference between sample means for paired data is significantly different from the hypothesized difference between population means.

Analyze Sample Data

Using sample data, find the standard deviation, standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

* Standard deviation. Compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) (sd) of the differences computed from n matched pairs.

sd = sqrt [ (Σ(di - d)2 / (n - 1) ]

where di is the difference for pair *i*, d is the sample mean of the differences, and n is the number of paired values.

* Standard error. Compute the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=standard%20error) (SE) of the sampling distribution of d.

SE = sd \* sqrt{ ( 1/n ) \* [ (N - n) / ( N - 1 ) ] }

where *s*d is the standard deviation of the sample difference, *N* is the number of matched pairs in the population, and *n* is the number of matched pairs in the sample. When the population size is much larger (at least 20 times larger) than the sample size, the standard error can be approximated by:

SE = sd / sqrt( n )

* Degrees of freedom. The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is: DF = n - 1 .
* Test statistic. The test statistic is a t statistic (t) defined by the following equation.

t = [ (x1 - x2) - D ] / SE = (d - D) / SE

where x1 is the mean of sample 1, x2 is the mean of sample 2, d is the mean difference between paired values in the sample, D is the hypothesized difference between population means, and SE is the standard error.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to assess the probability associated with the t statistic, having the degrees of freedom computed above. (See the sample problem at the end of this lesson for guidance on how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

Forty-four sixth graders were randomly selected from a school district. Then, they were divided into 22 matched pairs, each pair having equal IQ's. One member of each pair was randomly selected to receive special training. Then, all of the students were given an IQ test. Test results are summarized below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Pair | Training | No training | Difference, d | (d - d)2 | | 1 | 95 | 90 | 5 | 16 | | 2 | 89 | 85 | 4 | 9 | | 3 | 76 | 73 | 3 | 4 | | 4 | 92 | 90 | 2 | 1 | | 5 | 91 | 90 | 1 | 0 | | 6 | 53 | 53 | 0 | 1 | | 7 | 67 | 68 | -1 | 4 | | 8 | 88 | 90 | -2 | 9 | | 9 | 75 | 78 | -3 | 16 | | 10 | 85 | 89 | -4 | 25 | | 11 | 90 | 95 | -5 | 36 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Pair | Training | No training | Difference, d | (d - d)2 | | 12 | 85 | 83 | 2 | 1 | | 13 | 87 | 83 | 4 | 9 | | 14 | 85 | 83 | 2 | 1 | | 15 | 85 | 82 | 3 | 4 | | 16 | 68 | 65 | 3 | 4 | | 17 | 81 | 79 | 2 | 1 | | 18 | 84 | 83 | 1 | 0 | | 19 | 71 | 60 | 11 | 100 | | 20 | 46 | 47 | -1 | 4 | | 21 | 75 | 77 | -2 | 9 | | 22 | 80 | 83 | -3 | 16 | |

Σ(d - d)2 = 270   
d = 1

Do these results provide evidence that the special training helped or hurt student performance? Use an 0.05 level of significance. Assume that the mean differences are approximately normally distributed.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μd = 0   
Alternative hypothesis: μd ≠ 0

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [matched-pairs t-test](http://stattrek.com/Help/Glossary.aspx?Target=Matched-pairs%20t-test) of the null hypothesis.
* **Analyze sample data**. Using sample data, we compute the standard deviation of the differences (s), the standard error (SE) of the mean difference, the degrees of freedom (DF), and the t statistic test statistic (t).

s = sqrt [ (Σ(di - d)2 / (n - 1) ] = sqrt[ 270/(22-1) ] = sqrt(12.857) = 3.586   
  
SE = s / sqrt(n) = 3.586 / [ sqrt(22) ] = 3.586/4.69 = 0.765   
  
DF = n - 1 = 22 -1 = 21   
  
t = [ (x1 - x2) - D ] / SE = (d - D)/ SE = (1 - 0)/0.765 = 1.307

where di is the observed difference for pair *i*, d is mean difference between sample pairs, D is the hypothesized mean difference between population pairs, and n is the number of pairs.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that a t statistic having 21 degrees of freedom is more extreme than 1.307; that is, less than -1.307 or greater than 1.307.

We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t < -1.307) = 0.103, and P(t > 1.307) = 0.103. Thus, the P-value = 0.103 + 0.103 = 0.206.

* **Interpret results**. Since the P-value (0.206) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples consisted of paired data, and the mean differences were normally distributed. In addition, we used the approximation formula to compute the standard error, since the sample size was small relative to the population size.

Hypothesis Test for a Proportion

This lesson explains how to conduct a hypothesis test of a proportion, when the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* Each sample point can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
* The sample includes at least 10 successes and 10 failures.
* The population size is at least 20 times as big as the sample size.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [one-sample z-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20z-test) to determine whether the hypothesized population proportion differs significantly from the observed sample proportion.

Analyze Sample Data

Using sample data, find the test statistic and its associated P-Value.

* Standard deviation. Compute the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=standard%20deviation) (σ) of the sampling distribution.

σ = sqrt[ P \* ( 1 - P ) / n ]

where P is the hypothesized value of population proportion in the null hypothesis, and n is the sample size.

* Test statistic. The test statistic is a z-score (z) defined by the following equation.

z = (p - P) / σ

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and σ is the standard deviation of the sampling distribution.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to assess the probability associated with the z-score. (See sample problems at the end of this lesson for examples of how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

In this section, two hypothesis testing examples illustrate how to conduct a hypothesis test of a proportion. The first problem involves a a two-tailed test; the second problem, a one-tailed test.

Sample Planning Wizard

As you probably noticed, the process of testing a hypothesis about a proportion can be complex and time-consuming. Stat Trek's Sample Planning Wizard can do the same job quickly, easily, and error-free. In addition to conducting the hypothesis test, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to test a hypothesis, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Problem 1: Two-Tailed Test**

The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are very satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Among the sampled customers, 73 percent say they are very satisified. Based on these findings, can we reject the CEO's hypothesis that 80% of the customers are very satisfied? Use a 0.05 level of significance.

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P = 0.80   
Alternative hypothesis: P ≠ 0.80

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample proportion is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method, shown in the next section, is a [one-sample z-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20z-test).
* **Analyze sample data**. Using sample data, we calculate the standard deviation (σ) and compute the z-score test statistic (z).

σ = sqrt[ P \* ( 1 - P ) / n ] = sqrt [(0.8 \* 0.2) / 100] = sqrt(0.0016) = 0.04   
z = (p - P) / σ = (.73 - .80)/0.04 = -1.75

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and n is the sample size.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that the z-score is less than -1.75 or greater than 1.75.

We use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to find P(z < -1.75) = 0.04, and P(z > 1.75) = 0.04. Thus, the P-value = 0.04 + 0.04 = 0.08.

* **Interpret results**. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the sample included at least 10 successes and 10 failures, and the population size was at least 10 times the sample size.

**Problem 2: One-Tailed Test**  
  
Suppose the previous example is stated a little bit differently. Suppose the CEO claims that *at least* 80 percent of the company's 1,000,000 customers are very satisfied. Again, 100 customers are surveyed using simple random sampling. The result: 73 percent are very satisfied. Based on these results, should we accept or reject the CEO's hypothesis? Assume a significance level of 0.05.

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P >= 0.80   
Alternative hypothesis: P < 0.80

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected only if the sample proportion is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method, shown in the next section, is a [one-sample z-test](http://stattrek.com/Help/Glossary.aspx?Target=One-sample%20z-test).
* **Analyze sample data**. Using sample data, we calculate the standard deviation (σ) and compute the z-score test statistic (z).

σ = sqrt[ P \* ( 1 - P ) / n ] = sqrt [(0.8 \* 0.2) / 100] = sqrt(0.0016) = 0.04   
z = (p - P) / σ = (.73 - .80)/0.04 = -1.75

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and n is the sample size.

Since we have a [one-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=One-tailed%20test), the P-value is the probability that the z-score is less than -1.75. We use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to find P(z < -1.75) = 0.04. Thus, the P-value = 0.04.

* **Interpret results**. Since the P-value (0.04) is less than the significance level (0.05), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the sample included at least 10 successes and 10 failures, and the population size was at least 10 times the sample size.

Hypothesis Test of a Proportion (Small Sample)

This lesson explains how to test a hypothesis about a proportion when a simple random sample has fewer than 10 successes or 10 failures - a situation that often occurs with small samples. (In the [previous lesson](http://stattrek.com/Lesson5/Proportion.aspx?Tutorial=stat), we showed how to conduct a hypothesis test for a proportion when a simple random sample includes at least 10 successes and 10 failures.)

The approach described in this lesson is appropriate as long as the sample includes at least one success and one failure. The key steps are:

* Formulate the hypotheses to be tested. This means stating the [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null_hypothesis) and the [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative_hypothesis).
* Determine the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the proportion. If the sample proportion is the outcome of a [binomial experiment](http://stattrek.com/Help/Glossary.aspx?Target=Binomial_experiment), the sampling distribution will be binomial. If it is the outcome of a [hypergeometric experiment](http://stattrek.com/Help/Glossary.aspx?Target=Hypergeometric_experiment), the sampling distribution will be hypergeometric.
* Specify the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance_level). (Researchers often set the significance level equal to 0.05 or 0.01, although other values may be used.)
* Based on the hypotheses, the sampling distribution, and the significance level, define the [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_acceptance).
* Test the null hypothesis. If the sample proportion falls within the region of acceptance, accept the null hypothesis; otherwise, reject the null hypothesis.

The following hypothesis testing examples illustrate how this works. The first example involves a binomial experiment; and the second example, a hypergeometric experiment.

Sample Planning Wizard

The steps required to conduct a hypothesis test using a small sample are not trivial. They can be time-consuming and complex. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and accurately. In addition to conducting the hypothesis test, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to test a hypothesis, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

Example 1: Sampling With Replacement

Suppose an urn contains 30 marbles. Some marbles are red, and the rest are green. A researcher hypothesizes that the urn contains 15 or more red marbles. The researcher randomly samples five marbles, [with replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_with_replacement), from the urn. Two of the selected marbles are red, and three are green. Based on the sample results, should the researcher accept or reject the hypothesis. Use a significance level of 0.20.

*Solution:* There are five steps in conducting a hypothesis test, as described in the previous section. We work through each of the five steps below:

* **Formulate hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P >= 0.50   
Alternative hypothesis: P < 0.50

Note that these hypotheses constitute a [one-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test). The null hypothesis will be rejected only if the sample proportion is too small.

* **Determine sampling distribution**. Since we sampled with replacement, the sample proportion can be considered an outcome of a binomial experiment. And based on the null hypothesis, we assume that at least 15 of 30 marbles are red. Thus, the true population proportion is assumed to be 15/30 or 0.50.   
    
  Given those inputs (a binomial distribution where the true population proportion is equal to 0.50), the sampling distribution of the proportion can be determined. It appears in the table below. (Previously, we showed [how to compute binomial probabilities](http://stattrek.com/Lesson2/Binomial.aspx?Tutorial=stat) that form the body of the table.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of red marbles in sample** | **Sample proportion** | **Binomial probability** | **Cumulative probability** |
| 0 | 0.0 | 0.03125 | 0.03125 |
| 1 | 0.2 | 0.15625 | 0.1875 |
| 2 | 0.4 | 0.3125 | 0.5 |
| 3 | 0.6 | 0.3125 | 0.8125 |
| 4 | 0.8 | 0.15625 | 0.96875 |
| 5 | 1.0 | 0.03125 | 1.00 |

* **Specify significance level**. The significance level was set at 0.20. (This means that the probability of making a [Type I error](http://stattrek.com/Help/Glossary.aspx?Target=Type_I_error) is 0.20, assuming that the null hypothesis is true.)
* **Define the region of acceptance**. From the sampling distribution (see above table), we see that it is not possible to define a region of acceptance for which the significance level is *exactly* 0.20.   
    
  However, we can define a region of acceptance for which the significance level would be *no more than* 0.20. From the table, we see that if the true population proportion is equal to 0.50, we would be very unlikely to pick 0 or 1 red marble in our sample of 5 marbles. The probability of selecting 1 or 0 red marbles would be 0.1875. Therefore, if we let the significance level equal 0.1875, we can define the [region of rejection](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_rejection) as any sampled outcome that includes only 0 or 1 red marble (i.e., a sampled proportion equal to 0 or 0.20). We can define the region of acceptance as any sampled outcome that includes at least 2 red marbles. This is equivalent to a sampled proportion that is greater than or equal to 0.40.
* **Test the null hypothesis**. Since the sample proportion (0.40) is within the region of acceptance, we cannot reject the null hypothesis.

Example 2: Sampling Without Replacement

The Acme Advertising company has 25 clients. Account executives at Acme claim that 80 percent of these clients are very satisfied with the service they receive. To test that claim, Acme's CEO commissions a survey of 10 clients. Survey participants are randomly sampled, [without replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_without_replacement), from the client population. Six of the ten sampled customers (i.e., 60 percent) say that they are very satisfied. Based on the sample results, should the CEO accept or reject the hypothesis that 80 percent of Acme's clients are very satisfied. Use a significance level of 0.10.

*Solution:* There are five steps in conducting a hypothesis test, as described in the previous section. We work through each of the five steps below:

* **Formulate hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P >= 0.80   
Alternative hypothesis: P < 0.80

Note that these hypotheses constitute a [one-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test). The null hypothesis will be rejected only if the sample proportion is too small.

* **Determine sampling distribution**. Since we sampled without replacement, the sample proportion can be considered an outcome of a hypergeometric experiment. And based on the null hypothesis, we assume that at least 80 percent of the 25 clients (i.e. 20 clients) are very satisfied.   
    
  Given those inputs (a hypergeometric distribution where 20 of 25 clients are very satisfied), the sampling distribution of the proportion can be determined. It appears in the table below. (Previously, we showed [how to compute hypergeometric probabilities](http://stattrek.com/Lesson2/Hypergeometric.aspx?Tutorial=stat) that form the body of the table.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of satisfied clients in sample** | **Sample proportion** | **Hypergeometric probability** | **Cumulative probability** |
| 4 or less | 0.4 or less | 0.00 | 0.00 |
| 5 | 0.5 | 0.00474 | 0.00474 |
| 6 | 0.6 | 0.05929 | 0.06403 |
| 7 | 0.7 | 0.23715 | 0.30119 |
| 8 | 0.8 | 0.38538 | 0.68656 |
| 9 | 0.9 | 0.25692 | 0.94348 |
| 10 | 1.0 | 0.05652 | 1.00 |

* **Specify significance level**. The significance level was set at 0.10. (This means that the probability of making a [Type I error](http://stattrek.com/Help/Glossary.aspx?Target=Type_I_error) is 0.10, assuming that the null hypothesis is true.)
* **Define the region of acceptance**. From the sampling distribution (see above table), we see that it is not possible to define a region of acceptance for which the significance level is *exactly* 0.10.   
    
  However, we can define a region of acceptance for which the significance level would be *no more than* 0.10. From the table, we see that if the true proportion of very satisfied clients is equal to 0.80, we would be very unlikely to have fewer than 7 very satisfied clients in our sample. The probability of having 6 or fewer very satisfied clients in the sample would be 0.064. Therefore, if we let the significance level equal 0.064, we can define the [region of rejection](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_rejection) as any sampled outcome that includes 6 or fewer very satisfied customers. We can define the region of acceptance as any sampled outcome that includes 7 or more very satisfied customers. This is equivalent to a sample proportion that is greater than or equal to 0.70.
* **Test the null hypothesis**. Since the sample proportion (0.60) is outside the region of acceptance, we cannot accept the null hypothesis at the 0.064 level of significance.

Hypothesis Test: Difference Between Proportions

This lesson explains how to conduct a hypothesis test to determine whether the difference between two proportions is significant. The test procedure, called the **two-proportion z-test**, is appropriate when the following conditions are met:

* The sampling method for each population is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The samples are [independent](http://stattrek.com/Help/Glossary.aspx?Target=Independent).
* Each sample includes at least 10 successes and 10 failures.
* Each population is at least 20 times as big as its sample.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The table below shows three sets of hypotheses. Each makes a statement about the difference *d* between two population proportions, P1 and P2. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | P1 - P2 = 0 | P1 - P2 ≠ 0 | 2 |
| 2 | P1 - P2 > 0 | P1 - P2 < 0 | 1 |
| 3 | P1 - P2 < 0 | P1 - P2 > 0 | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population proportions (i.e., d = 0), the null and alternative hypothesis for a two-tailed test are often stated in the following form.

H0: P1 = P2  
Ha: P1 ≠ P2

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the two-proportion z-test (described in the next section) to determine whether the hypothesized difference between population proportions differs significantly from the observed sample difference.

Analyze Sample Data

Using sample data, complete the following computations to find the test statistic and its associated P-Value.

* **Pooled sample proportion.** Since the null hypothesis states that P1=P2, we use a pooled sample proportion (p) to compute the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=standard%20error) of the sampling distribution.

p = (p1 \* n1 + p2 \* n2) / (n1 + n2)

where p1 is the sample proportion from population 1, p2 is the sample proportion from population 2, n1 is the size of sample 1, and n2 is the size of sample 2.

* **Standard error.** Compute the standard error (SE) of the sampling distribution difference between two proportions.

SE = sqrt{ p \* ( 1 - p ) \* [ (1/n1) + (1/n2) ] }

where p is the pooled sample proportion, n1 is the size of sample 1, and n2 is the size of sample 2.

* **Test statistic.** The test statistic is a z-score (z) defined by the following equation.

z = (p1 - p2) / SE

where p1 is the proportion from sample 1, p2 is the proportion from sample 2, and SE is the standard error of the sampling distribution.

* **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to assess the probability associated with the z-score. (See sample problems at the end of this lesson for examples of how this is done.)

The analysis described above is a two-proportion z-test.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

In this section, two sample problems illustrate how to conduct a hypothesis test for the difference between two proportions. The first problem involves a a two-tailed test; the second problem, a one-tailed test.

**Problem 1: Two-Tailed Test**

Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P1 = P2  
Alternative hypothesis: P1 ≠ P2

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the proportion from population 1 is too big or if it is too small.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. The test method is a two-proportion z-test.
* **Analyze sample data**. Using sample data, we calculate the pooled sample proportion (p) and the standard error (SE). Using those measures, we compute the z-score test statistic (z).

p = (p1 \* n1 + p2 \* n2) / (n1 + n2) = [(0.38 \* 100) + (0.51 \* 200)] / (100 + 200) = 140/300 = 0.467   
  
SE = sqrt{ p \* ( 1 - p ) \* [ (1/n1) + (1/n2) ] }   
SE = sqrt [ 0.467 \* 0.533 \* ( 1/100 + 1/200 ) ] = sqrt [0.003733] = 0.061   
  
z = (p1 - p2) / SE = (0.38 - 0.51)/0.061 = -2.13

where p1 is the sample proportion in sample 1, where p2 is the sample proportion in sample 2, n1 is the size of sample 1, and n2 is the size of sample 2.

Since we have a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that the z-score is less than -2.13 or greater than 2.13.

We use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to find P(z < -2.13) = 0.017, and P(z > 2.13) = 0.017. Thus, the P-value = 0.017 + 0.017 = 0.034.

* **Interpret results**. Since the P-value (0.034) is less than the significance level (0.05), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, each population was at least 10 times larger than its sample, and each sample included at least 10 successes and 10 failures.

**Problem 2: One-Tailed Test**

Suppose the previous example is stated a little bit differently. Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is more effective for women than for men. To test this claim, they choose a a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we conclude that the drug is more effective for women than for men? Use a 0.01 level of significance.

*Solution:* The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P1 >= P2  
Alternative hypothesis: P1 < P2

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the proportion of women catching cold (p1) is sufficiently smaller than the proportion of men catching cold (p2).

* **Formulate an analysis plan**. For this analysis, the significance level is 0.01. The test method is a two-proportion z-test.
* **Analyze sample data**. Using sample data, we calculate the pooled sample proportion (p) and the standard error (SE). Using those measures, we compute the z-score test statistic (z).

p = (p1 \* n1 + p2 \* n2) / (n1 + n2) = [(0.38 \* 100) + (0.51 \* 200)] / (100 + 200) = 140/300 = 0.467   
  
SE = sqrt{ p \* ( 1 - p ) \* [ (1/n1) + (1/n2) ] }   
SE = sqrt [ 0.467 \* 0.533 \* ( 1/100 + 1/200 ) ] = sqrt [0.003733] = 0.061   
  
z = (p1 - p2) / SE = (0.38 - 0.51)/0.061 = -2.13

where p1 is the sample proportion in sample 1, where p2 is the sample proportion in sample 2, n1 is the size of sample 1, and n2 is the size of sample 2.

Since we have a [one-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), the P-value is the probability that the z-score is less than -2.13. We use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx) to find P(z < -2.13) = 0.017. Thus, the P-value = 0.017.

* **Interpret results**. Since the P-value (0.017) is greater than the significance level (0.01), we cannot reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, each population was at least 10 times larger than its sample, and each sample included at least 10 successes and 10 failures.

Region of Acceptance

In this lesson, we describe how to find the [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region%20of%20acceptance) for a hypothesis test.

One-Tailed and Two-Tailed Hypothesis Tests

The steps taken to define the region of acceptance will vary, depending on whether the [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null_hypothesis) and the [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative_hypothesis) call for one- or two-tailed hypothesis tests. So we begin with a brief review.

The table below shows three sets of hypotheses. Each makes a statement about how the population mean μ is related to a specified value *M*. (In the table, the symbol ≠ means " not equal to ".)

|  |  |  |  |
| --- | --- | --- | --- |
| **Set** | **Null hypothesis** | **Alternative hypothesis** | **Number of tails** |
| 1 | μ = M | μ ≠ M | 2 |
| 2 | μ > M | μ < M | 1 |
| 3 | μ < M | μ > M | 1 |

The first set of hypotheses (Set 1) is an example of a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), since an extreme value on either side of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are [one-tailed tests](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

How to Find the Region of Acceptance

We define the region of acceptance in such a way that the chance of making a [Type I error](http://stattrek.com/Help/Glossary.aspx?Target=Type_I_error) is equal to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance_level). Here is how that is done.

