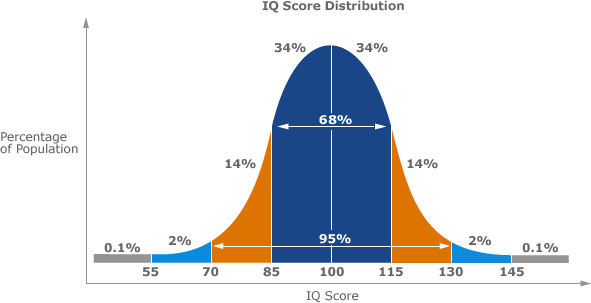
## **Statistics Overview**

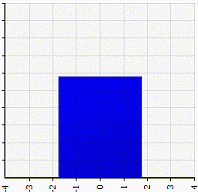
Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a [statistical population](https://en.wikipedia.org/wiki/Statistical_population) or a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of [surveys](https://en.wikipedia.org/wiki/Statistical_survey) and [experiments](https://en.wikipedia.org/wiki/Experimental_design)

The most common basic statistics terms you’ll come across are the [mean, mode and median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/). These are all what are known as “Measures of [Central Tendency](https://www.statisticshowto.datasciencecentral.com/central-tendency-2/).” Also important in this early chapter of statistics is the [shape of a distribution](https://www.statisticshowto.datasciencecentral.com/shapes-of-distributions/). This tells us something about how data is spread out around the [mean](https://www.statisticshowto.datasciencecentral.com/mean) or [median](https://www.statisticshowto.datasciencecentral.com/median). Perhaps the most common distribution you’ll see is the [**normal distribution**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), sometimes called a bell curve. Heights, weights, and many other things found in nature tend to be shaped like this:



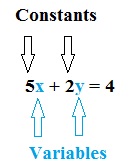
*IQ scores fit a bell curve shape.*

On the other end of the scale, you can also get a **flat distribution**. With this shape, the odds of anything happening are equal. For example, a [uniform distribution](https://www.statisticshowto.datasciencecentral.com/uniform-distribution/) can represent choosing a particular card from a standard deck; all the cards have a 1/52 chance of being chosen. Or tossing a coin, where you have a 50% chance of tossing a heads or a tails.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/shape_uniform.gif)

*A uniform distribution.*

## **Types of Variables in Statistics and Research**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/variable.jpg)

A “variable” in algebra really just means one thing—an unknown value. However, in statistics, you’ll come across dozens of types of variables in statistics. In most cases, the word still means that you’re dealing with something that’s unknown, but—unlike in algebra—that unknown isn’t always a number. Some variable types are used more than others. For example, you’ll be much more likely to come across [continuous variables](https://www.statisticshowto.datasciencecentral.com/discrete-vs-continuous-variables/) than you would [dummy variables](https://www.statisticshowto.datasciencecentral.com/dummy-variables/).