* Define a test statistic. Here, the test statistic is the sample measure used to estimate the population parameter that appears in the null hypothesis. For example, suppose the null hypothesis is

H0: μ = M

The test statistic, used to estimate *M*, would be *m*. If *M* were a population mean, *m* would be the sample mean; if *M* were a population proportion, *m* would be the sample proportion; if *M* were a difference between population means, *m* would be the difference between sample means; and so on.

* Given the significance level α , find the upper limit (UL) of the region of acceptance. There are three possibilities, depending on the form of the null hypothesis.
  + If the null hypothesis is μ < M: The upper limit of the region of acceptance will be equal to the value for which the cumulative probability of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) is equal to one minus the significance level. That is, P( m < UL ) = 1 - α .
  + If the null hypothesis is μ = M: The upper limit of the region of acceptance will be equal to the value for which the cumulative probability of the sampling distribution is equal to one minus the significance level divided by 2. That is, P( m < UL ) = 1 - α/2 .
  + If the null hypothesis is μ > M: The upper limit of the region of acceptance is equal to plus infinity, unless the test statistic were a proportion or a percentage. The upper limit is 1 for a proportion, and 100 for a percentage.
* In a similar way, we find the lower limit (LL) of the range of acceptance. Again, there are three possibilities, depending on the form of the null hypothesis.
  + If the null hypothesis is μ < M: The lower limit of the region of acceptance is equal to minus infinity, unless the test statistic is a proportion or a percentage. The lower limit for a proportion or a percentage is zero.
  + If the null hypothesis is μ = M: The lower limit of the region of acceptance will be equal to the value for which the cumulative probability of the sampling distribution is equal to the significance level divided by 2. That is, P( m < LL ) = α/2 .
  + If the null hypothesis is μ > M: The lower limit of the region of acceptance will be equal to the value for which the cumulative probability of the sampling distribution is equal to the significance level. That is, P( m < LL ) = α .

The region of acceptance is defined by the range between LL and UL.

Test Your Understanding

In this section, two hypothesis testing examples illustrate how to define the region of acceptance. The first problem shows a two-tailed test with a mean score; and the second problem, a one-tailed test with a proportion.

Sample Planning Wizard

As you probably noticed, defining the region of acceptance can be complex and time-consuming. Stat Trek's Sample Planning Wizard can do the same job quickly, easily, and error-free. In addition, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need find the region of acceptance, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Problem 1**

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. Suppose a random sample of 50 engines is tested. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes.

Consider the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. Find the region of acceptance. Based on the region of acceptance, would you reject the null hypothesis?

*Solution:* The process of defining a region of acceptance to test a hypothesis takes four steps. We work through those steps below:

* **Formulate hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: μ = 300 minutes  
Alternative hypothesis: μ ≠ 300 minutes

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample mean is too big or if it is too small.

* **Identify the test statistic**. In this example, the test statistic is the mean run time of the 50 engines in the sample - 295 minutes.
* **Define the region of acceptance**. To define the region of acceptance, we need to understand the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the test statistic. And we need to derive some probabilities. Those points are covered below.
  + **Specify the sampling distribution**. Since the sample size is large (greater than or equal to 40), we assume that the sampling distribution of the mean is normal, based on the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem).
  + **Define the mean of the sampling distribution**. We assume that the mean of the sampling distribution is equal to the mean value that appears in the null hypothesis - 300 minutes.
  + **Compute the standard error of the sample mean**. Here the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the sample mean (sx) is:

sx = s \* sqrt( 1/n )  
sx = 20 \* sqrt[ 1/50 ] = 2.83

where *s* is the sample standard deviation, *n* is the sample size, and *N* is the population size. (In this example, we assume that the population size is very large, so the [finite population correction](http://stattrek.com/statistics/dictionary.aspx?Definition=finite_population_correction) is equal to about one.)

* + **Find the lower limit of the region of acceptance.** Given a two-tailed hypothesis, the lower limit (LL) will be equal to the value for which the cumulative probability of the sampling distribution is equal to the significance level divided by 2. That is, P( x < LL ) = α/2 = 0.05/2 = 0.025. To find this lower limit, we use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx). We input the following entries into the calculator: cumulative probability = 0.025, mean = 300, and standard deviation = 2.83. The calculator tells us that the lower limit is 294.45, given those inputs.
  + **Find the upper limit of the region of acceptance.** Given a two-tailed hypothesis, the upper limit (UL) will be equal to the value for which the cumulative probability of the sampling distribution is equal to one minus the significance level divided by 2. That is, P( x< UL ) = 1 - α/2 = 1 - 0.025 = 0.975. To find this upper limit, we use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx). We input the following entries into the calculator: cumulative probability = 0.975, mean = 300, and standard deviation = 2.83. The calculator tells us that the upper limit is 305.55, given those inputs.

Thus, we have determined that the region of acceptance is defined by the values between 294.45 and 305.55.

* **Accept or reject the null hypothesis**. The sample mean in this example was 295 minutes. This value falls within the region of acceptance. Therefore, we cannot reject the null hypothesis that a new engine runs for 300 minutes on a gallon of gasoline.

**Problem 2**  
  
Suppose the CEO of a large software company claims that *at least* 80 percent of the company's 1,000,000 customers are very satisfied. A survey of 100 randomly sampled customers finds that 73 percent are very satisfied. To test the CEO's hypothesis, find the region of acceptance. Assume a significance level of 0.05.

*Solution:* The process of defining a region of acceptance to test a hypothesis takes four steps. We work through those steps below:

* **Formulate hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: P > 0.80   
Alternative hypothesis: P < 0.80

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the sample proportion is too small.

* **Identify the test statistic**. In this example, the test statistic is the proportion of sampled customers who say they are very satisfied; i.e., 0.73.
* **Define the region of acceptance**. To define the region of acceptance, we need to understand the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the test statistic. And we need to derive some probabilities. Those points are covered below.
  + **Specify the sampling distribution**. Since the sample size is large (greater than or equal to 40), we assume that the sampling distribution of the proportion is normal, based on the central limit theorem.
  + **Define the mean of the sampling distribution**. We assume that the mean of the sampling distribution is equal to the hypothesized population proportion, which appears in the null hypothesis - 0.80.
  + **Compute the standard deviation of the sampling distribution** . Here the standard deviation of the sampling distribution sp is:

sp = sqrt[ P' \* ( 1 - P' ) / n ] \* ( N - n ) / ( N - 1 ) ]   
sp = sqrt[ 0.8 \* 0.2 / 100 ] \* ( 999,900) / (999,999 ) ] = sqrt(0.0016) \* 0.9999 = 0.04

where P' is the test value specified in the null hypothesis, n is the sample size, and N is the population size.

* + **Find the lower limit of the region of acceptance.** Given a one-tailed hypothesis, the lower limit (LL) will be equal to the value for which the cumulative probability of the sampling distribution is equal to the significance level. That is, P( x < LL ) = α = 0.05. To find this lower limit, we use the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx). We input the following entries into the calculator: cumulative probability = 0.05, and mean = 0.80. The calculator tells us that the lower limit is 0.734, given those inputs.
  + **Find the upper limit of the region of acceptance.** Since we have a one-tailed hypothesis in which the null hypothesis states that the satisfaction level is 0.80 or more, any proportion greater than 0.80 is consistent with the null hypothesis. Therefore, the upper limit is 1.0 (since the highest possible proportion is 1.0).

Thus, we have determined that the region of acceptance is defined by the values between 0.734 and 1.00.

* **Accept or reject the null hypothesis**. The sample proportion in this example was a satisfaction level of 0.73. This value falls outside the region of acceptance. Therefore, we reject the null hypothesis that 80 percent of the utility's customers are very satisfied.

# Power of a Hypothesis Test

The probability of *not* committing a [Type II error](http://stattrek.com/Help/Glossary.aspx?Target=Type%20II%20error) is called the **power** of a hypothesis test.

## Effect Size

To compute the power of the test, one offers an alternative view about the "true" value of the population parameter, assuming that the null hypothesis is false. The **effect size** is the difference between the true value and the value specified in the null hypothesis.

Effect size = True value - Hypothesized value

For example, suppose the null hypothesis states that a population mean is equal to 100. A researcher might ask: What is the probability of rejecting the null hypothesis if the true population mean is equal to 90? In this example, the effect size would be 90 - 100, which equals -10.

## Factors That Affect Power

The power of a hypothesis test is affected by three factors.

* Sample size (n). Other things being equal, the greater the sample size, the greater the power of the test.
* Significance level (α). The higher the significance level, the higher the power of the test. If you increase the significance level, you reduce the [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region%20of%20acceptance). As a result, you are more likely to reject the null hypothesis. This means you are less likely to accept the null hypothesis when it is false; i.e., less likely to make a Type II error. Hence, the power of the test is increased.
* The "true" value of the parameter being tested. The greater the difference between the "true" value of a parameter and the value specified in the null hypothesis, the greater the power of the test. That is, the greater the effect size, the greater the power of the test.

## Test Your Understanding

**Problem 1**

Other things being equal, which of the following actions will reduce the power of a hypothesis test?

I. Increasing sample size.   
II. Increasing significance level.   
III. Increasing beta, the probability of a Type II error.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above   
(E) None of the above

**Solution**

The correct answer is (C). Increasing sample size makes the hypothesis test more sensitive - more likely to reject the null hypothesis when it is, in fact, false. Increasing the significance level reduces the [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region%20of%20acceptance), which makes the hypothesis test more likely to reject the null hypothesis, thus increasing the power of the test. Since, by definition, power is equal to one minus beta, the power of a test will get smaller as beta gets bigger.

**Problem 2**

Suppose a researcher conducts an experiment to test a hypothesis. If she doubles her sample size, which of the following will increase?

I. The power of the hypothesis test.   
II. The effect size of the hypothesis test.   
III. The probability of making a Type II error.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above   
(E) None of the above

**Solution**

The correct answer is (A). Increasing sample size makes the hypothesis test more sensitive - more likely to reject the null hypothesis when it is, in fact, false. Thus, it increases the power of the test. The effect size is not affected by sample size. And the probability of making a [Type II error](http://stattrek.com/Help/Glossary.aspx?Target=Type%20II%20error) gets smaller, not bigger, as sample size increases.

How to Find the Power of a Statistical Test

When a researcher designs a study to test a hypothesis, he/she should compute the [power](http://stattrek.com/Help/Glossary.aspx?Target=Power) of the test (i.e., the likelihood of avoiding a Type II error).

How to Compute the Power of a Hypothesis Test

To compute the power of a hypothesis test, use the following three-step procedure.

* Define the [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_acceptance). Previously, we showed [how to compute the region of acceptance](http://stattrek.com/Lesson5/Region-of-Acceptance.aspx) for a hypothesis test.
* Specify the critical parameter value. The **critical parameter value** is an alternative to the value specified in the null hypothesis. The difference between the critical parameter value and the value from the null hypothesis is called the **effect size**. That is, the effect size is equal to the critical parameter value minus the value from the null hypothesis.
* Compute power. Assume that the true population parameter is equal to the critical parameter value, rather than the value specified in the null hypothesis. Based on that assumption, compute the probability that the sample estimate of the population parameter will fall outside the region of acceptance. That probability is the power of the test.

The following examples illustrate how this works. The first example involves a mean score; and the second example, a proportion.

Sample Planning Wizard

The steps required to compute the power of a hypothesis test are not trivial. They can be time-consuming and complex. Stat Trek's Sample Planning Wizard does this work for you - quickly, easily, and error-free. In addition to constructing a confidence interval, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to construct a confidence interval, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

Example 1: Power of the Hypothesis Test of a Mean Score

Two inventors have developed a new, energy-efficient lawn mower engine. One inventor says that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. Suppose a random sample of 50 engines is tested. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. The inventor tests the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes, using a 0.05 level of significance.

The other inventor says that the new engine will run continuously for only 290 minutes on a gallon of gasoline. Find the power of the test to reject the null hypothesis, if the second inventor is correct.

*Solution:* The steps required to compute power are presented below.

* **Define the region of acceptance**. In a previous lesson, we showed that the region of acceptance for this problem consists of the values between 294.45 and 305.55 (see [previous lesson](http://stattrek.com/Lesson5/Region-of-Acceptance.aspx#example1)).
* **Specify the critical parameter value**. The null hypothesis tests the hypothesis that the run time of the engine is 300 minutes. We are interested in determining the probability that the hypothesis test will reject the null hypothesis, if the true run time is actually 290 minutes. Therefore, the critical parameter value is 290. (Another way to express the critical parameter value is through effect size. The effect size is equal to the critical parameter value minus the hypothesized value. Thus, effect size is equal to 290 - 300 or -10.)
* **Compute power**. The power of the test is the probability of rejecting the null hypothesis, assuming that the true population mean is equal to the critical parameter value. Since the region of acceptance is 294.45 to 305.55, the null hypothesis will be rejected when the sampled run time is less than 294.45 or greater than 305.55.   
    
  Therefore, we need to compute the probability that the sampled run time will be less than 294.45 or greater than 305.55. To do this, we make the following assumptions:
  + The sampling distribution of the mean is normally distributed. (Because the sample size is relatively large, this assumption can be justified by the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem).)
  + The mean of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) is the critical parameter value, 290.
  + The [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the sampling distribution is 2.83. The standard error of the sampling distribution was computed in a previous lesson (see [previous lesson](http://stattrek.com/Lesson5/Region-of-Acceptance.aspx#example1)).

Given these assumptions, we first assess the probability that the sample run time will be less than 294.45. This is easy to do, using the [Normal Calculator](http://stattrek.com/Tables/Normal.aspx). We enter the following values into the calculator: value = 294.45; mean = 290; and standard deviation = 2.83. Given these inputs, we find that the cumulative probability is 0.94207551. This means the probability that the sample mean will be less than 294.45 is 0.942.   
  
Next, we assess the probability that the sample mean is greater than 305.55. Again, we use the [Normal Calculator](http://stattrek.com/Tables/Normal.aspx). We enter the following values into the calculator: normal random variable = 305.55; mean = 290; and standard deviation = 2.83. Given these inputs, we find that the probability that the sample mean is less than 305.55 (i.e., the cumulative probability) is 0.99999998. Thus, the probability that the sample mean is greater than 305.55 is 1 - 0.99999998 or 0.00000002.   
  
The power of the test is the sum of these probabilities: 0.94207551 + 0.00000002 = 0.94207553. This means that if the true average run time of the new engine were 290 minutes, we would correctly reject the hypothesis that the run time was 300 minutes 94.2 percent of the time. Hence, the probability of a Type II error would be very small. Specifically, it would be 1 minus 0.942 or 0.058.

Example 2: Power of the Hypothesis Test of a Proportion

A major corporation offers a large bonus to all of its employees if at least 80 percent of the corporation's 1,000,000 customers are very satisfied. The company conducts a survey of 100 randomly sampled customers to determine whether or not to pay the bonus. The null hypothesis states that the proportion of very satisfied customers is at least 0.80. If the null hypothesis cannot be rejected, given a significance level of 0.05, the company pays the bonus.

Suppose the true proportion of satisfied customers is 0.75. Find the power of the test to reject the null hypothesis.

*Solution:* The steps required to compute power are presented below.

* **Define the region of acceptance**. In a previous lesson, we showed that the region of acceptance for this problem consists of the values between 0.734 and 1.00. (see [previous lesson](http://stattrek.com/Lesson5/Region-of-Acceptance.aspx#example2)).
* **Specify the critical parameter value**. The null hypothesis tests the hypothesis that the proportion of very satisfied customers is 0.80. We are interested in determining the probability that the hypothesis test will reject the null hypothesis, if the true satisfaction level is 0.75. Therefore, the critical parameter value is 0.75. (Another way to express the critical parameter value is through effect size. The effect size is equal to the critical parameter value minus the hypothesized value. Thus, effect size is equal to [0.75 - 0.80] or - 0.05.)
* **Compute power**. The power of the test is the probability of rejecting the null hypothesis, assuming that the true population proportion is equal to the critical parameter value. Since the region of acceptance is 0.734 to 1.00, the null hypothesis will be rejected when the sample proportion is less than 0.734.   
    
  Therefore, we need to compute the probability that the sample proportion will be less than 0.734. To do this, we take the following steps:
  + Assume that the sampling distribution of the mean is normally distributed. (Because the sample size is relatively large, this assumption can be justified by the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central_limit_theorem).)
  + Assume that the mean of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) is the critical parameter value, 0.75. (This assumption is justified because, for the purpose of calculating power, we assume that the true population proportion is equal to the critical parameter value. And the mean of all possible sample proportions is equal to the population proportion. Hence, the mean of the sampling distribution is equal to the critical parameter value.)
  + Compute the standard deviation of the sampling distribution. In a [previous lesson](http://stattrek.com/AP-Statistics-4/Standard-Error.aspx?Tutorial=stat), we showed that the standard deviation of the sample estimate of a proportion σP is:

σP = sqrt[ P \* ( 1 - P; ) / n ]

where P is the population proportion and n is the sample size. Therefore,

σP = sqrt[ ( 0.75 \* 0.25 ) / 100 ] = 0.0433

* Following these steps, we can assess the probability that the probability that the sample proportion will be less than 0.734. This is easy to do, using the [Normal Calculator](http://stattrek.com/Tables/Normal.aspx). We enter the following values into the calculator: normal random variable = 0.734; mean = 0.75; and standard deviation = 0.0433. Given these inputs, we find that the cumulative probability is 0.36. This means that if the true population proportion is 0.75, then the probability that the sample proportion will be less than 0.734 is 0.36. Thus, the power of the test is 0.36, which means that the probability of making a Type II error is 1 - 0.36, which equals 0.64.

Chi-Square Goodness of Fit Test

This lesson explains how to conduct a **chi-square goodness of fit test**. The test is applied when you have one [categorical variable](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company. The [sample problem](http://stattrek.com/chi-square-test/goodness-of-fit.aspx?tutorial=stat#example1) at the end of the lesson considers this example.

When to Use the Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is appropriate when the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The variable under study is [categorical](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable).
* The expected value of the number of sample observations in each [level](http://stattrek.com/Help/Glossary.aspx?Target=Level) of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) (H0) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis) (Ha). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

For a chi-square goodness of fit test, the hypotheses take the following form.

|  |
| --- |
| H0: The data are consistent with a specified distribution.  Ha: The data are *not* consistent with a specified distribution. |

Typically, the null hypothesis (H0) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (Ha) is that *at least* one of the specified proportions is not true.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [chi-square goodness of fit test](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20goodness%20of%20fit%20test) to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square goodness of fit test is described in the next section, and demonstrated in the sample problem at the end of this lesson.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic.

* Degrees of freedom. The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is equal to the number of levels (k) of the categorical variable minus 1: DF = k - 1 .
* Expected frequency counts. The expected frequency counts at each level of the categorical variable are equal to the sample size times the hypothesized proportion from the null hypothesis

Ei = npi

where Ei is the expected frequency count for the *i*th level of the categorical variable, n is the total sample size, and pi is the hypothesized proportion of observations in level *i*.

* Test statistic. The test statistic is a chi-square random variable (Χ2) defined by the following equation.

Χ2 = Σ [ (Oi - Ei)2 / Ei ]

where Oi is the observed frequency count for the *i*th level of the categorical variable, and Ei is the expected frequency count for the *i*th level of the categorical variable.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  + Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
  + Alternative hypothesis: At least one of the proportions in the null hypothesis is false.
* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square goodness of fit test](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20goodness%20of%20fit%20test) of the null hypothesis.
* **Analyze sample data**. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](http://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = k - 1 = 3 - 1 = 2   
  
(Ei) = n \* pi  
(E1) = 100 \* 0.30 = 30  
(E2) = 100 \* 0.60 = 60  
(E3) = 100 \* 0.10 = 10   
  
Χ2 = Σ [ (Oi - Ei)2 / Ei ]   
Χ2 = [ (50 - 30)2 / 30 ] + [ (45 - 60)2 / 60 ] + [ (5 - 10)2 / 10 ]  
Χ2 = (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58

where DF is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample, Ei is the expected frequency count for level i, Oi is the observed frequency count for level i, and Χ2 is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

We use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to find P(Χ2 > 19.58) = 0.0001.

* **Interpret results**. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.

Chi-Square Test of Homogeneity

This lesson explains how to conduct a **chi-square test of homogeneity**. The test is applied to a single [categorical variable](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from two or more different populations. It is used to determine whether frequency counts are distributed identically across different populations.

For example, in a survey of TV viewing preferences, we might ask respondents to identify their favorite program. We might ask the same question of two different populations, such as males and females. We could use a chi-square test for homogeneity to determine whether male viewing preferences differed significantly from female viewing preferences. The [sample problem](http://stattrek.com/chi-square-test/homogeneity.aspx?tutorial=stat#example1) at the end of the lesson considers this example.

When to Use Chi-Square Test for Homogeneity

The test procedure described in this lesson is appropriate when the following conditions are met:

* For each population, the sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The variable under study is [categorical](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable).
* If sample data are displayed in a [contingency table](http://stattrek.com/Help/Glossary.aspx?Target=Contingency%20table) (Populations x Category levels), the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

Suppose that data were sampled from *r* populations, and assume that the categorical variable had *c* levels. At any specified level of the categorical variable, the null hypothesis states that each population has the same proportion of observations. Thus,

|  |
| --- |
| H0: Plevel 1 of population 1 = Plevel 1 of population 2 = . . . = Plevel 1 of population r H0: Plevel 2 of population 1 = Plevel 2 of population 2 = . . . = Plevel 2 of population r . . .  H0: Plevel c of population 1 = Plevel c of population 2 = . . . = Plevel c of population r |

The alternative hypothesis (Ha) is that *at least* one of the null hypothesis statements is false.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [chi-square test for homogeneity](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20homogeneity) to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square test for homogeneity is described in the next section.

Analyze Sample Data

Using sample data from the contingency tables, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic. The analysis described in this section is illustrated in the sample problem at the end of this lesson.

* **Degrees of freedom.** The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is equal to:

DF = (r - 1) \* (c - 1)

where r is the number of populations, and c is the number of levels for the categorical variable.

* **Expected frequency counts.** The expected frequency counts are computed separately for each population at each level of the categorical variable, according to the following formula.

Er,c = (nr \* nc) / n

where Er,c is the expected frequency count for population r at level *c* of the categorical variable, nr is the total number of observations from population r, nc is the total number of observations at treatment level *c*, and n is the total sample size.

* **Test statistic.** The test statistic is a chi-square random variable (Χ2) defined by the following equation.

Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]

where Or,c is the observed frequency count in population *r* for level *c* of the categorical variable, and Er,c is the expected frequency count in population *r* for level *c* of the categorical variable.

* **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

In a study of the television viewing habits of children, a developmental psychologist selects a random sample of 300 first graders - 100 boys and 200 girls. Each child is asked which of the following TV programs they like best: The Lone Ranger, Sesame Street, or The Simpsons. Results are shown in the [contingency table](http://stattrek.com/Help/Glossary.aspx?Target=Contingency%20table) below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Viewing Preferences** | | | **Row total** |
| **Lone Ranger** | **Sesame Street** | **The Simpsons** |
| **Boys** | 50 | 30 | 20 | 100 |
| **Girls** | 50 | 80 | 70 | 200 |
| **Column total** | 100 | 110 | 90 | 300 |

Do the boys' preferences for these TV programs differ significantly from the girls' preferences? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  + Null hypothesis: The null hypothesis states that the proportion of boys who prefer the Lone Ranger is identical to the proportion of girls. Similarly, for the other programs. Thus,

|  |
| --- |
| H0: Pboys who prefer Lone Ranger = Pgirls who prefer Lone Ranger H0: Pboys who prefer Sesame Street = Pgirls who prefer Sesame Street H0: Pboys who prefer The Simpsons = Pgirls who prefer The Simpsons |

* + Alternative hypothesis: At least one of the null hypothesis statements is false.
* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square test for homogeneity](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20homogeneity).
* **Analyze sample data**. Applying the chi-square test for homogeneity to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](http://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2   
  
Er,c = (nr \* nc) / n  
E1,1 = (100 \* 100) / 300 = 10000/300 = 33.3  
E1,2 = (100 \* 110) / 300 = 11000/300 = 36.7  
E1,3 = (100 \* 90) / 300 = 9000/300 = 30.0  
E2,1 = (200 \* 100) / 300 = 20000/300 = 66.7  
E2,2 = (200 \* 110) / 300 = 22000/300 = 73.3  
E2,3 = (200 \* 90) / 300 = 18000/300 = 60.0  
  
  
Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]   
Χ2 = (50 - 33.3)2/33.3 + (30 - 36.7)2/36.7 + (20 - 30)2/30  
+ (50 - 66.7)2/66.7 + (80 - 73.3)2/73.3 + (70 - 60)2/60  
Χ2 = (16.7)2/33.3 + (-6.7)2/36.7 + (-10.0)2/30 + (-16.7)2/66.7 + (3.3)2/73.3 + (10)2/60  
Χ2 = 8.38 + 1.22 + 3.33 + 4.18 + 0.61 + 1.67 = 19.39

where DF is the degrees of freedom, r is the number of populations, c is the number of levels of the categorical variable, nr is the number of observations from population *r*, nc is the number of observations from level *c* of the categorical variable, n is the number of observations in the sample, Er,c is the expected frequency count in population *r* for level *c*, and Or,c is the observed frequency count in population *r* for level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.39.

We use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to find P(Χ2 > 19.39) = 0.0000. (The actual P-value, of course, is not exactly zero. If the Chi-Square Distribution Calculator reported more than four decimal places, we would find that the actual P-value is a very small number that is less than 0.00005 and greater than zero.)

* **Interpret results**. Since the P-value (0.0000) is less than the significance level (0.05), we reject the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and the expected frequency count was at least 5 in each population at each level of the categorical variable.

Chi-Square Test for Independence

This lesson explains how to conduct a **chi-square test for independence**. The test is applied when you have two [categorical variables](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The [sample problem](http://stattrek.com/chi-square-test/independence.aspx?tutorial=stat#example1) at the end of the lesson considers this example.

When to Use Chi-Square Test for Independence

The test procedure described in this lesson is appropriate when the following conditions are met:

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The variables under study are each [categorical](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable).
* If sample data are displayed in a [contingency table](http://stattrek.com/Help/Glossary.aspx?Target=Contingency%20table), the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Suppose that Variable A has *r* levels, and Variable B has *c* levels. The [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

|  |
| --- |
| H0: Variable A and Variable B are independent.  Ha: Variable A and Variable B are not independent. |

The [alternative hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis) is that knowing the level of Variable A *can* help you predict the level of Variable B.

**Note:** Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use the [chi-square test for independence](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20independence) to determine whether there is a significant relationship between two categorical variables.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

* **Degrees of freedom.** The [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is equal to:

DF = (r - 1) \* (c - 1)

where r is the number of levels for one catagorical variable, and c is the number of levels for the other categorical variable.

* **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r \* c expected frequencies, according to the following formula.

Er,c = (nr \* nc) / n

where Er,c is the expected frequency count for level *r* of Variable A and level *c* of Variable B, nr is the total number of sample observations at level r of Variable A, nc is the total number of sample observations at level *c* of Variable B, and n is the total sample size.

* **Test statistic.** The test statistic is a chi-square random variable (Χ2) defined by the following equation.

Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]

where Or,c is the observed frequency count at level *r* of Variable A and level *c* of Variable B, and Er,c is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

* **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the [contingency table](http://stattrek.com/Help/Glossary.aspx?Target=Contingency%20table) below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Voting Preferences** | | | **Row total** |
| **Republican** | **Democrat** | **Independent** |
| **Male** | 200 | 150 | 50 | 400 |
| **Female** | 250 | 300 | 50 | 600 |
| **Column total** | 450 | 450 | 100 | 1000 |

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an alternative hypothesis.

H0: Gender and voting preferences are independent.   
Ha: Gender and voting preferences are not independent.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square test for independence](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20independence).
* **Analyze sample data**. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](http://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2   
  
Er,c = (nr \* nc) / n  
E1,1 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,2 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,3 = (400 \* 100) / 1000 = 40000/1000 = 40  
E2,1 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,2 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,3 = (600 \* 100) / 1000 = 60000/1000 = 60  
  
  
Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]   
Χ2 = (200 - 180)2/180 + (150 - 180)2/180 + (50 - 40)2/40  
+ (250 - 270)2/270 + (300 - 270)2/270 + (50 - 60)2/60  
Χ2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60  
Χ2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, nr is the number of observations from level *r* of gender, nc is the number of observations from level *c* of voting preference, n is the number of observations in the sample, Er,c is the expected frequency count when gender is level *r* and voting preference is level *c*, and Or,c is the observed frequency count when gender is level *r* voting preference is level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the [Chi-Square Distribution Calculator](http://stattrek.com/Tables/ChiSquare.aspx) to find P(Χ2 > 16.2) = 0.0003.

* **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variables under study were categorical, and the expected frequency count was at least 5 in each cell of the contingency table.

How to Describe Data Patterns in Statistics

Graphic displays are useful for seeing patterns in data. Patterns in data are commonly described in terms of: center, spread, shape, and unusual features. Some common distributions have special descriptive labels, such as symmetric, bell-shaped, skewed, etc.

Center



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Graphically, the **center** of a distribution is located at the [median](http://stattrek.com/Help/Glossary.aspx?Target=median) of the distribution. This is the point in a graphic display where about half of the observations are on either side. In the chart to the right, the height of each column indicates the frequency of observations. Here, the observations are centered over 4.

Spread

The **spread** of a distribution refers to the variability of the data. If the observations cover a wide [range](http://stattrek.com/Help/Glossary.aspx?Target=range), the spread is larger. If the observations are clustered around a single value, the spread is smaller.



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |  | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| Less spread |  | More spread |

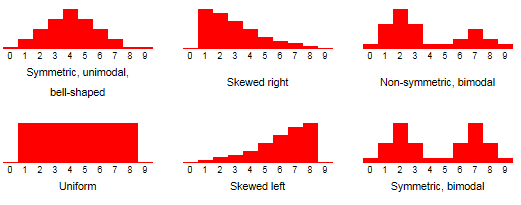
Consider the figures above. In the figure on the left, data values range from 3 to 7; whereas in the figure on the right, values range from 1 to 9. The figure on the right is more variable, so it has the greater spread.

Shape

The shape of a distribution is described by the following characteristics.

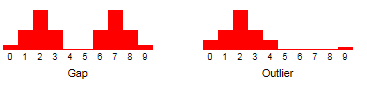
* **Symmetry**. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other.
* **Number of peaks**. Distributions can have few or many peaks. Distributions with one clear peak are called **unimodal**, and distributions with two clear peaks are called **bimodal**. When a symmetric distribution has a single peak at the center, it is referred to as **bell-shaped**.
* **Skewness**. When they are displayed graphically, some distributions have many more observations on one side of the graph than the other. Distributions with fewer observations on the right (toward higher values) are said to be **skewed right**; and distributions with fewer observations on the left (toward lower values) are said to be **skewed left**.
* **Uniform**. When the observations in a set of data are equally spread across the range of the distribution, the distribution is called a **uniform distribution**. A uniform distribution has no clear peaks.

Here are some examples of distributions and shapes.

Unusual Features

Sometimes, statisticians refer to unusual features in a set of data. The two most common unusual features are gaps and outliers.

* **Gaps**. Gaps refer to areas of a distribution where there are no observations. The first figure below has a gap; there are no observations in the middle of the distribution.
* **Outliers**. Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called outliers. The second figure below illustrates a distribution with an outlier. Except for one lonely observation (the outlier on the extreme right), all of the observations fall between 0 and 4. As a "rule of thumb", an extreme value is often considered to be an outlier if it is at least 1.5 [interquartile ranges](http://stattrek.com/Help/Glossary.aspx?Target=Interquartile%20range) below the first [quartile](http://stattrek.com/Help/Glossary.aspx?Target=quartile) (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3).



Bar Charts and Histograms

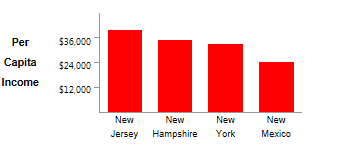
Like [dotplots](http://stattrek.com/Help/Glossary.aspx?Target=Dotplot), **bar charts** and **histograms** are used to compare the sizes of different groups.

Bar Charts

A bar chart is made up of columns plotted on a graph. Here is how to read a bar chart.

* The columns are positioned over a label that represents a [categorical variable](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable).
* The height of the column indicates the size of the group defined by the column label.

The bar chart below shows average per capita income for the four "New" states - New Jersey, New York, New Hampshire, and New Mexico.

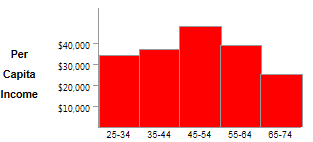


Histograms

Like a bar chart, a histogram is made up of columns plotted on a graph. Usually, there is no space between adjacent columns. Here is how to read a histogram.

* The columns are positioned over a label that represents a [quantitative variable](http://stattrek.com/Help/Glossary.aspx?Target=Quantitative%20variable).
* The column label can be a single value or a range of values.
* The height of the column indicates the size of the group defined by the column label.

The histogram below shows per capita income for five age groups.



The Difference Between Bar Charts and Histograms

Here is the main difference between bar charts and histograms. With bar charts, each column represents a group defined by a categorical variable; and with histograms, each column represents a group defined by a quantitative variable.

One implication of this distinction: it is always appropriate to talk about the [skewness](http://stattrek.com/Help/Glossary.aspx?Target=skewness) of a histogram; that is, the tendency of the observations to fall more on the low end or the high end of the X axis.

With bar charts, however, the X axis does not have a low end or a high end; because the labels on the X axis are categorical - not quantitative. As a result, it is not appropriate to comment on the skewness of a bar chart.

Test Your Understanding

**Problem 1**

Consider the histograms below.



Which of the following statements are true?

I. Both data sets are symmetric.   
II. Labels on the X axis are quantitative.

(A) I only   
(B) II only   
(C) I and II   
(D) Neither is true.   
(E) There is insufficient information to answer this question.

**Solution**

The correct answer is (C). Both histograms are mirror images around their center, so both are [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry). With bar charts, the labels on the X axis are categorical; with histograms, the labels are quantitative. Both of these charts are histograms. Therefore, their labels are quantitative.

Stemplots (aka, Stem and Leaf Plots)

Although a [histogram](http://stattrek.com/Help/Glossary.aspx?Target=histogram) shows how observations are distributed across groups, it does not show the exact values of individual observations. A different kind of graphical display, called a **stemplot** or a **stem and leaf plot**, does show exact values of individual observations.

Stemplots

A stemplot is used to display quantitative data, generally from small data sets (50 or fewer observations). The stemplot below shows IQ scores for 30 sixth graders.

|  |  |
| --- | --- |
| **Stems** 15 14 13 12 11 10 9 8  Key: 11 | **Leaves** 1   2 6 4 5 7 9 1 2 2 2 5 7 9 9 0 2 3 4 4 5 7 8 9 9 1 1 4 7 8  7 represents an IQ score of 117 |

In a stemplot, the entries on the left are called stems; and the entries on the right are called leaves. In the example above, the stems are tens (8 represents 80, 9 represents 90, 10 represents 100, and so on); and the leaves are ones. However, the stems and leaves could be other units - millions, thousands, ones, tenths, etc.

Some stemplots include a key to help the user interpret the display correctly. The key in the stemplot above indicates that a stem of 11 with a leaf of 7 represents an IQ score of 117.

Looking at the example above, you should be able to quickly describe the distribution of IQ scores. Most of the scores are clustered between 90 and 109, with the center falling in the neighborhood of 100. The scores range from a low of 81 (two students have an IQ of 81) to a high of 151. The high score of 151 might be classified as an [outlier](http://stattrek.com/Help/Glossary.aspx?Target=Outlier).

**Note:** In the example above, the stems and leaves are explicitly labeled for educational purposes. In the real world, however, stemplots usually do not include explicit labels for the stems and leaves.

Test Your Understanding

**Problem 1**

The stemplot below shows the number of hot dogs eaten by contestants in a recent hot dog eating contest. Assume that the stems represents tens and the leaves represent ones.

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

Which of the following statements is true?

I. The range is 70.   
II. The median is 46.

(A) I only   
(B) II only   
(C) I and II   
(D) Neither is true.   
(E) There is insufficient information to answer this question.

**Solution**

The correct answer is (C). The [range](http://stattrek.com/Help/Glossary.aspx?Target=Range) is equal to the biggest value minus the smallest value. The biggest value is 81, and the smallest value is 11; so the range is equal to 81 -11 or 70. Since the data set has an even number of values, the [median](http://stattrek.com/Help/Glossary.aspx?Target=Median) is the average of the middle two values - 45 and 47. That is, the median is (45 + 47)/2 or 46.

Data Collection Methods

To derive conclusions from data, we need to know how the data were collected; that is, we need to know the method(s) of data collection.

Methods of Data Collection

For this tutorial, we will cover four methods of data collection.

* **Census**. A census is a study that obtains data from every member of a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population). In most studies, a census is not practical, because of the cost and/or time required.
* **Sample survey**. A sample survey is a study that obtains data from a subset of a population, in order to estimate population attributes.
* **Experiment**. An experiment is a controlled study in which the researcher attempts to understand cause-and-effect relationships. The study is "controlled" in the sense that the researcher controls (1) how subjects are assigned to groups and (2) which treatments each group receives.  
    
  In the analysis phase, the researcher compares group scores on some [dependent variable](http://stattrek.com/Help/Glossary.aspx?Target=Dependent%20variable). Based on the analysis, the researcher draws a conclusion about whether the treatment ( [independent variable](http://stattrek.com/Help/Glossary.aspx?Target=Independent%20variable)) had a causal effect on the dependent variable.
* **Observational study**. Like experiments, observational studies attempt to understand cause-and-effect relationships. However, unlike experiments, the researcher is not able to control (1) how subjects are assigned to groups and/or (2) which treatments each group receives.

Data Collection Methods: Pros and Cons

Each method of data collection has advantages and disadvantages.

* **Resources**. When the population is large, a sample survey has a big resource advantage over a census. A well-designed sample survey can provide very precise estimates of population parameters - quicker, cheaper, and with less manpower than a census.
* **Generalizability**. Generalizability refers to the appropriateness of applying findings from a study to a larger population. Generalizability requires random selection. If participants in a study are randomly selected from a larger population, it is appropriate to generalize study results to the larger population; if not, it is not appropriate to generalize.   
    
  Observational studies do not feature random selection; so generalizing from the results of an observational study to a larger population can be a problem.
* **Causal inference**. Cause-and-effect relationships can be teased out when subjects are randomly assigned to groups. Therefore, experiments, which allow the researcher to control assignment of subjects to treatment groups, are the best method for investigating causal relationships.

Test Your Understanding

**Problem**

Which of the following statements are true?

I. A sample survey is a type of experiment.   
II. An observational study requires fewer resources than an experiment.   
III. The best method for investigating causal relationships is an observational study.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above.   
(E) None of the above.

**Solution**

The correct answer is (E). Unlike an experiment, a sample survey does not require the researcher to assign treatments to survey respondents. Therefore, a sample survey is not necessarily an experiment. A sample survey could be an observational study, rather than an experiment. An observational study may or may not require fewer resources (time, money, manpower) than an experiment. The best method for investigating causal relationships is an experiment - not an observational study - because an experiment features randomized assignment of subjects to treatment groups.

Survey Sampling Methods

**Sampling method** refers to the way that observations are selected from a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) to be in the [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) for a [sample survey](http://stattrek.com/Help/Glossary.aspx?Target=Sample%20survey).

Population Parameter vs. Sample Statistic

The reason for conducting a sample survey is to estimate the value of some attribute of a population.

* **Population parameter**. A population parameter is the true value of a population attribute.
* **Sample statistic**. A sample statistic is an estimate, based on sample data, of a population parameter.

Consider this example. A public opinion pollster wants to know the percentage of voters that favor a flat-rate income tax. The *actual* percentage of all the voters is a population parameter. The *estimate* of that percentage, based on sample data, is a sample statistic.

The quality of a sample statistic (i.e., accuracy, precision, representativeness) is strongly affected by the way that sample observations are chosen; that is., by the sampling method.

Probability vs. Non-Probability Samples

As a group, sampling methods fall into one of two categories.

* **Probability samples**. With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.
* **Non-probability samples**. With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

Non-probability sampling methods offer two potential advantages - convenience and cost. The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from population parameters. Only probability sampling methods permit that kind of analysis.

Non-Probability Sampling Methods

Two of the main types of non-probability sampling methods are voluntary samples and convenience samples.

* **Voluntary sample**. A voluntary sample is made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey.   
    
  Suppose, for example, that a news show asks viewers to participate in an on-line poll. This would be a volunteer sample. The sample is chosen by the viewers, not by the survey administrator.
* **Convenience sample**. A convenience sample is made up of people who are easy to reach.   
    
  Consider the following example. A pollster interviews shoppers at a local mall. If the mall was chosen because it was a convenient site from which to solicit survey participants and/or because it was close to the pollster's home or business, this would be a convenience sample.

Probability Sampling Methods

The main types of probability sampling methods are simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic random sampling. The key benefit of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid.

* **Simple random sampling**. Simple random sampling refers to any sampling method that has the following properties.
  + The population consists of N objects.
  + The sample consists of n objects.
  + If all possible samples of n objects are equally likely to occur, the sampling method is called simple random sampling.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the N population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.

* **Stratified sampling**. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called **strata**.   
    
  As a example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.
* **Cluster sampling**. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.   
    
  Note the difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters.
* **Multistage sampling**. With multistage sampling, we select a sample by using combinations of different sampling methods.   
    
  For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.
* **Systematic random sampling**. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* elements on the population list. Thereafter, we select every *kth* element on the list.   
    
  This method is different from simple random sampling since every possible sample of *n* elements is not equally likely.

Test Your Understanding

**Problem**

An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.

Is this an example of a simple random sample?

(A) Yes, because each buyer in the sample was randomly sampled.   
(B) Yes, because each buyer in the sample had an equal chance of being sampled.   
(C) Yes, because car buyers of every brand were equally represented in the sample.   
(D) No, because every possible 400-buyer sample did not have an equal chance of being chosen.   
(E) No, because the population consisted of purchasers of four different brands of car.

**Solution**

The correct answer is (D). A [simple random sample](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling) requires that every [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) of size *n* (in this problem, *n* is equal to 400) has an equal chance of being selected. In this problem, there was a 100 percent chance that the sample would include 100 purchasers of each brand of car. There was zero percent chance that the sample would include, for example, 99 Ford buyers, 101 Honda buyers, 100 Toyota buyers, and 100 GM buyers. Thus, all possible samples of size 400 did not have an equal chance of being selected; so this cannot be a simple random sample.

The fact that each buyer in the sample was randomly sampled is a necessary condition for a simple random sample, but it is not sufficient. Similarly, the fact that each buyer in the sample had an equal chance of being selected is characteristic of a simple random sample, but it is not sufficient. The sampling method in this problem used random sampling and gave each buyer an equal chance of being selected; but the sampling method was actually [stratified random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Stratified%20sampling).

The fact that car buyers of every brand were equally represented in the sample is irrelevant to whether the sampling method was simple random sampling. Similarly, the fact that population consisted of buyers of different car brands is irrelevant.

Bias in Survey Sampling

In survey sampling, **bias** refers to the tendency of a sample [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic) to systematically over- or under-estimate a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter).

Bias Due to Unrepresentative Samples

A good [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) is **representative**. This means that each sample point represents the attributes of a known number of [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) elements.

Bias often occurs when the survey sample does not accurately represent the population. The bias that results from an unrepresentative sample is called **selection bias**. Some common examples of selection bias are described below.

* **Undercoverage**. Undercoverage occurs when some members of the population are inadequately represented in the sample. A classic example of undercoverage is the *Literary Digest* voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from undercoverage of low-income voters, who tended to be Democrats.  
    
  How did this happen? The survey relied on a [convenience sample](http://stattrek.com/Help/Glossary.aspx?Target=Convenience%20sample), drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent. Undercoverage is often a problem with convenience samples.
* **Nonresponse bias**. Sometimes, individuals chosen for the sample are unwilling or unable to participate in the survey. Nonresponse bias is the bias that results when respondents differ in meaningful ways from nonrespondents. The *Literary Digest* survey illustrates this problem. Respondents tended to be Landon supporters; and nonrespondents, Roosevelt supporters. Since only 25% of the sampled voters actually completed the mail-in survey, survey results overestimated voter support for Alfred Landon.  
    