**Click on any bold variable name to learn more about that particular type.**

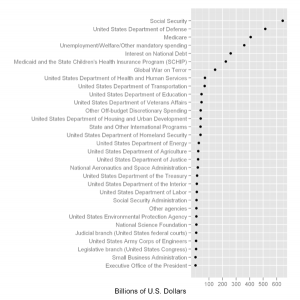
### **Common Types of Variables**

* [**Categorical variable**](https://www.statisticshowto.datasciencecentral.com/what-is-a-categorical-variable/): variables than can be put into categories. For example, the category “Toothpaste Brands” might contain the variables Colgate and Aquafresh.
* [**Confounding variable**](https://www.statisticshowto.datasciencecentral.com/confounding-variable/): extra variables that have a hidden effect on your experimental results.
* **Continuous variable**: a variable with infinite number of values, like “time” or “weight”.
* [**Control variable**](https://www.statisticshowto.datasciencecentral.com/control-variable/): a factor in an experiment which must be held constant. For example, in an experiment to determine whether light makes plants grow faster, you would have to control for soil quality and water.
* [**Dependent variable**](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/): the outcome of an experiment. As you change the independent variable, you watch what happens to the dependent variable.
* [**Discrete variable**](https://www.statisticshowto.datasciencecentral.com/discrete-vs-continuous-variables/): a variable that can only take on a certain number of values. For example, “number of cars in a parking lot” is discrete because a car park can only hold so many cars.
* [**Independent variable**](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/): a variable that is not affected by anything that you, the researcher, does. Usually plotted on the x-axis.
* [**Lurking variable**](https://www.statisticshowto.datasciencecentral.com/lurking-variable/)**: a “hidden” variable the affects the relationship between the independent and dependent variables.**
* A [**measurement variable**](https://www.statisticshowto.datasciencecentral.com/measurement-variable/) has a number associated with it. It’s an “amount” of something, or a”number” of something.
* [**Nominal variable**](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/): another name for categorical variable.
* [**Ordinal variable**](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/): similar to a categorical variable, but there is a clear order. For example, income levels of low, middle, and high could be considered ordinal.
* [**Qualitative variable**](https://www.statisticshowto.datasciencecentral.com/qualitative-variable/): a broad category for any variable that can’t be counted (i.e. has no numerical value). Nominal and ordinal variables fall under this umbrella term.
* [**Quantitative variable**](https://www.statisticshowto.datasciencecentral.com/what-are-quantitative-variables-and-quantitative-data/)**:** A broad category that includes any variable that can be counted, or has a numerical value associated with it. Examples of variables that fall into this category include discrete variables and ratio variables.
* [**Random variables**](https://www.statisticshowto.datasciencecentral.com/random-variable/) are associated with random processes and give numbers to outcomes of random events.
* A [**ranked variable**](https://www.statisticshowto.datasciencecentral.com/ranked-variable/) is an ordinal variable; a variable where every data point can be put in order (1st, 2nd, 3rd, etc.).
* [**Ratio variables**](https://www.statisticshowto.datasciencecentral.com/ratio-scale/): similar to interval variables, but has a meaningful zero.