  The *Literary Digest* experience illustrates a common problem with mail surveys. Response rate is often low, making mail surveys vulnerable to nonresponse bias.
* **Voluntary response bias**. Voluntary response bias occurs when sample members are self-selected volunteers, as in [voluntary samples](http://stattrek.com/Help/Glossary.aspx?Target=Voluntary%20sample). An example would be call-in radio shows that solicit audience participation in surveys on controversial topics (abortion, affirmative action, gun control, etc.). The resulting sample tends to overrepresent individuals who have strong opinions.

**Random sampling** is a procedure for sampling from a population in which (a) the selection of a sample unit is based on chance and (b) every element of the population has a known, non-zero probability of being selected. Random sampling helps produce representative samples by eliminating voluntary response bias and guarding against undercoverage bias. All probability sampling methods rely on random sampling.

Bias Due to Measurement Error

A poor measurement process can also lead to bias. In survey research, the measurement process includes the environment in which the survey is conducted, the way that questions are asked, and the state of the survey respondent.

**Response bias** refers to the bias that results from problems in the measurement process. Some examples of response bias are given below.

* **Leading questions**. The wording of the question may be loaded in some way to unduly favor one response over another. For example, a satisfaction survey may ask the respondent to indicate where she is satisfied, dissatisfied, or very dissatified. By giving the respondent one response option to express satisfaction and two response options to express dissatisfaction, this survey question is biased toward getting a dissatisfied response.
* **Social desirability**. Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes or illegal activities in a survey, particularly if survey results are not confidential. Instead, their responses may be biased toward what they believe is socially desirable.

Sampling Error and Survey Bias

A survey produces a sample statistic, which is used to estimate a population parameter. If you repeated a survey many times, using different samples each time, you might get a different sample statistic with each replication. And each of the different sample statistics would be an estimate for the *same* population parameter.

If the statistic is unbiased, the average of all the statistics from all possible samples will equal the true population parameter; even though any individual statistic may differ from the population parameter. The variability among statistics from different samples is called **sampling error**.

Increasing the sample size tends to reduce the sampling error; that is, it makes the sample statistic less variable. However, increasing sample size does not affect survey bias. A large sample size cannot correct for the methodological problems (undercoverage, nonresponse bias, etc.) that produce survey bias. The *Literary Digest* example discussed above illustrates this point. The sample size was very large - over 2 million surveys were completed; but the large sample size could not overcome problems with the sample - undercoverage and nonresponse bias.

Test Your Understanding

**Problem**

Which of the following statements are true?

I. Random sampling is a good way to reduce response bias.   
II. To guard against bias from undercoverage, use a convenience sample.   
III. Increasing the sample size tends to reduce survey bias.   
IV. To guard against nonresponse bias, use a mail-in survey.

(A) I only   
(B) II only   
(C) III only   
(D) IV only   
(E) None of the above.

**Solution**

The correct answer is (E). None of the statements is true. [Random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Random%20sampling) provides strong protection against bias from [undercoverage](http://stattrek.com/Help/Glossary.aspx?Target=Undercoverage) bias and [voluntary response bias](http://stattrek.com/Help/Glossary.aspx?Target=Voluntary%20response%20bias); but it is not effective against [response bias](http://stattrek.com/Help/Glossary.aspx?Target=Response%20bias). A [convenience sample](http://stattrek.com/Help/Glossary.aspx?Target=Convenience%20sample) does not protect against undercoverage bias; in fact, it sometimes causes undercoverage bias. Increasing sample size does not affect survey [bias](http://stattrek.com/Help/Glossary.aspx?Target=Bias). And finally, using a mail-in survey does not prevent [nonresponse bias](http://stattrek.com/Help/Glossary.aspx?Target=Nonresponse%20bias). In fact, mail-in surveys are quite vulnerable to nonresponse bias.

Introduction to Survey Sampling

**Sampling** refers to the process of choosing a [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) of elements from a total [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) of elements.

Probability vs. Non-Probability Sampling

Statisticians distinguish between two broad categories of sampling.

* **Probability sampling**. With probability sampling, every element of the population has a known probability of being included in the sample.
* **Non-probability sampling**. With non-probability sampling, we cannot specify the probability that each element will be included in the sample.

Each approach has advantages and disadvantages. The main advantages of non-probability sampling are convenience and cost. However, with non-probability samples, we cannot make probability statements about our sample statistics. For example, we cannot compute a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval) for an estimation problem or a [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_acceptance) for a hypothesis test.

Probability samples, in contrast, allow us to make probability statements about sample statistics. We can estimate the extent to which a sample [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic) is likely to differ from a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter). The remainder of this tutorial focuses on probability sampling.

Quality of Survey Results

When researchers describe the quality of survey results, they may use one or more of the following terms.

* **Accuracy**. Accuracy refers to how close a sample [statistic](http://stattrek.com/Help/Glossary.aspx?Target=Statistic) is to a population [parameter](http://stattrek.com/Help/Glossary.aspx?Target=Parameter). Thus, if you know that a sample mean is 99 and the true population mean is 100, you can make a statement about the sample accuracy. For example, you might say the sample mean is accurate to within 1 unit.
* **Precision**. Precision refers to how close estimates from different samples are to each other. For example, the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error) is a measure of precision. When the standard error is small, estimates from different samples will be close in value; and vice versa. Precision is inversely related to standard error. When the standard error is small, sample estimates are more precise; when the standard error is large, sample estimates are less precise.
* **Margin of error**. The margin of error expresses the maximum expected difference between the true population parameter and a sample estimate of that parameter. To be meaningful, the margin of error should be qualified by a probability statement. For example, a pollster might report that 50% of voters will choose the Democratic candidate. To indicate the quality of the survey result, the pollster might add that the margin of error is +5%, with a [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level) of 90%. This means that if the same sampling method were applied to different samples, the true percentage of Democratic voters would fall within the margin of error 90% of the time.   
    
  The margin of error is equal to half of the width of the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval). In a previous lesson, the tutorial described [how to construct a confidence interval](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx?Tutorial=Stat).

Sample Design

A **sample design** can be described by two factors.

* **Sampling method**. Sampling method refers to the rules and procedures by which some elements of the population are included in the sample. Some common sampling methods are described elsewhere in the tutorial (see [simple random sampling](http://stattrek.com/Lesson6/SRS.aspx), [stratified sampling](http://stattrek.com/Lesson6/STR.aspx), and [cluster sampling](http://stattrek.com/Lesson6/CLS.aspx).)
* **Estimator**. The estimation process for calculating sample statistics is called the estimator. Different sampling methods may use different estimators. For example, the formula for computing a mean score with a simple random sample is different from the formula for computing a mean score with a stratified sample. Similarly, the formula for the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error) may vary from one sampling method to the next.

The "best" sample design depends on survey objectives and on survey resources. For example, a researcher might select the most economical design that provides a desired level of precision. Or, if the budget is limited, a researcher might choose the design that provides the greatest precision without going over budget. Or other factors might guide the choice of sample design.

Analysis of Simple Random Samples

**Simple random sampling** refers to a sampling method that has the following properties.

* The population consists of *N* objects.
* The sample consists of *n* objects.
* All possible samples of *n* objects are equally likely to occur.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define a [confidence interval](http://stattrek.com/statistics/dictionary.aspx?definition=confidence_interval) around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the *N* population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects *n* numbers. Population members having the selected numbers are included in the sample.

Notation

The following notation is helpful, when we talk about simple random sampling.

* σ: The known [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation) of the population.
* σ2: The known [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) of the population.
* *P*: The true population [proportion](http://stattrek.com/Help/Glossary.aspx?Target=Proportion).
* *N*: The number of observations in the population.
* x: The sample estimate of the population mean.
* *s*: The sample estimate of the standard deviation of the population.
* *s2*: The sample estimate of the population variance.
* *p*: The proportion of successes in the sample.
* *n*: The number of observations in the sample.
* *SD*: The standard deviation of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).
* *SE*: The [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). (This is an estimate of the standard deviation of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).)
* Σ = Summation symbol, used to compute sums over the sample. ( To illustrate its use, Σ xi = x1 + x2 + x3 + ... + xm-1 + xm )

The Variability of the Estimate

The [precision](http://stattrek.com/Help/Glossary.aspx?Target=Precision) of a [sample design](http://stattrek.com/Help/Glossary.aspx?Target=Sample_design) is directly related to the variability of the estimate. Two common measures of variability are the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) (SD) of the estimate and the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error) (SE) of the estimate. The tables below show how to compute both measures, assuming that the sample method is simple random sampling.

The first table shows how to compute variability for a mean score. Note that the table shows four sample designs. In two of the designs, the true population variance is known; and in two, it is estimated from sample data. Also, in two of the designs, the researcher sampled with replacement; and in two, without replacement.

|  |  |  |
| --- | --- | --- |
| **Population variance** | **Replacement strategy** | **Variability** |
| Known | With replacement | SD = sqrt [ σ2 / n ] |
| Known | Without replacement | SD = sqrt { [ ( N - n ) / ( N - 1 ) ] \* σ2 / n } |
| Estimated | With replacement | SE = sqrt [ s2 / n ] |
| Estimated | Without replacement | SE = sqrt { [ ( N - n ) / ( N - 1 ) ] \* s2 / n } |

The next table shows how to compute variability for a proportion. Like the previous table, this table shows four sample designs. In two of the designs, the true population proportion is known; and in two, it is estimated from sample data. Also, in two of the designs, the researcher sampled with replacement; and in two, without replacement.

|  |  |  |
| --- | --- | --- |
| **Population proportion** | **Replacement strategy** | **Variability** |
| Known | With replacement | SD = sqrt [ P \* ( 1 - P ) / n ] |
| Known | Without replacement | SD = sqrt { [ ( N - n ) / ( N - 1 ) ] \* P \* ( 1 - P ) / n } |
| Estimated | With replacement | SE = sqrt [ p \* ( 1 - p ) / ( n - 1 ) ] |
| Estimated | Without replacement | SE = sqrt [ [ ( N - n ) / ( N - 1 ) ] \* p \* ( 1 - p ) / n ] |

Sample Problem

This section presents a sample problem that illustrates how to analyze survey data when the sampling method is simple random sampling. (In a [subsequent lesson](http://stattrek.com/Lesson6/SamplingMethod.aspx?Tutorial=stat), we re-visit this problem and see how simple random sampling compares to other sampling methods.)

Sample Planning Wizard

The analysis of data collected via simple random sampling can be complex and time-consuming. Stat Trek's Sample Planning Wizard can help. The Wizard computes survey precision, sample size requirements, costs, etc., as well as estimates population parameters and tests hypotheses. It also creates a summary report that lists key findings and documents analytical techniques. Whenever you work with simple random samples, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Problem 1**

At the end of every school year, the state administers a reading test to a simple random sample drawn without replacement from a population of 20,000 third graders. This year, the test was administered to 36 students selected via simple random sampling. The test score from each sampled student is shown below:

50, 55, 60, 62, 62, 65, 67, 67, 70, 70, 70, 70, 72, 72, 73, 73, 75, 75,   
75, 78, 78, 78, 78, 80, 80, 80, 82, 82, 85, 85, 85, 88, 88, 90, 90, 90

Using sample data, estimate the mean reading achievement level in the population. Find the [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error) and the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval). Assume a 95% [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level).

*Solution:* Previously we described [how to compute the confidence interval for a mean score](http://stattrek.com/Lesson4/Mean.aspx?Tutorial=stat). We follow that process below.

* Identify a sample statistic. Since we are trying to estimate a population mean, we choose the sample mean as the sample statistic. The sample mean is:

x = Σ ( xi ) / n   
x = ( 50 + 55 + 60 + ... + 90 + 90 + 90 ) / 36 = 75

Therefore, based on data from the simple random sample, we estimate that the mean reading achievement level in the population is equal to 75.

* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 95% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx?Tutorial=stat) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard error of the sampling distribution. First, we estimate the variance of the test scores (s2). And then, we compute the standard error (SE).

s2 = Σ ( xi - x )2 / ( n - 1 )  
s2 = [ (50 - 75)2 + (55 - 75)2 + (60 - 75)2 + ... + (90 - 75)2 + (90 - 75)2 ] / 29 = 98.97

SE = sqrt { [ ( N - n ) / ( N - 1 ) ] \* s2 / n }   
SE = sqrt [ ( 0.998 ) \* 98.97 / 36 ] = 1.66

* + Find critical value. The critical value is a factor used to compute the margin of error. Based on the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem), we can assume that the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the mean is normally distributed. Therefore, we express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 95/100 = 0.05
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.975. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 1.96.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.96 \* 1.66 = 3.25
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 95% confidence interval is 71.75 to 78.25. And the margin of error is equal to 3.25. That is, we are 95% confident that the true population mean is in the range defined by 75 + 3.25.

Stratified Random Sampling

**Stratified random sampling** refers to a sampling method that has the following properties.

* The population consists of *N* elements.
* The population is divided into *H* groups, called **strata**.
* Each element of the population can be assigned to one, and only one, stratum.
* The number of observations within each stratum *Nh* is known, and N = N1 + N2 + N3 + ... + NH-1 + NH.
* The researcher obtains a [probability sample](http://stattrek.com/Help/Glossary.aspx?Target=Probability_sampling) from each stratum.

In this tutorial, we will assume that the researcher draws a [simple random sample](http://stattrek.com/Help/Glossary.aspx?Target=Simple_random_sampling) from each stratum.

Advantages and Disadvantages

Stratified sampling offers several advantages over simple random sampling.

* A stratified sample can provide greater precision than a simple random sample of the same size.
* Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money.
* A stratified sample can guard against an "unrepresentative" sample (e.g., an all-male sample from a mixed-gender population).
* We can ensure that we obtain sufficient sample points to support a separate analysis of any subgroup.

The main disadvantage of a stratified sample is that it may require more administrative effort than a simple random sample.

Proportionate Versus Disproportionate Statification

All stratified sampling designs fall into one of two categories, each of which has strengths and weaknesses as described below.

* **Proportionate stratification**. With proportionate stratification, the sample size of each stratum is proportionate to the population size of the stratum. This means that each stratum has the same [sampling fraction](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_fraction).
  + Proportionate stratification provides equal or better precision than a simple random sample of the same size.
  + Gains in precision are greatest when when values within strata are [homogeneous](http://stattrek.com/Help/Glossary.aspx?Target=Homogeneous).
  + Gains in precision accrue to all survey measures.
* **Disproportionate stratification**. With disproportionate stratification, the sampling fraction may vary from one stratum to the next.
  + The precision of the design may be very good or very poor, depending on how sample points are allocated to strata. The way to maximize precision through disproportionate stratification is discussed in a subsequent lesson (see [Statistics Tutorial: Sample Size Within Strata](http://stattrek.com/Lesson6/SampleSizeStrata.aspx)).
  + If variances differ across strata, disproportionate stratification can provide better precision than proportionate stratification, when sample points are correctly allocated to strata.
  + With disproportionate stratification, the researcher can maximize precision for a single important survey measure. However, gains in precision may not accrue to other survey measures.

**Recommendation** If costs and variances are about equal across strata, choose proportionate stratification over disproportionate stratification. If the variances or costs differ across strata, consider disproportionate stratification.

How to Analyze Stratified Random Samples

In this lesson, we describe how to analyze survey data from stratified random samples.

Notation

The following notation is helpful, when we talk about analyzing data from stratified samples.

* H: The number of [strata](http://stattrek.com/Help/Glossary.aspx?Target=Strata) in the population.
* N: The number of observations in the population.
* Nh: The number of observations in stratum *h* of the population.
* Ph: The true [proportion](http://stattrek.com/Help/Glossary.aspx?Target=Proportion) in stratum *h* of the population.
* σ2: The known [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) of the population.
* σ: The known [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard_deviation) of the population.
* σh: The known standard deviation in stratum *h* of the population.
* x: The sample estimate of the population mean.
* xh: The mean of observations from stratum *h* of the sample.
* ph: The proportion of successes in stratum *h* of the sample.
* *s*h: The sample estimate of the population standard deviation in stratum *h*.
* *s*h2: The sample estimate of the population variance in stratum *h*.
* *n*: The number of observations in the sample.
* nh: The number of observations in stratum *h* of the sample.
* *SD*: The standard deviation of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).
* *SE*: The [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). (This is an estimate of the standard deviation of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).)
* Σ: Summation symbol. ( To illustrate the use of the symbol, Σ xh = x1 + x2 + ... + xH-1 + xH )

How to Analyze Data From Stratified Samples

When it comes to analyzing data from stratified samples, there is good new and there is bad news.

First, the bad news. Different [sampling methods](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_method) use different formulas to estimate population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter) and to estimate [standard errors](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). The formulas that we have used so far in this tutorial work for simple random samples, but they are not right for stratified samples.

Now, the good news. Once you know the correct formulas, you can readily estimate population parameters and standard errors. And once you have the standard error, the procedures for computing other things (e.g., [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error), [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval), and [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_acceptance)) are largely the same for stratified samples as for simple random samples. The next two sections provide formulas that can be used with stratified sampling. The sample problem at the end of this lesson shows how to use these formulas to analyze data from stratified samples.

Measures of Central Tendency

The table below shows formulas that can be used with stratified sampling to estimate a population mean and a population proportion.

|  |  |
| --- | --- |
| **Population parameter** | **Formula for sample estimate** |
| Mean | Σ( Nh / N ) \* xh |
| Proportion | Σ( Nh / N ) \* ph |

Note that Nh/N is the [sampling fraction](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_fraction). Thus, to compute a sample estimate of the population mean or population proportion, we need to know the sampling fraction (i.e., we need to know the relative size of each [stratum](http://stattrek.com/Help/Glossary.aspx?Target=Strata)).

The Variability of the Estimate

The [precision](http://stattrek.com/Help/Glossary.aspx?Target=Precision) of a [sample design](http://stattrek.com/Help/Glossary.aspx?Target=Sample_design) is directly related to the variability of the estimate, which is measured by the [standard deviation](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20deviation) or [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). The tables below show how to compute the standard deviation (SD) and standard error (SE), assuming that the sample method is stratified random sampling.

The first table shows how to compute the varibility for a mean score. Note that the table shows four sample designs. In two of the designs, the true population variance is known; and in two, it is estimated from sample data. Also, in two of the designs, the researcher sampled with replacement; and in two, without replacement.

|  |  |  |
| --- | --- | --- |
| **Population variance** | **Replacement strategy** | **Variability** |
| Known | With replacement | SD = (1 / N) \* sqrt [ Σ ( Nh2 \* σh2 / nh ) ] |
| Known | Without replacement | SD = (1 / N) \* sqrt { Σ [ Nh3/( Nh - 1) ] \* ( 1 - nh / Nh ) \* σh2 / nh } |
| Estimated | With replacement | SE = (1 / N) \* sqrt [ Σ ( Nh2 \* sh2 / nh ) ] |
| Estimated | Without replacement | SE = (1 / N) \* sqrt { Σ [ Nh2 \* ( 1 - nh/Nh ) \* sh2 / nh ] } |

The next table shows how to compute the variability for a proportion. Like the previous table, this table shows four sample designs. In this case, however, the designs are based on whether the true population proportion is known and whether the design calls for sampling with or without replacement.

|  |  |  |
| --- | --- | --- |
| **Population proportion** | **Replacement strategy** | **Variability** |
| Known | With replacement | SD = (1 / N) \* sqrt { Σ [ Nh2 \* Ph \* ( 1 - Ph ) / nh ] } |
| Known | Without replacement | SD = (1 / N) \* sqrt ( Σ { [ Nh3/( Nh - 1) ] \* ( 1 - nh / Nh ) \* Ph \* ( 1 - Ph ) / nh } ) |
| Estimated | With replacement | SE = (1 / N) \* sqrt { Σ [ Nh2 \* ph \* ( 1 - ph ) / ( nh - 1 ) ] } |
| Estimated | Without replacement | SE = (1 / N) \* sqrt { Σ [ Nh2 \* ( 1 - nh/Nh ) \* ph \* ( 1 - ph ) / ( nh - 1 ) ] } |

Sample Problem

This section presents a sample problem that illustrates how to analyze survey data when the sampling method is proportionate stratified sampling. (In a [subsequent lesson](http://stattrek.com/Lesson6/SamplingMethod.aspx), we re-visit this problem and see how stratified sampling compares to other sampling methods.)

Sample Planning Wizard

The analysis of data collected via stratified random sampling can be complex and time-consuming. Stat Trek's Sample Planning Wizard can help. The Wizard computes survey precision, sample size requirements, costs, etc., as well as estimates population parameters and tests hypotheses. It also creates a summary report that lists key findings and documents analytical techniques. Whenever you work with stratified random samples, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Problem 1**

At the end of every school year, the state administers a reading test to a sample of third graders. The school system has 20,000 third graders, half boys and half girls.

This year, a proportionate stratified sample was used to select 36 students for testing. Because the population is half boy and half girl, one stratum consisted of 18 boys; the other, 18 girls. Test scores from each sampled student are shown below:

|  |  |
| --- | --- |
| **Boys** | 50, 55, 60, 62, 62, 65, 67, 67, 70, 70, 73, 73, 75, 78, 78, 80, 85, 90 |
| **Girls** | 70, 70, 72, 72, 75, 75, 78, 78, 80, 80, 82, 82, 85, 85, 88, 88, 90, 90 |

Using sample data, estimate the mean reading achievement level in the population. Find the [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error) and the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval). Assume a 95% [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level).

*Solution:* Previously we described [how to compute the confidence interval for a mean score](http://stattrek.com/Lesson4/Mean.aspx?Tutorial=Stat). We follow that process below.

* Identify a sample statistic. For this problem, we use the overall sample mean to estimate the population mean. To compute the overall sample mean, we need to compute the sample means for each stratum. The stratum mean for boys is equal to:

xboys = Σ ( xi ) / n   
xboys = ( 50 + 55 + 60 + ... + 80 + 85 + 90 ) / 18 = 70

The stratum mean for girls is computed similarly. It is equal to 80. Therefore, overall sample mean is:

x = Σ( Nh / N ) \* xh  
x = ( 10,000 / 20,000 ) \* 70 + ( 10,000 / 20,000 ) \* 80 = 75

Therefore, based on data from the sample strata, we estimate that the mean reading achievement level in the population is equal to 75.

* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 95% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx?Tutorial=Stat) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard error of the sampling distribution. First, we estimate the variance of the test scores (sh2) within each stratum. And then, we compute the standard error (SE). For boys, the within-stratum sample variance is equal to:

sh2 = Σ ( xi - xh )2 / ( n - 1 )  
sh2 = [ (50 - 70)2 + (55 - 70)2 + (60 - 70)2 + ... + (85 - 70)2 + (90 - 70)2 ] / 17 = 105.41

The within-stratum sample variance for girls is computed similarly. It is equal to 45.41.   
  
Using results from the above computations, we compute the standard error (SE):

SE = (1 / N) \* sqrt { Σ [ Nh2 \* ( 1 - nh/Nh ) \* sh2 / nh ] }   
SE = (1 / 20,000) \* sqrt { [ 100,000,000 \* ( 1 - 18/10,000 ) \* 105.41 / 18 ] + [ 100,000,000 \* ( 1 - 18/10,000 ) \* 45.41 / 18 ] }   
SE = (1 / 20,000) \* sqrt { 99,820,000 \* 105.41 / 18 ] + [ 99,820,000 \* 45.41 / 18 ] } = 1.45

Thus, the standard error of the sampling distribution of the mean is 1.45.

* + Find critical value. The critical value is a factor used to compute the margin of error. Based on the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem), we can assume that the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the mean is normally distributed. Therefore, we express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 95/100 = 0.05
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.975. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 1.96.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.96 \* 1.45 = 2.84
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 95% confidence interval is 72.16 to 77.84. And the margin of error is equal to 2.84. That is, we are 95% confident that the true population mean is in the range defined by 75 + 2.84.

# What is Cluster Sampling?

**Cluster sampling** refers to a sampling method that has the following properties.

* The population is divided into *N* groups, called **clusters**.
* The researcher randomly selects *n* clusters to include in the sample.
* The number of observations within each cluster *Mi* is known, and M = M1 + M2 + M3 + ... + MN-1 + MN.
* Each element of the population can be assigned to one, and only one, cluster.

This tutorial covers two types of cluster sampling methods.

* **One-stage sampling**. All of the elements within selected clusters are included in the sample.
* **Two-stage sampling**. A subset of elements within selected clusters are randomly selected for inclusion in the sample.

## Cluster Sampling: Advantages and Disadvantages

Assuming the sample size is constant across sampling methods, cluster sampling generally provides less [precision](http://stattrek.com/Help/Glossary.aspx?Target=Precision) than either [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple_random_sampling) or [stratified sampling](http://stattrek.com/Help/Glossary.aspx?Target=Stratified_sampling). This is the main disadvantage of cluster sampling.

Given this disadvantage, it is natural to ask: Why use cluster sampling? Sometimes, the cost per sample point is less for cluster sampling than for other sampling methods. Given a fixed budget, the researcher may be able to use a bigger sample with cluster sampling than with the other methods. When the increased sample size is sufficient to offset the loss in precision, cluster sampling may be the best choice.

How to Analyze Data from Cluster Samples

In this lesson, we describe how to analyze survey data when the sampling method is cluster sampling.

Notation

The following notation is helpful, when we talk about analyzing data from cluster samples.

* N = The number of [clusters](http://stattrek.com/Help/Glossary.aspx?Target=Cluster) in the population.
* Mi = The number of observations in the *i*th cluster.
* Xi = The population mean for the *i*th cluster
* M = The total number of observations in the population = Σ Mi.
* P = The population [proportion](http://stattrek.com/Help/Glossary.aspx?Target=Proportion).
* Pi = The population proportion for the *i*th cluster
* *n* = The number of clusters in the sample.
* mi = The number of sample observations from the *i*th cluster.
* xij = The measurement for the *j*th observation from the *i*th cluster.
* xi = The sample estimate of the population mean for the *i*th cluster = Σ ( xij / mi ) summed over *j*.
* p = The sample estimate of the population proportion.
* pi = The sample estimate of the population proportion for the *i*th cluster.
* s2i = The sample estimate of the population [variance](http://stattrek.com/Help/Glossary.aspx?Target=Variance) within cluster *i*.
* ti = The estimated total for the *i*th cluster = Σ ( Mi / mi ) \* xij = Mi \* xi .
* tmean = The sample estimate of the population total = ( N / n ) \* Σ ti .
* tprop(i) = The sample estimate of the number of successes in population i = Mi \* pi .
* tprop = The sample estimate of the number of successes in the population = ( N / n ) \* Σ tprop(i) .
* *SE*: The [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). (This is an estimate of the standard deviation of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution).)
* Σ = Summation symbol, used to compute sums over the sample. ( To illustrate its use, Σ xi = x1 + x2 + x3 + ... + xm-1 + xm )

How to Analyze Data From Cluster Samples

Different [sampling methods](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_method) use different formulas to estimate population [parameters](http://stattrek.com/Help/Glossary.aspx?Target=Parameter) and to estimate [standard errors](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). The formulas that we have used so far in this tutorial work for simple random samples and for stratified samples, but they are not right for cluster samples.

The next two sections of this lesson show the correct formulas to use with cluster samples. With these formulas, you can readily estimate population parameters and standard errors. And once you have the standard error, the procedures for computing other things (e.g., [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error), [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval), and [region of acceptance](http://stattrek.com/Help/Glossary.aspx?Target=Region_of_acceptance)) are largely the same for cluster samples as for simple random samples. The sample problem at the end of this lesson shows how to use these formulas to analyze data from cluster samples.

Measures of Central Tendency

The table below shows formulas that can be used with [one-stage](http://stattrek.com/Help/Glossary.aspx?Target=One_stage_sampling) and [two-stage](http://stattrek.com/Help/Glossary.aspx?Target=Two_stage_sampling) cluster samples to estimate a population mean and a population proportion.

|  |  |  |
| --- | --- | --- |
| **Population parameter** | **Sample estimate: One-stage** | **Sample estimate: Two-stage** |
| Mean | [ ( N / ( n \* M ) ] \* Σ ( Mi \* Xi ) | [ ( N / ( n \* M ) ] \* Σ ( Mi \* xi ) |
| Proportion | [ ( N / ( n \* M ) ] \* Σ ( Mi \* Pi ) | [ ( N / ( n \* M ) ] \* Σ ( Mi \* pi ) |

These formulas produce [unbiased estimates](http://stattrek.com/Help/Glossary.aspx?Target=Unbiased_estimator) of the population parameters.

The Variability of the Estimate

The [precision](http://stattrek.com/Help/Glossary.aspx?Target=Precision) of a [sample design](http://stattrek.com/Help/Glossary.aspx?Target=Sample_design) is directly related to the variability of the estimate, which is measured by the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error). The tables below show how to compute the standard error (SE), when the sampling method is cluster sampling.

The first table shows how to compute the standard error for a mean score, given one- or two-stage sampling.

|  |  |
| --- | --- |
| **Number of stages** | **Standard error of mean score** |
| One | ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* xi - tmean / N )2 / ( n - 1 ) } |
| Two | ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* xi - tmean / N )2 / ( n - 1 )  + ( N / n ) \* Σ [ ( 1 - mi / Mi ) \* Mi2 \* si2 / mi ] } |

The next table shows how to compute the standard error for a proportion. Like the previous table, this table shows equations for one- and two-stage designs. It also shows how the equations differ when the true population proportions are known versus when they are estimated based on sample data.

|  |  |  |
| --- | --- | --- |
| **Number of stages** | **Population proportion** | **Standard error of proportion** |
| One | Known | ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* Pi - tprop / N )2 / ( n - 1 ) } |
| One | Estimated | ( 1 / M ) \* sqrt { [ ( N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* pi - tprop / N )2 / ( n - 1 ) } |
| Two | Known | ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* Pi - tprop / N )2 } / ( n - 1 )  + ( N / n ) \* Σ [ ( 1 - mi / Mi ) \* Mi2 \* Pi \* ( 1 - Pi ) / mi ] } |
| Two | Estimated | ( 1 / M ) \* sqrt [ ( N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* pi - tprop / N )2 } / ( n - 1 )  + ( N / n ) \* Σ [ ( 1 - mi / Mi ) \* Mi2 \* pi \* ( 1 - pi ) / ( mi - 1 ) ] } |

Sample Problem

This section presents a sample problem that illustrates how to analyze survey data when the sampling method is one-stage cluster sampling. (In a [subsequent lesson](http://stattrek.com/Lesson6/SamplingMethod.aspx), we re-visit this problem and see how cluster sampling compares to other sampling methods.)

Sample Planning Wizard

The analysis of data collected via cluster sampling can be complex and time-consuming. Stat Trek's Sample Planning Wizard can help. The Wizard computes survey precision, sample size requirements, costs, etc., as well as estimates population parameters and tests hypotheses. It also creates a summary report that lists key findings and documents analytical techniques. Whenever you work with cluster sampling, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

**Example 1**

At the end of every school year, the state administers a reading test to a sample of third graders. The school system has 20,000 third graders, grouped in 1000 separate classes. Assume that each class has 20 students. This year, the test was administered to each student in 36 randomly-sampled classes. Thus, this is one-stage cluster sampling, with classes serving as clusters. The average test score from each sampled cluster Xi is shown below:

|  |
| --- |
| 55, 60, 65, 67, 67, 70, 70, 70, 72, 72, 72, 72, 73, 73, 75, 75, 75, 75,  75, 77, 77, 78, 78, 78, 78, 80, 80, 80, 80, 80, 80, 83, 83, 85, 85, 85 |

Using sample data, estimate the mean reading achievement level in the population. Find the [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error) and the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval). Assume a 95% [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level).

*Solution:* Previously we described [how to compute the confidence interval for a mean score](http://stattrek.com/Lesson4/Mean.aspx?Tutorial=stat). Below, we apply that process to the present cluster sampling problem.

* Identify a sample statistic. For this problem, we use the sample mean to estimate the population mean, and we use the equation from the "Measures of Central Tendency" table to compute the sample mean.

x = [ ( N / ( n \* M ) ] \* Σ ( Mi \* Xi )  
x = [ ( 1000 / ( 36 \* 20,000 ) ] \* Σ ( 20 \* Xi ) = Σ ( Xi ) / 36   
x = ( 55 + 60 + 65 + ... + 85 + 85 + 85 ) / 36 = 75

Therefore, based on data from the cluster sample, we estimate that the mean reading achievement level in the population is equal to 75.

* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 95% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx?Tutorial=stat) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard error of the sampling distribution. Since we used one-stage cluster sampling, the standard error is:

SE = ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* Xi - tmean / N )2 / ( n - 1 ) }   
where tmean = ( N / n ) \* Σ ( Mi \* Xi )

Except for tmean, all of the terms on the right side of the above equation are known. Therefore, to compute SE, we must first compute tmean. The formula for tmean is:

tmean = ( N / n ) \* Σ ti  
tmean = ( N / n ) \* ΣΣ [( Mi / mi ) \* xij ]   
tmean = ( 1000 / 36 ) \* ΣΣ [( 20 / 20 ) \* xij ]   
tmean = ( 27.778 ) \* ΣΣ ( xij ) = ( 27.778 ) \* 20 \* Σ ( Xi )   
tmean = ( 27.778 ) \* 20 \* ( 55 + 60 + 65 + ... + 85 + 85 + 85 ) = 1,500,000

After we compute tmean, all of the terms on the right side of the SE equation are known, so we plug the known values into the standard error equation. As shown below, the standard error is 1.1.

SE = ( 1 / M ) \* sqrt { [ N2 \* ( 1 - n/N ) / n ] \* Σ ( Mi \* Xi - tmean / N )2 / ( n - 1 ) }   
SE = ( 1 /20,000 ) \* sqrt { [ 10002 \* ( 1 - 36/1000 ) / 36 ] \* Σ ( 20 \* Xi - 1,500,000 / 1000 )2 / ( 35 ) }   
SE = ( 1 /20,000 ) \* sqrt { [ 10002 \* ( 1 - 36/1000 ) / 36 ] \*   
( 20 \* 55 - 1,500,000 / 1000 )2 / ( 35 ) + ( 20 \* 60 - 1,500,000 / 1000 )2 / ( 35 )   
+ ... +   
( 20 \* 85 - 1,500,000 / 1000 )2 / ( 35 ) + ( 20 \* 85 - 1,500,000 / 1000 )2 / ( 35 ) }   
SE = ( 1 /20,000 ) \* sqrt [ [ 10002 \* ( 1 - 36/1000 ) / 36 ] \* 18,217.143 ]   
SE = 1.1

* + Find critical value. The critical value is a factor used to compute the margin of error. Based on the [central limit theorem](http://stattrek.com/Help/Glossary.aspx?Target=Central%20limit%20theorem), we can assume that the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling%20distribution) of the mean is normally distributed. Therefore, we express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 95/100 = 0.05
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.975. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 1.96.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.96 \* 1.1 = 2.16
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 95% confidence interval is 72.84 to 77.16. And the margin of error is equal to 2.16. That is, we are 95% confident that the true population mean is in the range defined by 75 + 2.16.

## When to Use Cluster Sampling

Cluster sampling should be used only when it is economically justified - when reduced costs can be used to overcome losses in precision. This is most likely to occur in the following situations.

* Constructing a complete list of population elements is difficult, costly, or impossible. For example, it may not be possible to list all of the customers of a chain of hardware stores. However, it would be possible to randomly select a subset of stores (stage 1 of cluster sampling) and then interview a random sample of customers who visit those stores (stage 2 of cluster sampling).
* The population is concentrated in "natural" clusters (city blocks, schools, hospitals, etc.). For example, to conduct personal interviews of operating room nurses, it might make sense to randomly select a sample of hospitals (stage 1 of cluster sampling) and then interview all of the operating room nurses at that hospital. Using cluster sampling, the interviewer could conduct many interviews in a single day at a single hospital. Simple random sampling, in contrast, might require the interviewer to spend all day traveling to conduct a single interview at a single hospital.

Even when the above situations exist, it is often unclear which sampling method should be used. Test different options, using hypothetical data if necessary. Choose the most cost-effective approach; that is, choose the sampling method that delivers the greatest precision for the least cost.

## Sample Planning Wizard

The computations involved in testing different sample designs can be complex and time-consuming. Stat Trek's Sample Planning Wizard can help. The Wizard computes survey precision, sample size requirements, costs, etc., allowing you to compare alternative designs quickly, easily, and error-free. The Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you want to quickly find the most precise, cost-effective sample design, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

## The Difference Between Strata and Clusters

Although [strata](http://stattrek.com/Help/Glossary.aspx?Target=Strata) and clusters are both non-overlapping subsets of the population, they differ in several ways.

* All strata are represented in the sample; but only a subset of clusters are in the sample.
* With stratified sampling, the best survey results occur when elements within strata are internally [homogeneous](http://stattrek.com/Help/Glossary.aspx?Target=Homogeneous). However, with cluster sampling, the best results occur when elements within clusters are internally [heterogeneous](http://stattrek.com/Help/Glossary.aspx?Target=Heterogeneous).

Sample Size: Simple Random Samples

This lesson describes how to find the smallest sample size that provides the desired precision, when the sampling method is simple random sampling.

Factors That Influence Sample Size

The "right" sample size for a particular application depends on many factors, including the following:

* Cost considerations (e.g., maximum budget, desire to minimize cost).
* Administrative concerns (e.g., complexity of the design, research deadlines).
* Minimum acceptable level of precision.
* Confidence level.
* Variability within the population or subpopulation (e.g., stratum, cluster) of interest.
* Sampling method.

These factors interact in complex ways. Although a consideration of all the variations is beyond the scope of this tutorial, the remainder of this lesson covers a situation that commonly occurs with simple random samples: How to find the smallest sample size that provides the required precision.

The next lesson covers two situations that commonly occur with stratified random samples:

* [How to get the most precision from a stratified sample, given a fixed sample size](http://stattrek.com/sample-size/SampleSizeStrata.aspx#neyman).
* [How to get the most precision from a stratified sample, given a fixed budget](http://stattrek.com/sample-size/SampleSizeStrata.aspx#optimum).

For these, and for other situations not covered in the tutorial, consider using the Sample Planning Wizard (described below).

Sample Planning Wizard

Stat Trek's Sample Planning Wizard can help you find the right sample size quickly, easily, and accurately. You specify your main goal - maximize precision, minimize cost, stay within budget, etc. Based on your goal, the Wizard prompts you for the necessary inputs and handles all computations automatically, allowing you to compare alternative designs and sample sizes. The Wizard creates a summary report that lists key findings and describes analytical techniques. Whenever you want to quickly find the most precise, cost-effective sample design, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

How to Choose Sample Size for a Simple Random Sample

Consider the following problem. You are conducting a survey. The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple_random_sampling), [without replacement](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_without_replacement). You want your survey to provide a specified level of precision.

To choose the right sample size for a simple random sample, you need to define the following inputs.

* Specify the desired [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error) *ME*. This is your measure of precision.
* Specify [alpha](http://stattrek.com/Help/Glossary.aspx?Target=Alpha).
  + For a hypothesis test, alpha is the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level).
  + For an estimation problem, alpha is: 1 - [Confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20level).
* Find the critical [standard score](http://stattrek.com/Help/Glossary.aspx?Target=Standard_score) *z*.
  + For an [estimation problem](http://stattrek.com/Help/Glossary.aspx?Target=Estimation) or for a [two-tailed hypothesis test](http://stattrek.com/Help/Glossary.aspx?Target=Two_tailed_test), the critical standard score (z) is the value for which the cumulative probability is 1 - alpha/2.
  + For a [one-tailed hypothesis test](http://stattrek.com/Help/Glossary.aspx?Target=One_tailed_test), the critical standard score (z) is the value for which the cumulative probability is 1 - alpha.
* Unless the population size is very large, you need to specify the size of the population (N).

You will also need to know the variance of the population, σ2. Given these inputs, the following formulas find the smallest sample size that provides the desired level of precision.

|  |  |  |
| --- | --- | --- |
| **Sample statistic** | **Population size** | **Sample size** |
| Mean | Known | n = { z2 \* σ2 \* [ N / (N - 1) ] } / { ME2 + [ z2 \* σ2 / (N - 1) ] } |
| Mean | Unknown | n = ( z2 \* σ2 ) / ME2 |
| Proportion | Known | n = [ ( z2 \* p \* q ) + ME2 ] / [ ME2 + z2 \* p \* q / N ] |
| Proportion | Unknown | n = [ ( z2 \* p \* q ) + ME2 ] / ( ME2 ) |

This approach works when the sample size is relatively large (greater than or equal to 30). Use the first or third formulas when the population size is known. When the population size is large but unknown, use the second or fourth formulas.

For proportions, the sample size requirements vary, based on the value of the proportion. If you are unsure of the right value to use, set *p* equal to 0.5. This will produce a conservative sample size estimate; that is, the sample size will produce *at least* the precision called for and may produce better precision.

Sample Problem

At the end of every school year, the state administers a reading test to a simple random sample drawn without replacement from a population of 100,000 third graders. Over the last five years, students who took the test correctly answered 75% of the test questions.

What sample size should you use to achieve a margin of error equal to plus or minus 4%, with a confidence level of 95%?

*Solution:* To solve this problem, we follow the steps outlined above.

* Specify the margin of error. This was given in the problem definition. The margin of error is plus or minus 4% or 0.04.
* Specify the confidence level. This was also given. The confidence level is 95% or 0.95.
* Compute alpha. Alpha is equal to one minus the confidence level. Thus, alpha = 1 - 0.95 = 0.05.
* Determine the critical standard score (z). Since this is an [estimation problem](http://stattrek.com/Help/Glossary.aspx?Target=Estimation), the critical standard score is the value for which the cumulative probability is 1 - alpha/2 = 1 - 0.05/2 = 0.975.   
    
  To find that value, we use the [Normal Calculator](http://stattrek.com/Tables/Normal.aspx). Recall that the distribution of standard scores has a mean of 0 and a standard deviation of 1. Therefore, we plug the following entries into the normal calculator: Value = 0.975; Mean = 0; and Standard deviation = 1. The calulator tells us that the value of the standard score is 1.96.
* And finally, we assume that the population proportion *p* is equal to its past value over the previous 5 years. That value is 0.75. Given these inputs, we can find the smallest sample size *n* that will provide the required margin of error.

n = [ (z2 \* p \* q ) + ME2 ] / [ ME2 + z2 \* p \* q / N ]   
n = [ (1.96)2 \* 0.75 \* 0.25 + 0.0016] / [ 0.0016 + (1.96)2 \* 0.75 \* 0.25 / 100,000 ]  
n = (0.7203 + 0.0016) / ( 0.0016 + 0.0000072) = 449.2

Therefore, to achieve a margin of error of plus or minus 4 percent, we will need to survey 450 students, using simple random sampling.

Sample Size: Stratified Random Samples

The precision and cost of a stratified design is influenced by the way that sample elements are allocated to strata.

How to Assign Sample to Strata

One approach is [proportionate stratification](http://stattrek.com/Help/Glossary.aspx?Target=Proportionate_stratification). With proportionate stratification, the sample size of each stratum is proportionate to the population size of the stratum. Strata sample sizes are determined by the following equation :

nh = ( Nh / N ) \* n

where nh is the sample size for stratum *h*, Nh is the population size for stratum *h*, N is total population size, and n is total sample size.

Another approach is [disproportionate stratification](http://stattrek.com/Help/Glossary.aspx?Target=Disproportionate_stratification), which can be a better choice (e.g., less cost, more precision) if sample elements are assigned correctly to strata. To take advantage of disproportionate stratification, researchers need to answer such questions as:

* Given a fixed budget, how should sample be allocated to get the most precision from a stratified sample?
* Given a fixed sample size, how should sample be allocated to get the most precision from a stratified sample?
* Given a fixed budget, what is the most precision that I can get from a stratified sample?
* Given a fixed sample size, what is the most precision that I can get from a stratified sample?
* What is the smallest sample size that will provide a given level of survey precision?
* What is the minimum cost to achieve a given level of survey precision?
* Given a particular sample allocation plan, what level of precision can I expect?
* And so on.

Although a consideration of all these questions is beyond the scope of this tutorial, the remainder of this lesson does address the first two questions. (To answer the other questions, as well as the first two questions, consider using the [Sample Planning Wizard](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx).)