### **Less Common Types of Variables**

* [**Active Variable**](https://www.statisticshowto.datasciencecentral.com/active-variable/)**:** a variable that is manipulated by the researcher.
* [**Antecedent Variable**](https://www.statisticshowto.datasciencecentral.com/antecedent-variable/): a variable that comes before the independent variable.
* [**Attribute variable**](https://www.statisticshowto.datasciencecentral.com/attribute-variable/): another name for a categorical variable (in statistical software) or a variable that isn’t manipulated (in [design of experiments](https://www.statisticshowto.datasciencecentral.com/experimental-design/)).
* [**Binary variable**](https://www.statisticshowto.datasciencecentral.com/binary-variable-2/): a variable that can only take on two values, usually 0/1. Could also be yes/no, tall/short or some other two-variable combination.
* [**Collider Variable**](https://www.statisticshowto.datasciencecentral.com/collider/): a variable represented by a node on a causal graph that has paths pointing in as well as out.
* [**Covariate variable**](https://www.statisticshowto.datasciencecentral.com/covariate/): similar to an independent variable, it has an effect on the dependent variable but is usually not the variable of interest. See also: **[Noncomitant variable](https://www.statisticshowto.datasciencecentral.com/concomitant-variable/)**[.](https://www.statisticshowto.datasciencecentral.com/concomitant-variable/)
* [**Criterion variable**](https://www.statisticshowto.datasciencecentral.com/criterion-variable-2/): another name for a dependent variable, when the variable is used in non-experimental situations.
* [**Dichotomous variable**](https://www.statisticshowto.datasciencecentral.com/dichotomous-variable/): Another name for a binary variable.
* [**Dummy Variables**](https://www.statisticshowto.datasciencecentral.com/dummy-variables/): used in [regression analysis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/) when you want to assign relationships to unconnected categorical variables. For example, if you had the categories “has dogs” and “owns a car” you might assign a 1 to mean “has dogs” and 0 to mean “owns a car.”
* [**Endogenous variable**](https://www.statisticshowto.datasciencecentral.com/endogenous-variable/): similar to dependent variables, they are affected by other variables in the system. Used almost exclusively in [econometrics](http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf).
* [**Exogenous variable**](https://www.statisticshowto.datasciencecentral.com/endogenous-variable/): variables that affect others in the system.
* [**Explanatory Variable**](https://www.statisticshowto.datasciencecentral.com/explanatory-variable/): a type of independent variable. When a variable is independent, it is not affected at all by any other variables. When a variable isn’t independent for certain, it’s an explanatory variable.
* [**Extraneous variables**](https://www.statisticshowto.datasciencecentral.com/extraneous-variable/) are any variables that you are not intentionally studying in your experiment or test.
* A [**grouping variable**](https://www.statisticshowto.datasciencecentral.com/grouping-variable/) (also called a coding variable, group variable or by variable) sorts data within data files into categories or groups.
* [**Identifier Variables:**](https://www.statisticshowto.datasciencecentral.com/identifier-variables-2/) variables used to uniquely identify situations.
* [**Indicator variable**](https://www.statisticshowto.datasciencecentral.com/dummy-variables/): another name for a dummy variable.
* [**Interval variable**](https://www.statisticshowto.datasciencecentral.com/interval-scale/): a meaningful measurement between two variables. Also sometimes used as another name for a continuous variable.
* [**Intervening variable**](https://www.statisticshowto.datasciencecentral.com/intervening-variable/): a variable that is used to explain the relationship between variables.
* [**Latent Variable:**](https://www.statisticshowto.datasciencecentral.com/factor-analysis/#Latent) a hidden variable that can’t be measured or observed directly.
* [**Manifest variable**](https://www.statisticshowto.datasciencecentral.com/factor-analysis/#Manifest): a variable that can be directly observed or measured.
* [**Manipulated variable**](https://www.statisticshowto.datasciencecentral.com/manipulated-variable/): another name for independent variable.
* [**Mediating variable**](https://www.statisticshowto.datasciencecentral.com/mediator-variable/) or [intervening variable](https://www.statisticshowto.datasciencecentral.com/intervening-variable/): variables that explain how the relationship between variables happens. For example, it could explain the difference between the predictor and criterion.
* [**Moderating variable**](https://www.statisticshowto.datasciencecentral.com/moderating-variable/): changes the strength of an effect between independent and dependent variables. For example, psychotherapy may reduce stress levels for women more than men, so sex moderates the effect between psychotherapy and stress levels.
* [**Nuisance Variable**](https://www.statisticshowto.datasciencecentral.com/nuisance-variable/): an [extraneous variable](https://www.statisticshowto.datasciencecentral.com/extraneous-variable/) that increases variability overall.
* [**Observed Variable**](https://www.statisticshowto.datasciencecentral.com/observed-variables/): a measured variable (usually used in [SEM](https://www.statisticshowto.datasciencecentral.com/simultaneous-equations-model/)).
* [**Outcome variable**:](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/#DVOutcome) similar in meaning to a dependent variable, but used in a non-experimental study.
* [**Polychotomous variables**](https://www.statisticshowto.datasciencecentral.com/polychotomous-variable): variables that can have more than two values.
* [**Predictor variable**](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/): similar in meaning to the independent variable, but used in regression and in non-experimental studies.
* [**Responding variable**](https://www.statisticshowto.datasciencecentral.com/responding-variable/): an informal term for dependent variable, usually used in science fairs.
* [**Scale Variable**](https://www.statisticshowto.datasciencecentral.com/scale-variable/): basically, another name for a measurement variable.
* [**Study Variable (Research Variable)**](https://www.statisticshowto.datasciencecentral.com/research-variable-study/): can mean any variable used in a study, but does have a more formal definition when used in a clinical trial.
* [**Test Variable**](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/): another name for the [Dependent Variable](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/).
* [**Treatment variable**](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/): another name for independent variable.

## **Quantitative Variables (Numeric Variables)**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/quantitative-and-categorical-variables.png)

*Graph of categorical variables on the y-axis and quantitative/numerical data on the x-axis. Credit: Thupper|Wikimedia Commons*

**Two** types of [variables](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/types-of-variables/)are used in statistics: **Quantitative** and **categorical** (also called qualitative):

* **Quantitative variables** are numerical variables: counts, percents, or numbers.
* [**Categorical(Qualitative) variable**](https://www.statisticshowto.datasciencecentral.com/what-is-a-categorical-variable/)s are descriptions of groups or things, like “breeds of dog” or “voting preference”.