Sample Planning Wizard

Stat Trek's Sample Planning Wizard can help you find the right sample allocation plan for your stratified design. You specify your main goal - maximize precision, minimize cost, stay within budget, etc. Based on your goal, the Wizard prompts you for the necessary inputs and handles all computations automatically. It tells you the best sample size for each stratum. The Wizard creates a summary report that lists key findings, including the margin of error. And it describes analytical techniques. Whenever you want to quickly find the best sample allocation plan, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

How to Maximize Precision, Given a Stratified Sample With a Fixed Budget

The ideal sample allocation plan would provide the most precision for the least cost. **Optimal allocation** does just that. Based on optimal allocation, the best sample size for stratum *h* would be:

nh = n \* [ ( Nh \* σh ) / sqrt( ch ) ] / [ Σ ( Ni \* σi ) / sqrt( ci ) ]

where nh is the sample size for stratum *h*, n is total sample size, Nh is the population size for stratum *h*, σh is the standard deviation of stratum *h*, and ch is the direct cost to sample an individual element from stratum *h*. Note that ch does not include indirect costs, such as overhead costs.

The effect of the above equation is to sample more heavily from a stratum when

* The cost to sample an element from the stratum is low.
* The population size of the stratum is large.
* The variability within the stratum is large.

How to Maximize Precision, Given a Stratified Sample With a Fixed Sample Size

Sometimes, researchers want to find the sample allocation plan that provides the most precision, given a fixed sample size. The solution to this problem is a special case of optimal allocation, called **Neyman allocation**.

The equation for Neyman allocation can be derived from the equation for optimal allocation by assuming that the direct cost to sample an individual element is equal across strata. Based on Neyman allocation, the best sample size for stratum *h* would be:

nh = n \* ( Nh \* σh ) / [ Σ ( Ni \* σi ) ]

where nh is the sample size for stratum *h*, n is total sample size, Nh is the population size for stratum *h*, and σh is the standard deviation of stratum *h*.

Sample Problem

This section presents a sample problem that illustrates how to maximize precision, given a fixed sample size and a stratified sample. (In a [subsequent lesson](http://stattrek.com/Lesson6/SamplingMethod.aspx?Tutorial=stat), we re-visit this problem and see how stratified sampling compares to other sampling methods.)

**Problem 1**

At the end of every school year, the state administers a reading test to a sample of 36 third graders. The school system has 20,000 third graders, half boys and half girls. The results from last year's test are shown in the table below.

|  |  |  |
| --- | --- | --- |
| **Stratum** | **Mean score** | **Standard deviation** |
| Boys | 70 | 10.27 |
| Girls | 80 | 6.66 |

This year, the researchers plan to use a stratified sample, with one stratum consisting of boys and the other, girls. Use the results from last year to answer the following questions?

* To maximize precision, how many sampled students should be boys and how many should be girls?
* What is the mean reading achievement level in the population?
* Compute the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval).
* Find the [margin of error](http://stattrek.com/Help/Glossary.aspx?Target=Margin_of_error)

Assume a 95% [confidence level](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_level).

*Solution:* The first step is to decide how to allocate sample in order to maximize precision. Based on Neyman allocation, the best sample size for stratum *h* is:

nh = n \* ( Nh \* σh ) / [ Σ ( Ni \* σi ) ]

where nh is the sample size for stratum *h*, n is total sample size, Nh is the population size for stratum *h*, and σh is the standard deviation of stratum *h*. By this equation, the number of boys in the sample is:

nboys = 36 \* ( 10,000 \* 10.27 ) / [ ( 10,000 \* 10.27 ) + ( 10,000 \* 6.67 ) ] = 21.83

Therefore, to maximize precision, the total sample of 36 students should consist of 22 boys and (36 - 22) = 14 girls.

The remaining questions can be answered during the process of computing the [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence_interval). Previously, we described [how to compute a confidence interval](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx?Tutorial=stat). We employ that process below.

* Identify a sample statistic. For this problem, we use the overall sample mean to estimate the population mean. To compute the overall sample mean, we use the following equation (which was introduced in a [previous lesson](http://stattrek.com/Lesson6/STRAnalysis.aspx?Tutorial=stat)):

x = Σ ( Nh / N ) \* xh = ( 10,000/20,000 ) \* 70 + ( 10,000/20,000 ) \* 80 = 75

Therefore, based on data from the sample strata, we estimate that the mean reading achievement level in the population is equal to 75.

* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 95% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx?Tutorial=stat) when the sampling distribution is approximately normal. The key steps are shown below.
  + Find standard deviation or standard error. The equation to compute the standard error was introduced in a [previous lesson](http://stattrek.com/Lesson6/STRAnalysis.aspx?Tutorial=stat). We use that equation here:

SE = (1 / N) \* sqrt { Σ [ Nh2 \* ( 1 - nh/Nh ) \* sh2 / nh ] }   
SE = (1 / 20,000) \* sqrt { [ 10,0002 \* ( 1 - 22/10,000 ) \* (10.27)2 / 22 ] + [ 10,0002 \* ( 1 - 14/10,000 ) \* (6.66)2 / 14 ] } = 1.41

Thus, the standard deviation of the sampling distribution (i.e., the standard error) is 1.41.

* + Find critical value. The critical value is a factor used to compute the margin of error. We express the critical value as a [z score](http://stattrek.com/Help/Glossary.aspx?Target=z%20score). To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 99/100 = 0.01
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.05/2 = 0.975
    - The critical value is the z score having a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.975. From the [Normal Distribution Calculator](http://stattrek.com/Tables/Normal.aspx), we find that the critical value is 1.96.
  + Compute margin of error (ME): ME = critical value \* standard error = 1.96 \* 1.41 = 2.76
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level. Thus, with this sample design, we are 95% confident that the sample estimate of reading achievement is 75 + 2.76.

In summary, given a total sample size of 36 students, we can get the greatest precision from a stratified sample if we sample 22 boys and 14 girls. This results in a 95% confidence interval of 72.24 to 77.76. The margin of error is 2.76.

How to Choose the Best Sampling Method

The best [sampling method](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_method) is the sampling method that most effectively meets the particular goals of the study in question. The effectiveness of a sampling method depends on many factors. Because these factors interact in complex ways, the "best" sampling method is seldom obvious. Good researchers use the following strategy to identify the best sampling method.

* List the research goals (usually some combination of [accuracy](http://stattrek.com/Help/Glossary.aspx?Target=Accuracy), [precision](http://stattrek.com/Help/Glossary.aspx?Target=Precision), and/or cost).
* Identify potential sampling methods that *might* effectively achieve those goals.
* Test the ability of each method to achieve each goal.
* Choose the method that does the best job of achieving the goals.

The next section presents an example that illustrates this strategy.

Sample Planning Wizard

The computations involved in comparing different sampling methods can be complex and time-consuming. Stat Trek's Sample Planning Wizard can help. The Wizard computes survey precision, sample size requirements, costs, etc., allowing you to compare alternative sampling methods quickly, easily, and accurately. The Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you want to find the most precise, cost-effective sampling method, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. **>** [Learn more](http://stattrek.com/SPWizard/SamplePlanningWizardDescription.aspx)

How to Choose the Best Sampling Method

In this section, we illustrate how to choose the best sampling method by working through a sample problem. Here is the problem:

|  |
| --- |
| **Problem Statement** |
| At the end of every school year, the state administers a reading test to a sample of third graders. The school system has 20,000 third graders, half boys and half girls. There are 1000 third-grade classes, each with 20 students.  The maximum budget for this research is $3600. The only expense is the cost to proctor each test session. This amounts to $100 per session.  The purpose of the study is to estimate the reading proficiency of third graders, based on sample data. School administrators want to maximize the precision of this estimate without exceeding the $3600 budget. What sampling method should they use? |

As noted earlier, finding the "best" sampling method is a four-step process. We work through each step below.

* List goals. This study has two main goals: (1) maximize precision and (2) stay within budget.
* Identify potential sampling methods. This tutorial has covered three basic sampling methods - simple random sampling, stratified sampling, and cluster sampling. In addition, we've described some variations on the basic methods (e.g., proportionate vs. disproportionate stratification, [one-stage](http://stattrek.com/Help/Glossary.aspx?Target=One_stage_sampling) vs. [two-stage cluster sampling](http://stattrek.com/Help/Glossary.aspx?Target=Two_stage_sampling), sampling with replacement versus sampling without replacement).

Because one of the main goals is to maximize precision, we can eliminate some of these alternatives. Sampling without replacement always provides equal or better precision than sampling with replacement, so we will focus only on sampling without replacement. Also, as long as the same clusters are sampled, one-stage cluster sampling always provides equal or better precision than two-stage cluster sampling, so we will focus only on one-stage cluster sampling. (Note: For cluster sampling in this example, the cost is the same whether we sample all students or only some students from a particular cluster; so in this example, two-stage sampling offers no cost advantage over one-stage sampling.)

This leaves us with four potential sampling methods - simple random sampling, proportionate stratified sampling, disproportionate stratified sampling, and one-stage cluster sampling. Each of these uses sampling without replacement. Because of the need to maximize precision, we will use [Neyman allocation](http://stattrek.com/Help/Glossary.aspx?Target=Neyman_allocation) with our disproportionate stratified sample.

* Test methods. A key part of the analysis is to test the ability of each potential sampling method to satisfy the research goals. Specifically, we will want to know the level of precision and the cost associated with each potential method. For our test, we use the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard_error) to measure precision. The smaller the standard error, the greater the precision.

To avoid getting bogged down in the computational details of the analysis, we will use results from sample problems that have appeared in previous lessons. Those results are summarized in the table below. (To review the analyses that produced this output, click the "See analysis" links in the last column of the table.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sampling method** | **Cost** | **Standard error** | **Sample size** | **Analytical details** |
| Simple random sampling | $3600 | 1.66 | 36 | [See analysis](http://stattrek.com/Lesson6/SRS.aspx?Tutorial=stat#example1) |
| Proportionate stratified sampling | $3600 | 1.45 | 36 | [See analysis](http://stattrek.com/Lesson6/STRAnalysis.aspx?Tutorial=stat#example1) |
| Disproportionate stratified sampling | $3600 | 1.41 | 36 | [See analysis](http://stattrek.com/Lesson6/SampleSizeStrata.aspx?Tutorial=stat#example1) |
| One-stage cluster sampling | $3600 | 1.10 | 720 | [See analysis](http://stattrek.com/Lesson6/CLSAnalysis.aspx?Tutorial=stat#example1) |

Because the budget is $3600 and because each test session costs $100 (for the proctor), there can be at most 36 test sessions. For the first three methods, students in the sample might come from 36 different schools, which would mean that each test session could have only one student. Thus, for simple random sampling and stratified sampling, the sample size might be only 36 students. For cluster sampling, in contrast, each of the 36 test sessions will have a full class of 20 students; so the sample size will be 36 \* 20 = 720 students.

* Choose best method. In this example, the cost of each sampling method is identical, so none of the methods has an advantage on cost. However, the methods do differ with respect to precision (as measured by standard error). Cluster sampling provides the most precision (i.e., the smallest standard error); so cluster sampling is the best method.

Although cluster sampling was "best" in this example, it may not be the best solution in other situations. Other sampling methods may be best in other situations. Use the four-step process described above to determine which method is best in any situation.

Scales of Measurement in Statistics

Measurement scales are used to categorize and/or quantify variables. This lesson describes the four scales of measurement that are commonly used in statistical analysis: nominal, ordinal, interval, and ratio scales.

Properties of Measurement Scales

Each scale of measurement satisfies one or more of the following properties of measurement.

* **Identity**. Each value on the measurement scale has a unique meaning.
* **Magnitude**. Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
* **Equal intervals**. Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
* **A minimum value of zero**. The scale has a true zero point, below which no values exist.

Nominal Scale of Measurement

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude.

Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

Ordinal Scale of Measurement

The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale.

An example of an ordinal scale in action would be the results of a horse race, reported as "win", "place", and "show". We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a mile.

Interval Scale of Measurement

The interval scale of measurement has the properties of identity, magnitude, and equal intervals.

A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.

With an interval scale, you know not only whether different values are bigger or smaller, you also know *how much* bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Ratio Scale of Measurement

The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and a minimum value of zero.

The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and the scale has a minimum value of zero.

Weight scales have a minimum value of zero because objects at rest can be weightless, but they cannot have negative weight.

Test Your Understanding

**Problem 1**

Consider the centigrade scale for measuring temperature. Which of the following measurement properties is satisfied by the centigrade scale?

I. Magnitude.   
II. Equal intervals.   
III. A minimum value of zero.

(A) I only   
(B) II only   
(C) III only   
(D) I and II   
(E) II and III

**Solution**

The correct answer is (D). The centigrade scale has the magnitude property because each value on the scale can be ranked as larger or smaller than any other value. And it has the equal intervals property because the scale is made up of equal units.

However, the centigrade scale does not have a minimum value of zero. Water freezes at zero degrees centigrade, but temperatures get colder than that. In the arctic, temperatures are almost always below zero.

Correlation Coefficient

**Correlation coefficients** measure the strength of association between two variables. The most common correlation coefficient, called the **Pearson product-moment correlation coefficient**, measures the strength of the *linear association* between variables.

In this tutorial, when we speak simply of a correlation coefficient, we are referring to the Pearson product-moment correlation. Generally, the correlation coefficient of a [sample](http://stattrek.com/Help/Glossary.aspx?Target=Sample) is denoted by *r*, and the correlation coefficient of a [population](http://stattrek.com/Help/Glossary.aspx?Target=Population) is denoted by ρ or *R*.

How to Interpret a Correlation Coefficient

The sign and the [absolute value](http://stattrek.com/Help/Glossary.aspx?Target=Absolute_value) of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.

* The value of a correlation coefficient ranges between -1 and 1.
* The greater the absolute value of a correlation coefficient, the stronger the *linear* relationship.
* The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
* The weakest linear relationship is indicated by a correlation coefficient equal to 0.
* A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
* A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

Keep in mind that the Pearson product-moment correlation coefficient only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero *linear* relationship. (It is possible for two variables to have zero linear relationship and a strong curvilinear relationship at the same time.)

Scatterplots and Correlation Coefficients

The [scatterplots](http://stattrek.com/Help/Glossary.aspx?Target=Scatterplot) below show how different patterns of data produce different degrees of correlation.

|  |  |  |
| --- | --- | --- |
| http://stattrek.com/Images/Sp9.jpg | http://stattrek.com/Images/Sp10.jpg | http://stattrek.com/Images/Sp11.jpg |
| **Maximum positive correlation (r = 1.0)** | **Strong positive correlation (r = 0.80)** | **Zero correlation (r = 0)** |
| http://stattrek.com/Images/Sp12.jpg | http://stattrek.com/Images/Sp13.jpg | http://stattrek.com/Images/Sp14.jpg |
| **Maximum negative correlation (r = -1.0)** | **Moderate negative correlation (r = -0.43)** | **Strong correlation & outlier (r = 0.71)** |

Several points are evident from the scatterplots.

* When the [slope](http://stattrek.com/Help/Glossary.aspx?Target=Slope) of the line in the plot is negative, the correlation is negative; and vice versa.
* The strongest correlations (r = 1.0 and r = -1.0 ) occur when data points fall *exactly* on a straight line.
* The correlation becomes weaker as the data points become more scattered.
* If the data points fall in a random pattern, the correlation is equal to zero.
* Correlation is affected by [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier). Compare the first scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

How to Calculate a Correlation Coefficient

If you look in different statistics textbooks, you are likely to find different-looking (but equivalent) formulas for computing a correlation coefficient. In this section, we present several formulas that you may encounter.

The most common formula for computing a product-moment correlation coefficient (r) is given below.

**Product-moment correlation coefficient.** The correlation r between two variables is:

r = Σ (xy) / sqrt [ ( Σ x2 ) \* ( Σ y2 ) ]

where Σ is the summation symbol, x = xi - x, xi is the x value for observation i, x is the mean x value, y = yi - y, yi is the y value for observation i, and y is the mean y value.

The formula below uses population means and population standard deviations to compute a population correlation coefficient (ρ) from population data.

**Population correlation coefficient.** The correlation ρ between two variables is:

ρ = [ 1 / N ] \* Σ { [ (Xi - μX) / σx ] \* [ (Yi - μY) / σy ] }

where N is the number of observations in the population, Σ is the summation symbol, Xi is the X value for observation i, μX is the population mean for variable X, Yi is the Y value for observation i, μY is the population mean for variable Y, σx is the population standard deviation of X, and σy is the population standard deviation of Y.

The formula below uses sample means and sample standard deviations to compute a correlation coefficient (r) from sample data.

**Sample correlation coefficient.** The correlation r between two variables is:

r = [ 1 / (n - 1) ] \* Σ { [ (xi - x) / sx ] \* [ (yi - y) / sy ] }

where n is the number of observations in the sample, Σ is the summation symbol, xi is the x value for observation i, x is the sample mean of x, yi is the y value for observation i, y is the sample mean of y, sx is the sample standard deviation of x, and sy is the sample standard deviation of y.

Each of the latter two formulas can be derived from the first formula. Use the first or second formula when you have data from the entire population. Use the third formula when you only have sample data, but want to estimate the correlation in the population. When in doubt, use the first formula.

Fortunately, you will rarely have to compute a correlation coefficient by hand. Many software packages (e.g., Excel) and most [graphing calculators](http://stattrek.com/AP/Calculator.aspx) have a correlation function that will do the job for you.

Test Your Understanding

**Problem 1**

A national consumer magazine reported the following correlations.

* The correlation between car weight and car reliability is -0.30.
* The correlation between car weight and annual maintenance cost is 0.20.

Which of the following statements are true?

I. Heavier cars tend to be less reliable.   
II. Heavier cars tend to cost more to maintain.   
III. Car weight is related more strongly to reliability than to maintenance cost.

(A) I only   
(B) II only   
(C) III only   
(D) I and II only   
(E) I, II, and III

**Solution**

The correct answer is (E). The correlation between car weight and reliability is negative. This means that reliability tends to decrease as car weight increases. The correlation between car weight and maintenance cost is positive. This means that maintenance costs tend to increase as car weight increases.

The strength of a relationship between two variables is indicated by the [absolute value](http://stattrek.com/Help/Glossary.aspx?Target=Absolute_value) of the correlation coefficient. The correlation between car weight and reliability has an absolute value of 0.30. The correlation between car weight and maintenance cost has an absolute value of 0.20. Therefore, the relationship between car weight and reliability is stronger than the relationship between car weight and maintenance cost.

What is Linear Regression?

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

In this tutorial, we focus on the case where there is only one independent variable. This is called simple regression (as opposed to multiple regression, which handles two or more independent variables).

Tip: The next lesson presents a [simple regression example](http://stattrek.com/regression/Regression-Example.aspx?Tutorial=stat) that shows how to apply the material covered in this lesson. Since this lesson is a little dense, you may benefit by also reading the next lesson.

Prerequisites for Regression

Simple linear regression is appropriate when the following conditions are satisfied.

* The dependent variable *Y* has a linear relationship to the independent variable *X*. To check this, make sure that the XY [scatterplot](http://stattrek.com/Help/Glossary.aspx?Target=Scatterplot) is linear and that the [residual plot](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot) shows a random pattern. (Don't worry. We'll cover residual plots in a [future lesson](http://stattrek.com/regression/residual-analysis.aspx).)
* For each value of X, the probability distribution of Y has the same standard deviation σ. When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.
* For any given value of X,
  + The Y values are independent, as indicated by a random pattern on the residual plot.
  + The Y values are roughly normally distributed (i.e., [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry) and [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution)). A little [skewness](http://stattrek.com/Help/Glossary.aspx?Target=Skewness) is ok if the sample size is large. A [histogram](http://stattrek.com/Help/Glossary.aspx?Target=Histogram) or a [dotplot](http://stattrek.com/Help/Glossary.aspx?Target=Dotplot) will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a [bivariate](http://stattrek.com/Help/Glossary.aspx?Target=Bivariate%20data) data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

Y = Β0 + Β1X

where Β0 is a constant, Β1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

ŷ = b0 + b1x

where b0 is a constant, b1 is the regression coefficient, x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a [graphing calculator](http://stattrek.com/AP/Calculator.aspx) - to find b0 and b1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b0 and b1 "by hand". Here are the equations.

b1 = Σ [ (xi - x)(yi - y) ] / Σ [ (xi - x)2]   
b1 = r \* (sy / sx)   
b0 = y - b1 \* x

where b0 is the constant in the regression equation, b1 is the regression coefficient, r is the correlation between x and y, xi is the *X* value of observation *i*, yi is the *Y* value of observation *i*, x is the mean of *X*, y is the mean of *Y*, sx is the standard deviation of *X*, and sy is the standard deviation of *Y*

Properties of the Regression Line

When the regression parameters (b0 and b1) are defined as described above, the regression line has the following properties.

* The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).
* The regression line passes through the mean of the *X* values (x) and through the mean of the *Y* values (y).
* The regression constant (b0) is equal to the [y intercept](http://stattrek.com/Help/Glossary.aspx?Target=Y%20intercept) of the regression line.
* The regression coefficient (b1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the [slope](http://stattrek.com/Help/Glossary.aspx?Target=Slope) of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The Coefficient of Determination

The **coefficient of determination** (denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

* The coefficient of determination ranges from 0 to 1.
* An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
* An R2 of 1 means the dependent variable can be predicted without error from the independent variable.
* An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

**Coefficient of determination.** The coefficient of determination (R2) for a linear regression model with one independent variable is:

R2 = { ( 1 / N ) \* Σ [ (xi - x) \* (yi - y) ] / (σx \* σy ) }2

where N is the number of observations used to fit the model, Σ is the summation symbol, xi is the x value for observation i, x is the mean x value, yi is the y value for observation i, y is the mean y value, σx is the standard deviation of x, and σy is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R2) is easily computed using the following formula: R2 = r2.

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Test Your Understanding

**Problem 1**

A researcher uses a regression equation to predict home heating bills (dollar cost), based on home size (square feet). The correlation between predicted bills and home size is 0.70. What is the correct interpretation of this finding?

(A) 70% of the variability in home heating bills can be explained by home size.   
(B) 49% of the variability in home heating bills can be explained by home size.   
(C) For each added square foot of home size, heating bills increased by 70 cents.   
(D) For each added square foot of home size, heating bills increased by 49 cents.   
(E) None of the above.