### **Examples of Quantitative Variables / Numeric Variables:**

* High school Grade Point Average (e.g. 4.0, 3.2, 2.1).
* Number of pets owned (e.g. 1, 2, 4).
* Bank account balance (e.g. $100, $987, $-42.
* Number of stars in a galaxy (e.g. 100, 2301, 1 trillion) .
* [Average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)number of lottery tickets sold (e.g. 25, 2,789, 2 million).
* How many cousins you have (e.g. 0, 12, 22).
* The amount in your paycheck (e.g. $200, $1,457, $2,222).

General rule of thumb:**if you can add it, it’s quantitative.** For example, a G.P.A. of 3.3 and a G.P.A. of 4.0 can be added together (3.3 + 4.0 = 7.3), so that means it’s quantitative. On the other hand, grades of A, B, or C can’t be added together unless you convert them to numbers, so A, B, and C, are not quantitative.

### **Examples of Categorical Variables**

* Class in college (e.g. freshman, sophomore, junior, senior).
* Party affiliation (e.g. Republican, Democrat, Independent).
* Type of pet owned (e.g. dog, cat, rodent, fish).
* Favourite author (e.g. Stephen King, James Patterson, Charles Dickens).
* Preferred airline (e.g. Southwest, Virgin, Quantas).
* Hair color (e.g. blond, brunette, black).
* Your race (e.g. Asian, Latino, black).
* Types of hats (e.g. sombrero, beanie, fedora).

As a general rule, **if you can’t add something**, then it’s categorical. For example, you can’t add cat + dog, or Republican + Democrat.

### **Categorical vs. Quantitative**

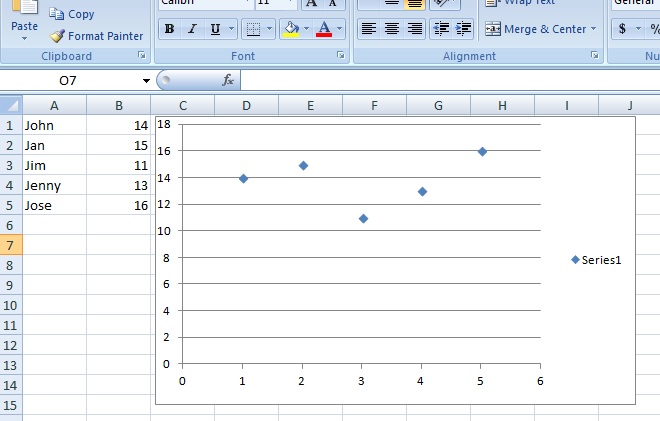
Watch this video on the difference between categorical(qualitative) and quantitative variables.

[](https://www.youtube.com/watch?v=muaVOWWMjXY)

### **What is a Quantitative Data Condition?**

**When you graph or plot statistical data, make sure you have quantitative data of known units**. If you don’t have known units, then you won’t be able to graph it. For example, the first list above states that “G.P.A.” is quantitative data. However, you won’t be able to graph G.P.A. versus another variable (say, race or sex) unless you actually have a unit, like 3.1 or 2.9. This sounds obvious, but with more complex data you should always check the quantitative data condition for missing or nonsensical information before you start a graph.

[Histograms](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/histogram-make-chart/), [boxplots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/)and [scatter plots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/) all require that you have quantitative (numerical data). If you try to graph categorical data with a histogram, boxplot or scatter plot, you’ll run into the same type of problem as if you try to graph numerical data with [pie charts](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/pie-chart/): **your graphs won’t make any sense**. The following scatter plot illustrates this point. I made a [scatter plot in Microsoft Excel](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/#excel) of categorical data (names) along with their ages in Excel. Excel didn’t recognize the categorical data and assigned numbers instead. The scatter plot is meaningless; no one will know that “1”, “2”, “3”, “4” and “5” refer to names and even if they do…the graph will be a mess if you have 100 names!:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/scatter-plot.jpg)

A workaround to this problem could be to assign numbers to names (e.g. John = 1, Jan = 2…), and include a key on the graph. However in this particular example, a scatter plot really isn’t the best choice for a graph— choose the [bar graph](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/) instead. A bar graph allows you to plot categories on one axis, so the quantitative data condition doesn’t have to be met for one axis.

## **Qualitative Variable (Categorical Variable):**

Watch the video or read on below:

[](https://www.youtube.com/watch?v=zcpRvHU5Ln8)

A qualitative variable, also called a categorical variable, are [variables](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/types-of-variables/)that are **not numerical**. It describes data that fits into categories. For example:

* Eye colors (variables include: blue, green, brown, hazel).
* States (variables include: Florida, New Jersey, Washington).
* Dog breeds (variables include: Alaskan Malamute, German Shepherd, Siberian Husky, Shih tzu).

These are all qualitative variables as they have no natural order. On the other hand, [quantitative variables](https://www.statisticshowto.datasciencecentral.com/what-are-quantitative-variables-and-quantitative-data/) have a value and they can be added, subtracted, divided or multiplied.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/dog.jpg)

*Breeds of dog are qualitative variables. How many dogs are quantitative variables.*

|  |  |
| --- | --- |
| **Quantitative Variable** | **Qualitative Variables** |
| Fractions | Cat breeds |
| Decimals | Cities |
| Odd Numbers | Fast Food Chains |
| Whole Numbers | College Major |
| Irrational Numbers | Fraternities |
| Ordered pairs (x,y) | Hair Color |
| Negative Numbers | Computer Brands |
| Map coordinates | Beer breweries |
| Positive Numbers | Pop music genre |
| Exponents | Tribe |

As a general rule, if you can apply some kind of math (like addition), it’s a [quantitative variable](https://www.statisticshowto.datasciencecentral.com/what-are-quantitative-variables-and-quantitative-data/). Otherwise, it’s qualitative. For example, you can’t add blue+green (unless you’re in an art class — even then you “mix” them, you don’t add them!).

**Numbers are sometimes assigned to qualitative variables** for [**data analysis**](http://www.hindawi.com/journals/jqre/2010/849043/), but they are still classified as qualitative variables despite the numerical classification. For example, a study may assign the number “1” to males and “2” to females.

### **Qualitative Variables and the Nominal Scale**

Qualitative variables aren’t ordered on a numerical scale in statistics so they are assigned [**nominal scales**](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/). The word “nominal” means “name”, which is exactly what qualitative variables are. A nominal scale is a scale where no ordering is possible or implied (except for alphabetical ordering like New York, Washington, West Virginia or Chelsea, Edinburgh, London). In other words, the nominal scale is where data is assigned to a category.

## **Discrete vs Continuous variables**

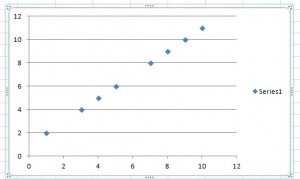
Watch the video, or read the article below:

[](https://www.youtube.com/watch?v=bRr8zDOkh9o)

In a nutshell, discrete variables are points plotted on a chart and a continuous variable can be plotted as a line.

### **What is a Discrete Variable?**

**Discrete variables**are countable in a finite amount of time. For example, you can count the change in your pocket. You can count the money in your bank account. You could also count the amount of money in everyone’s bank accounts. It might take you a long time to count that last item, but the point is—it’s still countable.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/scatter-plot-2.jpg)

*Discrete variables on a scatter plot.*

### **What is a Continuous Variable?**

**Continuous Variables** would (literally) take forever to count. In fact, you would get to “forever” and never finish counting them. For example, take age. You can’t count “age”.**Why not?** Because it would literally take forever. For example, you could be:  
25 years, 10 months, 2 days, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picosends…and so on.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/clock.jpg)

*Time is a continuous variable.*

You could turn age into a discrete variable and then you could count it. For example:

* A person’s age in years.
* A baby’s age in months.

Take a look at this article on [orders of magnitude of time](https://en.wikipedia.org/wiki/Orders_of_magnitude_(time))and you’ll see why time or age just isn’t countable. Try counting your age in Planctoseconds (good luck…see you at the end of time!).