**Solution**

The correct answer is (B). The coefficient of determination measures the proportion of variation in the dependent variable that is predictable from the independent variable. The coefficient of determination is equal to R2; in this case, (0.70)2 or 0.49. Therefore, 49% of the variability in heating bills can be explained by home size.

Simple Linear Regression Example

In this lesson, we apply regression analysis to some fictitious data, and we show how to interpret the results of our analysis.

**Note:** Regression computations are usually handled by a software package or a [graphing calculator](http://stattrek.com/AP/Calculator.aspx). For this example, however, we will do the computations "manually", since the gory details have educational value.

Problem Statement

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

* What linear regression equation best predicts statistics performance, based on math aptitude scores?
* If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
* How well does the regression equation fit the data?

How to Find the Regression Equation

In the table below, the xi column shows scores on the aptitude test. Similarly, the yi column shows statistics grades. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | Student | xi | yi | (xi - x) | (yi - y) | (xi - x)2 | (yi - y)2 | (xi - x)(yi - y) | |  | 1 | 95 | 85 | 17 | 8 | 289 | 64 | 136 | |  | 2 | 85 | 95 | 7 | 18 | 49 | 324 | 126 | |  | 3 | 80 | 70 | 2 | -7 | 4 | 49 | -14 | |  | 4 | 70 | 65 | -8 | -12 | 64 | 144 | 96 | |  | 5 | 60 | 70 | -18 | -7 | 324 | 49 | 126 | | **Sum** |  | 390 | 385 |  |  | 730 | 630 | 470 | | **Mean** |  | 78 | 77 |  |  |  |  |  | |

The regression equation is a linear equation of the form: ŷ = b0 + b1x . To conduct a regression analysis, we need to solve for b0 and b1. Computations are shown below.

|  |  |  |
| --- | --- | --- |
| b1 = Σ [ (xi - x)(yi - y) ] / Σ [ (xi - x)2]  b1 = 470/730 = 0.644 |  | b0 = y - b1 \* x b0 = 77 - (0.644)(78) = 26.768 |

Therefore, the regression equation is: ŷ = 26.768 + 0.644x .

How to Use the Regression Equation

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (*x*), perform the computation, and you have an estimated value (ŷ) for the dependent variable.

In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade would be:

ŷ = 26.768 + 0.644x = 26.768 + 0.644 \* 80 = 26.768 + 51.52 = 78.288

**Warning:** When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimates.

In this example, the aptitude test scores used to create the regression equation ranged from 60 to 95. Therefore, only use values inside that range to estimate statistics grades. Using values outside that range (less than 60 or greater than 95) is problematic.

How to Find the Coefficient of Determination

Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the [coefficient of determination](http://stattrek.com/Help/Glossary.aspx?Target=Coefficient%20of%20determination), which can be computed from the following formula.

R2 = { ( 1 / N ) \* Σ [ (xi - x) \* (yi - y) ] / (σx \* σy ) }2

where N is the number of observations used to fit the model, Σ is the summation symbol, xi is the x value for observation i, x is the mean x value, yi is the y value for observation i, y is the mean y value, σx is the standard deviation of x, and σy is the standard deviation of y. Computations for the sample problem of this lesson are shown below.

|  |  |  |
| --- | --- | --- |
| σx = sqrt [ Σ ( xi - x )2 / N ]  σx = sqrt( 730/5 ) = sqrt(146) = 12.083 |  | σy = sqrt [ Σ ( yi - y )2 / N ]  σy = sqrt( 630/5 ) = sqrt(126) = 11.225 |
| R2 = { ( 1 / N ) \* Σ [ (xi - x) \* (yi - y) ] / (σx \* σy ) }2 R2 = [ ( 1/5 ) \* 470 / ( 12.083 \* 11.225 ) ]2 = ( 94 / 135.632 )2 = ( 0.693 )2 = 0.48 | | |

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the [dependent variable](http://stattrek.com/Help/Glossary.aspx?Target=Dependent%20variable)) can be explained by the relationship to math aptitude scores (the [independent variable](http://stattrek.com/Help/Glossary.aspx?Target=Independent%20variable)). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Residual Analysis in Regression

Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.

Residuals

The difference between the observed value of the dependent variable (*y*) and the predicted value (*ŷ*) is called the **residual** (*e*). Each data point has one residual.

Residual = Observed value - Predicted value   
*e* = *y* - *ŷ*

Both the sum and the mean of the residuals are equal to zero. That is, Σ *e* = 0 and e = 0.

Residual Plots

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | x | 60 | 70 | 80 | 85 | 95 | | y | 70 | 65 | 70 | 95 | 85 | | ŷ | 65.411 | 71.849 | 78.288 | 81.507 | 87.945 | | e | 4.589 | -6.849 | -8.288 | 13.493 | -2.945 | |  | Image of residual plot |

The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

|  |  |  |
| --- | --- | --- |
| http://stattrek.com/Images/Sp16.jpg | http://stattrek.com/Images/Sp17.jpg | http://stattrek.com/Images/Sp18.jpg |
| **Random pattern** | **Non-random: U-shaped** | **Non-random: Inverted U** |

In the [next lesson](http://stattrek.com/AP-Statistics-1/Transformation.aspx?Tutorial=stat), we will work on a problem, where the residual plot shows a non-random pattern. And we will show how to "transform" the data to use a linear model with nonlinear data.

Test Your Understanding

In the context of [regression analysis](http://stattrek.com/Help/Glossary.aspx?Target=Regression), which of the following statements are true?

I. When the sum of the residuals is greater than zero, the data set is nonlinear.   
II. A random pattern of residuals supports a linear model.   
III. A random pattern of residuals supports a non-linear model.

(A) I only   
(B) II only   
(C) III only   
(D) I and II   
(E) I and III

**Solution**

The correct answer is (B). A random pattern of residuals supports a linear model; a non-random pattern supports a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.

Transformations to Achieve Linearity

When a [residual plot](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot) reveals a data set to be nonlinear, it is often possible to "transform" the raw data to make it more linear. This allows us to use [linear regression](http://stattrek.com/Help/Glossary.aspx?Target=Regression) techniques more effectively with nonlinear data.

What is a Transformation to Achieve Linearity?

Transforming a variable involves using a mathematical operation to change its measurement scale. Broadly speaking, there are two kinds of transformations.

* **Linear transformation.** A linear transformation preserves linear relationships between variables. Therefore, the [correlation](http://stattrek.com/Help/Glossary.aspx?Target=Correlation) between *x* and *y* would be unchanged after a linear transformation. Examples of a linear transformation to variable *x* would be multiplying *x* by a constant, dividing *x* by a constant, or adding a constant to *x*.
* **Nonlinear tranformation.** A nonlinear transformation changes (increases or decreases) linear relationships between variables and, thus, changes the correlation between variables. Examples of a nonlinear transformation of variable *x* would be taking the square root of *x* or the reciprocal of *x*.

In regression, a transformation to achieve linearity is a special kind of nonlinear transformation. It is a nonlinear transformation that *increases* the linear relationship between two variables.

Methods of Transforming Variables to Achieve Linearity

There are many ways to transform variables to achieve linearity for regression analysis. Some common methods are summarized below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Transformation(s)** | **Regression equation** | **Predicted value (ŷ)** |
| Standard linear regression | None | y = b0 + b1x | ŷ = b0 + b1x |
| Exponential model | Dependent variable = log(y) | log(y) = b0 + b1x | ŷ = 10b0 + b1x |
| Quadratic model | Dependent variable = sqrt(y) | sqrt(y) = b0 + b1x | ŷ = ( b0 + b1x )2 |
| Reciprocal model | Dependent variable = 1/y | 1/y = b0 + b1x | ŷ = 1 / ( b0 + b1x ) |
| Logarithmic model | Independent variable = log(x) | y= b0 + b1log(x) | ŷ = b0 + b1log(x) |
| Power model | Dependent variable = log(y)  Independent variable = log(x) | log(y)= b0 + b1log(x) | ŷ = 10b0 + b1log(x) |

Each row shows a different nonlinear transformation method. The second column shows the specific transformation applied to dependent and/or independent variables. The third column shows the regression equation used in the analysis. And the last column shows the "back transformation" equation used to restore the dependent variable to its original, non-transformed measurement scale.

In practice, these methods need to be tested on the data to which they are applied to be sure that they *increase* rather than *decrease* the linearity of the relationship. Testing the effect of a transformation method involves looking at [residual plots](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot) and correlation coefficients, as described in the following sections.

**Note:** The logarithmic model and the power model require the ability to work with [logarithms](http://stattrek.com/Help/Glossary.aspx?Target=Logarithm). Use a [graphic calculator](http://stattrek.com/AP/Calculator.aspx) to obtain the log of a number or to transform back from the logarithm to the original number. If you need it, the Stat Trek glossary has a brief [refresher on logarithms](http://stattrek.com/Help/Glossary.aspx?Target=Logarithm).

How to Perform a Transformation to Achieve Linearity

Transforming a data set to enhance linearity is a multi-step, trial-and-error process.

* Conduct a standard regression analysis on the raw data.
* Construct a residual plot.
  + If the plot pattern is random, do not transform data.
  + If the plot pattern is not random, continue.
* Compute the [coefficient of determination](http://stattrek.com/Help/Glossary.aspx?Target=coefficient%20of%20determination) (R2).
* Choose a transformation method (see above table).
* Transform the independent variable, dependent variable, or both.
* Conduct a regression analysis, using the transformed variables.
* Compute the coefficient of determination (R2), based on the transformed variables.
  + If the tranformed R2 is greater than the raw-score R2, the transformation was successful. Congratulations!
  + If not, try a different transformation method.

The best tranformation method (exponential model, quadratic model, reciprocal model, etc.) will depend on nature of the original data. The only way to determine which method is best is to try each and compare the result (i.e., [residual plots](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot), correlation coefficients).

A Transformation Example

Below, the table on the left shows data for independent and dependent variables - x and y, respectively. When we apply a linear regression to the untransformed raw data, the [residual plot](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot) shows a non-random pattern (a U-shaped curve), which suggests that the data are nonlinear.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | y | 2 | 1 | 6 | 14 | 15 | 30 | 40 | 74 | 75 | |  | residual plot showing non-random pattern |

Suppose we repeat the analysis, using a quadratic model to transform the dependent variable. For a quadratic model, we use the square root of y, rather than y, as the dependent variable. Using the transformed data, our regression equation is:

y't = b0 + b1x

where

yt = transformed dependent variable, which is equal to the square root of y  
y't = predicted value of the transformed dependent variable yt  
x = independent variable  
b0 = y-intercept of transformation regression line  
b1 = slope of transformation regression line

The table below shows the transformed data we analyzed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | yt | 1.41 | 1.00 | 2.45 | 3.74 | 3.87 | 5.48 | 6.32 | 8.60 | 8.66 | |  | residual plot showing random pattern |

Since the transformation was based on the quadratic model (yt = the square root of y), the transformation regression equation can be expressed in terms of the original units of variable Y as:

y' = ( b0 + b1x )2

where

y' = predicted value of y in its orginal units  
x = independent variable  
b0 = y-intercept of transformation regression line  
b1 = slope of transformation regression line

The residual plot (above right) shows residuals based on predicted raw scores from the transformation regression equation. The plot suggests that the transformation to achieve linearity was successful. The pattern of residuals is random, suggesting that the relationship between the independent variable (x) and the transformed dependent variable (square root of y) is linear. And the coefficient of determination was 0.96 with the transformed data versus only 0.88 with the raw data. The transformed data resulted in a better model.

Test Your Understanding

**Problem**

In the context of [regression analysis](http://stattrek.com/Help/Glossary.aspx?Target=Regression), which of the following statements is true?

I. A linear transformation increases the linear relationship between variables.   
II. A logarithmic model is the most effective transformation method.   
III. A residual plot reveals departures from linearity.

(A) I only   
(B) II only   
(C) III only   
(D) I and II only   
(E) I, II, and III

**Solution**

The correct answer is (C). A linear transformation neither increases nor decreases the linear relationship between variables; it preserves the relationship. A *nonlinear* transformation is used to increase the relationship between variables. The most effective transformation method depends on the data being transformed. In some cases, a logarithmic model may be more effective than other methods; but it other cases it may be less effective. Non-random patterns in a [residual plot](http://stattrek.com/Help/Glossary.aspx?Target=Residual%20plot) suggest a departure from linearity in the data being plotted.

Influential Points in Regression

Sometimes in regression analysis, a few data points have disproportionate effects on the slope of the regression equation. In this lesson, we describe how to identify those influential points.

Outliers

Data points that diverge in a big way from the overall pattern are called [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier). There are four ways that a data point might be considered an outlier.

* It could have an extreme X value compared to other data points.
* It could have an extreme Y value compared to other data points.
* It could have extreme X and Y values.
* It might be distant from the rest of the data, even without extreme X or Y values.

|  |  |
| --- | --- |
| **Extreme X value** | **Extreme Y value** |
| http://stattrek.com/Images/outlierX.gif | http://stattrek.com/Images/outlierY.gif |
| **Extreme X and Y** | **Distant data point** |
| http://stattrek.com/Images/outlierXY.gif | http://stattrek.com/Images/outliers.gif |

Influential Points

An influential point is an [outlier](http://stattrek.com/Help/Glossary.aspx?Target=Outlier) that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

This type of analysis is illustrated below. The scatter plots are identical, except that the plot on the right includes an outlier. The slope is flatter when the outlier is present (-3.32 vs. -4.10), so this outlier would be considered an influential point.

|  |  |  |
| --- | --- | --- |
| **Without Outlier** |  | **With Outlier** |
| http://stattrek.com/Images/Sp19.jpg |  | http://stattrek.com/Images/Sp20.jpg |
| Regression equation: ŷ = 104.78 - 4.10x Coefficient of determination: R2 = 0.94 |  | Regression equation: ŷ = 97.51 - 3.32x Coefficient of determination: R2 = 0.55 |

The charts below compare regression statistics for another data set with and without an outlier. Here, the chart on the right has a single outlier, located at the high end of the X axis (where x = 24). As a result of that single outlier, the slope of the regression line changes greatly, from -2.5 to -1.6; so the outlier would be considered an influential point.

|  |  |  |
| --- | --- | --- |
| **Without Outlier** |  | **With Outlier** |
| http://stattrek.com/Images/Sp21.jpg |  | http://stattrek.com/Images/Sp22.jpg |
| Regression equation: ŷ = 92.54 - 2.5x Slope: b0 = -2.5 Coefficient of determination: R2 = 0.46 |  | Regression equation: ŷ = 87.59 - 1.6x Slope: b0 = -1.6 Coefficient of determination: R2 = 0.52 |

Sometimes, an influential point will cause the [coefficient of determination](http://stattrek.com/Help/Glossary.aspx?Target=Coefficient%20of%20determination) to be bigger; sometimes, smaller. In the first example above, the coefficient of determination is smaller when the influential point is present (0.94 vs. 0.55). In the second example, it is bigger (0.46 vs. 0.52).

If your data set includes an influential point, here are some things to consider.

* An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
* Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

Test Your Understanding

In the context of [regression analysis](http://stattrek.com/Help/Glossary.aspx?Target=Regression), which of the following statements are true?

I. When the data set includes an influential point, the data set is nonlinear.   
II. Influential points always reduce the coefficient of determination.   
III. All outliers are influential data points.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above   
(E) None of the above

**Solution**

The correct answer is (E). Data sets with influential points can be linear or nonlinear. In this lesson, we went over an example in which an [influential point](http://stattrek.com/Help/Glossary.aspx?Target=Influential%20point) increased the coefficient of determination. With respect to regression, [outliers](http://stattrek.com/Help/Glossary.aspx?Target=Outlier) are influential only if they have a big effect on the regression equation. Sometimes, outliers do not have big effects. For example, when the data set is very large, a single outlier may not have a big effect on the regression equation.

Regression Slope: Confidence Interval

This lesson describes how to construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) around the [slope](http://stattrek.com/Help/Glossary.aspx?Target=Slope) of a [regression](http://stattrek.com/Help/Glossary.aspx?Target=Regression) line. We focus on the equation for simple linear regression, which is:

ŷ = b0 + b1x

where b0 is a constant, b1 is the slope (also called the regression coefficient), x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable.

Estimation Requirements

The approach described in this lesson is valid whenever the standard requirements for simple linear regression are met.

* The dependent variable *Y* has a linear relationship to the independent variable *X*.
* For each value of X, the probability distribution of Y has the same standard deviation σ.
* For any given value of X,
  + The Y values are independent.
  + The Y values are roughly normally distributed (i.e., [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry) and [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution)). A little [skewness](http://stattrek.com/Help/Glossary.aspx?Target=Skewness) is ok if the sample size is large.

Previously, we described [how to verify that regression requirements are met](http://stattrek.com/AP-Statistics-1/Regression.aspx#ReqressionPrerequisites).

The Variability of the Slope Estimate

To construct a [confidence interval](http://stattrek.com/Help/Glossary.aspx?Target=Confidence%20interval) for the slope of the regression line, we need to know the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) of the slope. Many statistical software packages and some graphing calculators provide the standard error of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: y = 76 + 35x .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| Constant | 76 | 30 | 2.53 | 0.01 |
| X | 35 | 20 | 1.75 | 0.04 |

In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as "SE Coeff". However, other software packages might use a different label for the standard error. It might be "StDev", "SE", "Std Dev", or something else.

If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

SE = sb1 = sqrt [ Σ(yi - ŷi)2 / (n - 2) ] / sqrt [ Σ(xi - x)2 ]

where yi is the value of the dependent variable for observation *i*, ŷi is estimated value of the dependent variable for observation *i*, xi is the observed value of the independent variable for observation *i*, x is the mean of the independent variable, and n is the number of observations.

How to Find the Confidence Interval for the Slope of a Regression Line

Previously, we described [how to construct confidence intervals](http://stattrek.com/AP-Statistics-4/Confidence-Interval.aspx#sixsteps). The confidence interval for the slope uses the same general approach. Note, however, that the critical value is based on a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) with *n* - 2 [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom).

* Identify a sample statistic. The sample statistic is the regression slope b1 calculated from sample data. In the table above, the regression slope is 35.
* Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
* Find the margin of error. Previously, we showed [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx), based on the [critical value](http://stattrek.com/Help/Glossary.aspx?Target=Critical%20value) and standard error. When calculating the margin of error for a regression slope, use a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) for the critical value, with [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) equal to *n* - 2.
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

In the next section, we work through a problem that shows how to use this approach to construct a confidence interval for the slope of a regression line.

Test Your Understanding

**Problem 1**

The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regression equation:** Annual bill = 0.55 \* Home size + 15 | | | | |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| Constant | 15 | 3 | 5.0 | 0.00 |
| Home size | 0.55 | 0.24 | 2.29 | 0.01 |

What is the 99% confidence interval for the slope of the regression line?

(A) 0.25 to 0.85   
(B) 0.02 to 1.08   
(C) -0.08 to 1.18   
(D) 0.20 to 1.30   
(E) 0.30 to 1.40

**Solution**

The correct answer is (C). Use the following four-step approach to construct a confidence interval.

* Identify a sample statistic. Since we are trying to estimate the slope of the true regression line, we use the regression coefficient for home size (i.e., the sample estimate of slope) as the sample statistic. From the regression output, we see that the slope coefficient is 0.55.
* Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 99% confidence level.
* Find the margin of error. Elsewhere on this site, we show [how to compute the margin of error](http://stattrek.com/AP-Statistics-4/Margin-Of-Error.aspx). The key steps applied to this problem are shown below.
  + Find standard deviation or standard error. The standard error is given in the regression output. It is 0.24.
  + Find critical value. The critical value is a factor used to compute the margin of error. With simple linear regression, to compute a confidence interval for the slope, the critical value is a [t score](http://stattrek.com/Help/Glossary.aspx?target=t_statistic) with [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) equal to *n* - 2. To find the critical value, we take these steps.
    - Compute alpha (α): α = 1 - (confidence level / 100) = 1 - 99/100 = 0.01
    - Find the critical probability (p\*): p\* = 1 - α/2 = 1 - 0.01/2 = 0.995
    - Find the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (df): df = *n* - 2 = 101 - 2 = 99.
    - The critical value is the t statistic having 99 degrees of freedom and a [cumulative probability](http://stattrek.com/Help/Glossary.aspx?Target=Cumulative%20probability) equal to 0.995. From the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx), we find that the critical value is 2.63.
  + Compute margin of error (ME): ME = critical value \* standard error = 2.63 \* 0.24 = 0.63
* Specify the confidence interval. The range of the confidence interval is defined by the *sample statistic* +*margin of error*. And the uncertainty is denoted by the confidence level.

Therefore, the 99% confidence interval for this sample is 0.55 + 0.63, which is -0.08 to 1.18

If we replicated the same study multiple times with different random samples and computed a confidence interval for each sample, we would expect 99% of the confidence intervals to contain the true slope of the regression line.

Hypothesis Test for Regression Slope

This lesson describes how to conduct a hypothesis test to determine whether there is a significant linear relationship between an independent variable *X* and a dependent variable *Y*. The test focuses on the [slope](http://stattrek.com/Help/Glossary.aspx?Target=Slope) of the [regression](http://stattrek.com/Help/Glossary.aspx?Target=Regression) line

Y = Β0 + Β1X

where Β0 is a constant, Β1 is the slope (also called the regression coefficient), X is the value of the independent variable, and Y is the value of the dependent variable.

If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.

Test Requirements

The approach described in this lesson is valid whenever the standard requirements for simple linear regression are met.

* The dependent variable *Y* has a linear relationship to the independent variable *X*.
* For each value of X, the probability distribution of Y has the same standard deviation σ.
* For any given value of X,
  + The Y values are independent.
  + The Y values are roughly normally distributed (i.e., [symmetric](http://stattrek.com/Help/Glossary.aspx?Target=Symmetry) and [unimodal](http://stattrek.com/Help/Glossary.aspx?Target=Unimodal%20distribution)). A little [skewness](http://stattrek.com/Help/Glossary.aspx?Target=Skewness) is ok if the sample size is large.

Previously, we described [how to verify that regression requirements are met](http://stattrek.com/AP-Statistics-1/Regression.aspx#ReqressionPrerequisites).