### **Discrete vs Continuous variables: Steps**

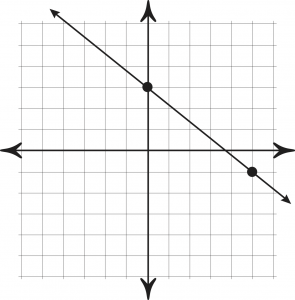
Step 1: Figure out how long it would take you to sit down and **count out** the possible values of your [variable](https://www.statisticshowto.datasciencecentral.com/variable/). For example, if your variable is “Temperature in Arizona,” how long would it take you to write every possible temperature? It would take you literally forever:

50°, 50.1°, 50.11°, 50.111°, 50.1111°, …

If you start counting now and never, ever, ever finish (i.e. the numbers go on and on until infinity), you have what’s called a **continuous variable.**

If your variable is “Number of Planets around a star,” then you can count all of the numbers out (there can’t be an infinite number of planets). That is a **discrete variable**.

Step 2: **Think about “hidden” numbers that you haven’t considered**. For example: is time a discrete or continuous variable? You might think it’s continuous (after all, time goes on forever, right?) but if we’re thinking about numbers on a wristwatch (or a stop watch), those numbers are **limited** by the numbers or number of decimal places that a manufacturer has decided to put into the watch. It’s unlikely that you’ll be given an ambiguous question like this in your elementary stats class but it’s worth thinking about!

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/01/graph-of-4-5x+3.png)

*This graph of -4/5x+3 has continuous variables — it could go on forever…*

## **Type of Statistics:**

1. Descriptive Statistics
2. [Inferential statistics](#Inferential_Statistics)

## **[Descriptive Statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/)**

[Descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/) are one of the fundamental “must know” with any set of data. It gives you a general idea of trends in your data including:

* The [mean, mode, median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/) and [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/).
* [Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/).
* [Skewness](https://www.statisticshowto.datasciencecentral.com/skewness/).
* Count, maximum and minimum.

Descriptive statistics is useful because it allows you to take a large amount of data and summarize it. For example, let’s say you had data on the incomes of one million people. No one is going to want to read a million pieces of data; if they did, they wouldn’t be able to glean any useful information from it. On the other hand, if you summarize it, it becomes useful: an average wage, or a median income, is much easier to understand than reams of data.

## **1. Sub-Areas**

Descriptive statistics can be further broken down into several sub-areas, like:

* [Measures of central tendency.](#_Central_Tendency)
* [Measures of dispersion](https://www.statisticshowto.datasciencecentral.com/dispersion/).
* [Charts & graphs](#_4._Descriptive_Statistics:).
* [Shapes of Distributions.](https://www.statisticshowto.datasciencecentral.com/shapes-of-distributions/)

## 

## [**Central Tendency**](https://www.statisticshowto.datasciencecentral.com/central-tendency-2/)

Central tendency (sometimes called “measures of location,” “central location,” or just “centre”) is a way to describe what’s typical for a set of data. Central tendency doesn’t tell you specifics about the individual pieces of data, but it does give you an overall picture of what is going on in the entire data set. There are three major ways to show central tendency: [mean, mode and median](#Mean_Median_Mode).

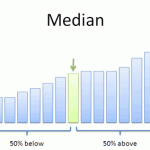
## **Central Tendency Measures**

**Mean**  
**The**[**mean**](#Mean_Average)**is the**[**average**](#Mean_Average)**of a set of numbers.** Add up all the numbers in a set of data and then divide by the number of items in the set. For example, the mean of 2 3 5 9 11 is:  
(2 + 3 + 5 + 9 + 11) / 5 = 30 / 5 = 6.

For more examples of finding the mean, see:  
[What is a mean?](#Mean_Average)

**Median**  
**The**[**median**](#Median)**is the middle of a set of numbers.** Think of it like the median in a road (that grassy area in the middle that separates traffic). Place your data in order, and the number in the exact center of a list is the median. For example:  
1 2 3 **4** 5 6 7  
The median is 4 because it’s in the center, with three numbers either side.