The test procedure consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

If there is a significant linear relationship between the independent variable *X* and the dependent variable *Y*, the slope will *not* equal zero.

H0: Β1 = 0   
Ha: Β1 ≠ 0

The [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

* Significance level. Often, researchers choose [significance levels](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* Test method. Use a linear regression t-test (described in the next section) to determine whether the slope of the regression line differs significantly from zero.

Analyze Sample Data

Using sample data, find the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the slope, the slope of the regression line, the degrees of freedom, the test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

* Standard error. Many statistical software packages and some graphing calculators provide the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: y = 76 + 35x .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| Constant | 76 | 30 | 2.53 | 0.01 |
| X | 35 | 20 | 1.75 | 0.04 |

* In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as "SE Coeff". However, other software packages might use a different label for the standard error. It might be "StDev", "SE", "Std Dev", or something else.   
    
  If you need to calculate the standard error of the slope (SE) by hand, use the following formula:
* SE = sb1 = sqrt [ Σ(yi - ŷi)2 / (n - 2) ] / sqrt [ Σ(xi - x)2 ]
* where yi is the value of the dependent variable for observation *i*, ŷi is estimated value of the dependent variable for observation *i*, xi is the observed value of the independent variable for observation *i*, x is the mean of the independent variable, and n is the number of observations.
* Slope. Like the standard error, the slope of the regression line will be provided by most statistics software packages. In the hypothetical output above, the slope is equal to 35.
* Degrees of freedom. For simple linear regression (one independent and one dependent variable), the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is equal to:

DF = n - 2

where n is the number of observations in the sample.

* Test statistic. The test statistic is a t statistic (t) defined by the following equation.

t = b1 / SE

where b1 is the slope of the sample regression line, and SE is the standard error of the slope.

* P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the [t Distribution Calculator](http://stattrek.com/Tables/T.aspx) to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](http://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

**Problem**

The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regression equation:** Annual bill = 0.55 \* Home size + 15 | | | | |
| **Predictor** | **Coef** | **SE Coef** | **T** | **P** |
| Constant | 15 | 3 | 5.0 | 0.00 |
| Home size | 0.55 | 0.24 | 2.29 | 0.01 |

Is there a significant linear relationship between annual bill and home size? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an alternative hypothesis.

H0: The slope of the regression line is equal to zero.   
Ha: The slope of the regression line is *not* equal to zero.

If the relationship between home size and electric bill is significant, the slope will *not* equal zero.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.
* **Analyze sample data**. To apply the linear regression t-test to sample data, we require the [standard error](http://stattrek.com/Help/Glossary.aspx?Target=Standard%20error) of the slope, the slope of the regression line, the degrees of freedom, the t statistic test statistic, and the P-value of the test statistic.

We get the slope (b1) and the standard error (SE) from the regression output.

b1 = 0.55 SE = 0.24

We compute the degrees of freedom and the t statistic test statistic, using the following equations.

DF = n - 2 = 101 - 2 = 99   
  
t = b1/SE = 0.55/0.24 = 2.29

where DF is the degrees of freedom, n is the number of observations in the sample, b1 is the slope of the regression line, and SE is the standard error of the slope.

Based on the t statistic test statistic and the [degrees of freedom](http://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](http://stattrek.com/Help/Glossary.aspx?Target=P-value). The P-value is the probability that a t statistic having 99 degrees of freedom is more extreme than 2.29. Since this is a [two-tailed test](http://stattrek.com/Help/Glossary.aspx?Target=Two-tailed%20test), "more extreme" means greater than 2.29 or less than -2.29. We use the [t Distribution Calculator](http://stattrek.com/Tables/t.aspx) to find P(t > 2.29) = 0.0121 and P(t < 2.29) = 0.0121. Therefore, the P-value is 0.0121 + 0.0121 or 0.0242.

* **Interpret results**. Since the P-value (0.0242) is less than the significance level (0.05), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention that this approach is only appropriate when the [standard requirements for simple linear regression](http://stattrek.com/AP-Statistics-1/Regression.aspx#ReqressionPrerequisites) are satisfied.

What is an Experiment?

In an experiment, a researcher manipulates one or more variables, while holding all other variables constant. By noting how the manipulated variables affect a response variable, the researcher can test whether a causal relationship exists between the manipulated variables and the response variable.

Parts of an Experiment

All experiments have independent variables, dependent variables, and experimental units.

* **Independent variable**. An independent variable (also called a **factor**) is an explanatory variable manipulated by the experimenter.   
    
  Each factor has two or more **levels** (i.e., different values of the factor). Combinations of factor levels are called **treatments**. The table below shows independent variables, factors, levels, and treatments for a hypothetical experiment.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  |  | Vitamin C | | | |  |  | 0 mg | 250 mg | 500 mg | | Vitamin E | 0 mg | Treatment 1 | Treatment 2 | Treatment 3 | | 400 mg | Treatment 4 | Treatment 5 | Treatment 6 | |

* In this hypothetical experiment, the researcher is studying the possible effects of Vitamin C and Vitamin E on health. There are two factors - dosage of Vitamin C and dosage of Vitamin E. The Vitamin C factor has three levels - 0 mg per day, 250 mg per day, and 500 mg per day. The Vitamin E factor has 2 levels - 0 mg per day and 400 mg per day. The experiment has six treatments. Treatment 1 is 0 mg of E and 0 mg of C, Treatment 2 is 0 mg of E and 250 mg of C, and so on.
* **Dependent variable**. In the hypothetical experiment above, the researcher is looking at the effect of vitamins on health. The dependent variable in this experiment would be some measure of health (annual doctor bills, number of colds caught in a year, number of days hospitalized, etc.).
* **Experimental units**. The recipients of experimental treatments are called experimental units. The experimental units in an experiment could be anything - people, plants, animals, or even inanimate objects.   
    
  In the hypothetical experiment above, the experimental units would probably be people (or lab animals). But in an experiment to measure the tensile strength of string, the experimental units might be pieces of string. When the experimental units are people, they are often called participants; when the experimental units are animals, they are often called subjects.

Characteristics of a Well-Designed Experiment

A well-designed experiment includes design features that allow researchers to eliminate extraneous variables as an explanation for the observed relationship between the independent variable(s) and the dependent variable. Some of these features are listed below.

* **Control**. Control refers to steps taken to reduce the effects of extraneous variables (i.e., variables other than the independent variable and the dependent variable). These extraneous variables are called **lurking variables**.  
    
  Control involves making the experiment as similar as possible for experimental units in each treatment condition. Three control strategies are control groups, placebos, and blinding.
  + **Control group**. A control group is a baseline group that receives no treatment or a neutral treatment. To assess treatment effects, the experimenter compares results in the treatment group to results in the control group.
  + **Placebo**. Often, participants in an experiment respond differently after they receive a treatment, even if the treatment is neutral. A neutral treatment that has no "real" effect on the dependent variable is called a **placebo**, and a participant's positive response to a placebo is called the **placebo effect**.   
      
    To control for the placebo effect, researchers often administer a neutral treatment (i.e., a placebo) to the control group. The classic example is using a sugar pill in drug research. The drug is considered effective only if participants who receive the drug have better outcomes than participants who receive the sugar pill.
  + **Blinding**. Of course, if participants in the control group know that they are receiving a placebo, the placebo effect will be reduced or eliminated; and the placebo will not serve its intended control purpose.   
      
    Blinding is the practice of not telling participants whether they are receiving a placebo. In this way, participants in the control and treatment groups experience the placebo effect equally. Often, knowledge of which groups receive placebos is also kept from people who administer or evaluate the experiment. This practice is called **double blinding**. It prevents the experimenter from "spilling the beans" to participants through subtle cues; and it assures that the analyst's evaluation is not tainted by awareness of actual treatment conditions.
* **Randomization**. Randomization refers to the practice of using chance methods (random number tables, flipping a coin, etc.) to assign experimental units to treatments. In this way, the potential effects of lurking variables are distributed at chance levels (hopefully roughly evenly) across treatment conditions.
* **Replication**. Replication refers to the practice of assigning each treatment to many experimental units. In general, the more experimental units in each treatment condition, the lower the variability of the dependent measures.

Confounding

**Confounding** occurs when the experimental controls do not allow the experimenter to reasonably eliminate plausible alternative explanations for an observed relationship between independent and dependent variables.

Consider this example. A drug manufacturer tests a new cold medicine with 200 participants - 100 men and 100 women. The men receive the drug, and the women do not. At the end of the test period, the men report fewer colds.

This experiment implements no controls! As a result, many variables are confounded, and it is impossible to say whether the drug was effective. For example, gender is confounded with drug use. Perhaps, men are less vulnerable to the particular cold virus circulating during the experiment, and the new medicine had no effect at all. Or perhaps the men experienced a placebo effect.

This experiment could be strengthened with a few controls. Women and men could be randomly assigned to treatments. One treatment group could receive a placebo, with blinding. Then, if the treatment group (i.e., the group getting the medicine) had sufficiently fewer colds than the control group, it would be reasonable to conclude that the medicine was effective in preventing colds.

Test Your Understanding

**Problem**

Which of the following statements are true?

I. Blinding controls for the effects of confounding.   
II. Randomization controls for effects of lurking variables.   
III. Each factor has one treatment level.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above.   
(E) None of the above.

**Solution**

The correct answer is (B). By randomly assigning experimental units to treatment levels, randomization spreads potential effects of [lurking variables](http://stattrek.com/Help/Glossary.aspx?Target=Lurking%20variable) roughly evenly across treatment levels. [Blinding](http://stattrek.com/Help/Glossary.aspx?Target=Blinding) ensures that participants in control and treatment conditions experience the [placebo effect](http://stattrek.com/Help/Glossary.aspx?Target=Placebo%20effect) equally, but it does not guard against [confounding](http://stattrek.com/Help/Glossary.aspx?Target=Confounding). And finally, each [factor](http://stattrek.com/Help/Glossary.aspx?Target=Factor) has *two* or more treatment levels. If a factor had only one treatment level, each participant in the experiment would get the same treatment on that factor. As a result, that factor would be [confounded](http://stattrek.com/Help/Glossary.aspx?Target=Confounding) with every other factor in the experiment.

Experimental Design

The term **experimental design** refers to a plan for assigning experimental units to [treatment](http://stattrek.com/Help/Glossary.aspx?Target=Treatment) conditions. A good experimental design serves three purposes.

* **Causation**. It allows the experimenter to make causal inferences about the relationship between [independent variables](http://stattrek.com/Help/Glossary.aspx?Target=Independent%20variable) and a [dependent variable](http://stattrek.com/Help/Glossary.aspx?Target=Dependent%20variable).
* **Control**. It allows the experimenter to rule out alternative explanations due to the [confounding](http://stattrek.com/Help/Glossary.aspx?Target=Confounding) effects of extraneous variables (i.e., variables other than the independent variables).
* **Variability**. It reduces variability within treatment conditions, which makes it easier to detect differences in treatment outcomes.

An Experimental Design Example

Consider the following hypothetical experiment. Acme Medicine is conducting an experiment to test a new vaccine, developed to immunize people against the common cold. To test the vaccine, Acme has 1000 volunteers - 500 men and 500 women. The participants range in age from 21 to 70.

In this lesson, we describe three experimental designs - a completely randomized design, a randomized block design, and a matched pairs design. And we show how each design might be applied by Acme Medicine to understand the effect of the vaccine, while ruling out confounding effects of other factors.

Completely Randomized Design

The **completely randomized design** is probably the simplest experimental design, in terms of data analysis and convenience. With this design, participants are randomly assigned to treatments.

|  |  |
| --- | --- |
| Treatment | |
| Placebo | Vaccine |
| 500 | 500 |

A completely randomized design layout for the Acme Experiment is shown in the table to the right. In this design, the experimenter randomly assigned participants to one of two treatment conditions. They received a [placebo](http://stattrek.com/Help/Glossary.aspx?Target=Placebo) or they received the vaccine. The same number of participants (500) were assigned to each treatment condition (although this is not required). The dependent variable is the number of colds reported in each treatment condition. If the vaccine is effective, participants in the "vaccine" condition should report significantly fewer colds than participants in the "placebo" condition.

A completely randomized design relies on [randomization](http://stattrek.com/Help/Glossary.aspx?Target=Randomization) to control for the effects of extraneous variables. The experimenter assumes that, on averge, extraneous factors will affect treatment conditions equally; so any significant differences between conditions can fairly be attributed to the independent variable.

Randomized Block Design

With a **randomized block design**, the experimenter divides participants into subgroups called **blocks**, such that the variability within blocks is less than the variability between blocks. Then, participants within each block are randomly assigned to treatment conditions. Because this design reduces variability and potential confounding, it produces a better estimate of treatment effects.

|  |  |  |
| --- | --- | --- |
| Gender | Treatment | |
| Placebo | Vaccine |
| Male | 250 | 250 |
| Female | 250 | 250 |

The table to the right shows a randomized block design for the Acme experiment. Participants are assigned to blocks, based on gender. Then, within each block, participants are randomly assigned to treatments. For this design, 250 men get the placebo, 250 men get the vaccine, 250 women get the placebo, and 250 women get the vaccine.

It is known that men and women are physiologically different and react differently to medication. This design ensures that each treatment condition has an equal proportion of men and women. As a result, differences between treatment conditions cannot be attributed to gender. This randomized block design removes gender as a potential source of variability and as a potential confounding variable.

In this Acme example, the randomized block design is an improvement over the completely randomized design. Both designs use randomization to implicitly guard against confounding. But only the randomized block design explicitly controls for gender.

Note 1: In some blocking designs, individual participants may receive multiple treatments. This is called using the participant *as his own control*. Using the participant as his own control is desirable in some experiments (e.g., research on learning or fatigue). But it can also be a problem (e.g., medical studies where the medicine used in one treatment might interact with the medicine used in another treatment).

Note 2: Blocks perform a similar function in experimental design as [strata](http://stattrek.com/Help/Glossary.aspx?Target=Strata) perform in sampling. Both divide observations into subgroups. However, they are not the same. Blocking is associated with experimental design, and stratification is associated with survey sampling.

Matched Pairs Design

|  |  |  |
| --- | --- | --- |
| Pair | Treatment | |
| Placebo | Vaccine |
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| ... | ... | ... |
| 499 | 1 | 1 |
| 500 | 1 | 1 |

A **matched pairs design** is a special case of the randomized block design. It is used when the experiment has only two treatment conditions; and participants can be grouped into pairs, based on some blocking variable. Then, within each pair, participants are randomly assigned to different treatments.

The table to the right shows a matched pairs design for the Acme experiment. The 1000 participants are grouped into 500 matched pairs. Each pair is matched on gender and age. For example, Pair 1 might be two women, both age 21. Pair 2 might be two women, both age 22, and so on.

For the Acme example, the matched pairs design is an improvement over the completely randomized design and the randomized block design. Like the other designs, the matched pairs design uses randomization to control for confounding. However, unlike the others, this design explicitly controls for two potential [lurking variables](http://stattrek.com/Help/Glossary.aspx?Target=Lurking%20variable) - age and gender.

Test Your Understanding

**Problem**

Which of the following statements are true?

I. A completely randomized design offers no control for lurking variables.   
II. A randomized block design controls for the placebo effect.   
III. In a matched pairs design, participants within each pair receive the same treatment.

(A) I only   
(B) II only   
(C) III only   
(D) All of the above.   
(E) None of the above.

**Solution**

The correct answer is (E). In a [completely randomized design](http://stattrek.com/Help/Glossary.aspx?Target=Completely%20randomized%20design), experimental units are randomly assigned to treatment conditions. [Randomization](http://stattrek.com/Help/Glossary.aspx?Target=Randomization) provides some control for [lurking variables](http://stattrek.com/Help/Glossary.aspx?Target=Lurking%20variable). By itself, a [randomized block design](http://stattrek.com/Help/Glossary.aspx?Target=Randomized%20block%20design) does not control for the [placebo effect](http://stattrek.com/Help/Glossary.aspx?Target=Placebo). To control for the placebo effect, the experimenter must include a placebo in one of the treatment levels. In a [matched pairs design](http://stattrek.com/Help/Glossary.aspx?Target=Matched%20pairs%20design), experimental units within each pair are assigned to *different* treatment levels.

Simulation of Random Events

**Simulation** is a way to model random events, such that simulated outcomes closely match real-world outcomes. By observing simulated outcomes, researchers gain insight on the real world.

Why use simulation?

Some situations do not lend themselves to precise mathematical treatment. Others may be difficult, time-consuming, or expensive to analyze. In these situations, simulation may approximate real-world results; yet, require less time, effort, and/or money than other approaches.

How to Conduct a Simulation

A simulation is useful only if it closely mirrors real-world outcomes. The steps required to produce a useful simulation are presented below.

1. Describe the possible outcomes.
2. Link each outcome to one or more random numbers.
3. Choose a source of random numbers.
4. Choose a random number.
5. Based on the random number, note the "simulated" outcome.
6. Repeat steps 4 and 5 multiple times; preferably, until the outcomes show a stable pattern.
7. Analyze the simulated outcomes and report results.

**Note:** When it comes to choosing a source of random numbers (Step 3 above), you have many options. Flipping a coin and rolling dice are low-tech but effective. Tables of random numbers (often found in the appendices of statistics texts) are another option. And good random number generators can be found on the internet.

Random Number Generator

In practice, flipping a coin or rolling dice to obtain random numbers can be cumbersome, particularly with large sample sizes. As an alternative, use Stat Trek's Random Number Generator. With the Random Number Generator, you can select up to 1000 random numbers quickly and easily. This tool is provided at no cost - free!! To access the Random Number Generator, simply click on the button below. It can also be found under the Stat Tools tab, which appears in the header of every Stat Trek web page.

|  |
| --- |
| [Random Number Generator](http://stattrek.com/Tables/Random.aspx) |

Simulation Example

In this section, we work through an example to show how to apply simulation methods to probability problems.

**Problem Description**

On average, suppose a baseball player hits a home run once in every 10 times at bat, and suppose he gets exactly two "at bats" in every game. Using simulation, estimate the likelihood that the player will hit 2 home runs in a single game.

**Solution**

Earlier we described seven steps required to produce a useful simulation. Let's apply those steps to this problem.

1. Describe the possible outcomes. For this problem, there are two outcomes - the player hits a home run or he doesn't.
2. Link each outcome to one or more random numbers. Since the player hits a home run in 10% of his at bats, 10% of the random numbers should represent a home run. For this problem, let's say that the digit "2" represents a home run and any other digit represents a different outcome.
3. Choose a source of random numbers. For this problem, we used Stat Trek's [Random Number Generator](http://stattrek.com/Tables/Random.aspx) to produce a list of 500 two-digit numbers (see below).
4. Choose a random number. The list below shows the random numbers that we generated.
5. Based on the random number, note the "simulated" outcome. In this example, each 2-digit number represents two "at-bats" in a single game. Since the digit "2" represents a home run, the number "22" represents two home runs in a single game. Any other 2-digit number represents a failure to hit consecutive home runs in the game.
6. Repeat steps 4 and 5 multiple times; preferably, until the outcomes show a stable pattern. In this example, the list of random numbers consists of 500 2-digit pairs; i.e., 500 repetitions of steps 4 and 5.
7. Analyze the simulated outcomes and report results. In the list, we found 6 occurrences of "22", which are highlighted in red in the table. In this simulation, each occurrence of "22" represents a game in which the player hit consecutive home runs.

|  |
| --- |
| **Random Numbers** |
| 42 99 02 65 04 14 30 09 70 88 89 85 95 40 53 67 25 50 48 79 86 92 76 24 53 39 08 73 78 17 72 81 08 01 68 94 43 43 95 12 36 90 28 88 34 69 18 69 91 79 14 82 26 94 15 26 19 41 74 02 17 20 38 84 74 30 34 96 09 46 61 41 02 93 94 90 00 71 84 98 30 82 80 11 92 97 81 29 85 44 40 05 83 22 04 86 13 33 00 99 74 75 27 43 68 22 59 20 66 00 24 01 96 84 19 14 57 26 47 58 51 73 06 08 49 52 70 15 79 35 65 28 40 77 93 73 33 24 25 22 32 03 89 03 62 13 85 16 23 28 12 61 16 75 45 37 15 54 36 18 45 64 31 31 06 80 32 75 99 27 91 25 98 05 55 32 27 16 51 45 89 31 78 90 82 05 11 39 80 83 01 20 10 67 97 33 72 09 98 78 39 56 57 54 63 35 21 35 93 18 17 48 55 60 44 92 21 07 77 42 46 86 41 49 76 96 36 62 38 11 64 07 04 58 23 56 29 37 87 37 59 47 83 77 21 63 10 95 87 10 42 71 12 88 06 52 42 99 02 65 04 14 30 09 70 88 89 85 95 40 53 67 25 50 48 79 86 92 76 24 53 39 08 73 78 17 72 81 08 01 68 94 43 43 95 12 36 90 28 88 34 69 18 69 91 79 14 82 26 94 15 26 19 41 74 02 17 20 38 84 74 30 34 96 09 46 61 41 02 93 94 90 00 71 84 98 30 82 80 11 92 97 81 29 85 44 40 05 83 22 04 86 13 33 00 99 74 75 27 43 68 22 59 20 66 00 24 01 96 84 19 14 57 26 47 58 51 73 06 08 49 52 70 15 79 35 65 28 40 77 93 73 33 24 25 22 32 03 89 03 62 13 85 16 23 28 12 61 16 75 45 37 15 54 36 18 45 64 31 31 06 80 32 75 99 27 91 25 98 05 55 32 27 16 51 45 89 31 78 90 82 05 11 39 80 83 01 20 10 67 97 33 72 09 98 78 39 56 57 54 63 35 21 35 93 18 17 48 55 60 44 92 21 07 77 42 46 86 41 49 76 96 36 62 38 11 64 07 04 58 23 56 29 37 87 37 59 47 83 77 |

The simulation predicts that this particular player will hit consecutive home runs 6 times in 500 games. Thus, the simulation suggests that there is a 1.2% chance that he will hit two home runs in a single game. The actual probability, based on the [multiplication rule](http://stattrek.com/Help/Glossary.aspx?Target=Multiplication%20rule), states that there is a 1.0% chance that this player will hit consecutive home runs in a game. While the simulation is not exact, it is very close. And, if we had generated a list with more random numbers, it likely would have been even closer.