For more about the median, see:  
[What is a median?](#Median)

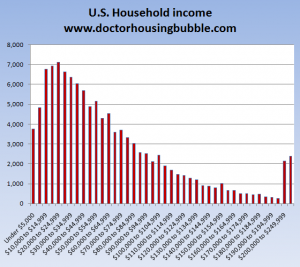
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/median.png)

**Mode**  
The [mode](#Mode) is the most common number in a set of data. For example, the mode of 1 2 2 3 5 6 is 2. Some data sets have no mode, like this one: 1 2 3 4 5 6. Others have multiple modes, like this one: 1 1 2 3 3.

For more on finding modes, see:  
[What is a Mode?](#Median)

**Outliers**  
[Outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/)are extremely high or extremely low values. Outliers can affect central tendency, especially the mean. For example, if you got paid three weeks in a row but took vacation in the fourth week, your pay checks might be: $300 $300 $300 $0. Your four week mean would be ($300 + $300 + $300 + $0) / 4 = $900/4 = $225. That outlier of zero dollars brought your mean down very low.

[Skewed Distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/)is a visual way to show the central tendency of a set of data.



*A left-skewed distribution.*

## **2. Difference Between Descriptive and Inferential Statistics**

Statistics can be broken down into two areas:

* **Descriptive statistics:** describes and summarizes data. You are just describing what the data shows: a trend, a specific feature, or a certain [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)(like a [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)or median).
* [**Inferential statistics**](https://www.statisticshowto.datasciencecentral.com/inferential-statistics/): uses statistics to make predictions.

Descriptive statistics just describes data. For example, descriptive statistics about a college could include: the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)SAT score for incoming freshmen; the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)income of parents; racial makeup of the student body. It says nothing about why the data might exist, or what trends you might be able to see from the data. When you take your data and start to make predictions about future behaviour or trends, that’s inferential statistics. Inferential statistics also allows you to take [sample](https://www.statisticshowto.datasciencecentral.com/sample/)data (e.g. from one university) and apply it to a larger [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/) (e.g. all universities in the country).

## **3.** **Excel Descriptive Statistics**

Using the descriptive statistics feature in Excel means that you won’t have to type in individual functions like MEAN or MODE. One button click will return a dozen different stats for your data set. If you want to calculate Excel descriptive statistics, you must have the Data Analysis Toolpak loaded in Excel. Click the “Data” tab in Excel. If you don’t see “Data analysis” on the right of the toolbar, you need to load the Toolpak first. See: [Load the Excel Data Analysis Toolpak.](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)

### **How to Calculate Excel Descriptive Statistics: Steps**

[Watch the video](https://www.youtube.com/watch?v=ZdcoTVYJNF4) or read the steps below:

Step 1:**Type your data into Excel,** in a single column. For example, if you have ten items in your data set, type them into cells A1 through A10.

Step 2:**Click the “Data” tab**and then click “Data Analysis” in the Analysis group.

Step 3:**Highlight “Descriptive Statistics”**in the pop-up Data Analysis window.

Step 4:**Type an input range into the “Input Range” text box.**For this example, type “A1:A10” into the box.

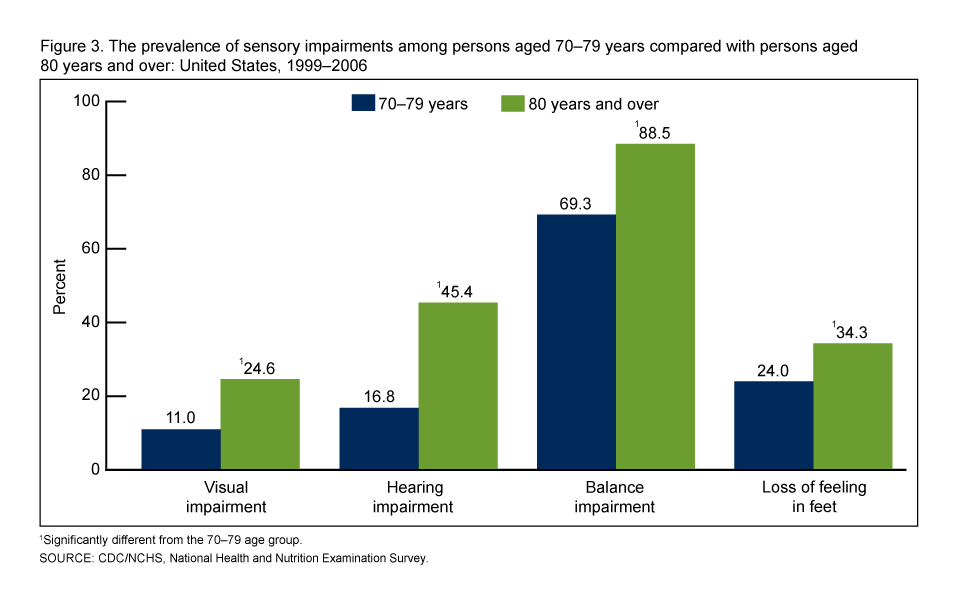
Step 5:**Check the “Labels in first row” check box**if you have titled the column in row 1, otherwise leave the box unchecked.

Step 6:**Type a cell location into the “Output Range” box.** For example, type “C1.” Make sure that two adjacent columns do not have data in them.

Step 7:**Click the “Summary Statistics” check box and then click “OK”**to display Excel descriptive statistics. A list of descriptive statistics will be returned in the column you selected as the Output Range.

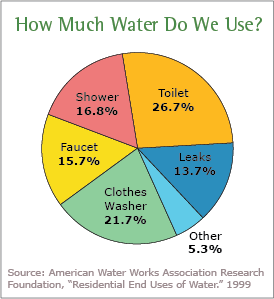
## **4. Descriptive Statistics: Charts, Graphs and Plots**

There are literally dozens of charts and graphs you can make from data. which one you choose depends upon what kind of data you have and what you want to display. For example, if you wanted to display relationships between data in categories, you could make a [bar graph.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/)



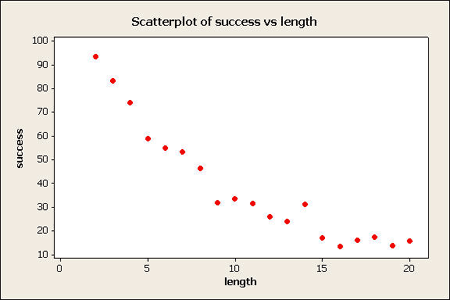
*Grouped bar graph. Image: CDC.*

A [pie chart](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/pie-chart/) would show you how categories in your data relate to the whole set.

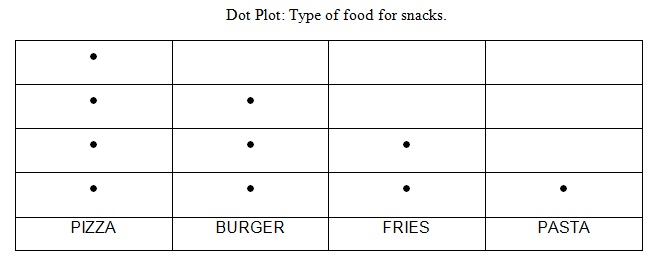


*Pie chart showing water consumption. Image courtesy of EPA.*

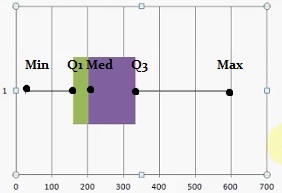
[Scatter plots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/#definition) are a good way to display data points.

 *Image: Penn State*

Less common, but useful in some cases, include [dot plots](https://www.statisticshowto.datasciencecentral.com/what-is-a-dot-plot/) and [box and whisker charts](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/#definition):

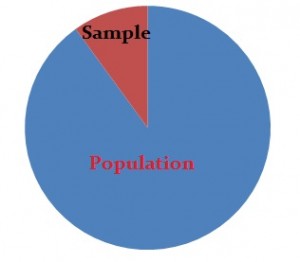
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/10/dot-plot-2.jpg)

*Simple dot plot showing the types of foods a group of friends eats.*

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/11/box-and-whiskers-graph-.jpg)

*Box and whiskers graph*

## **Inferential Statistics**



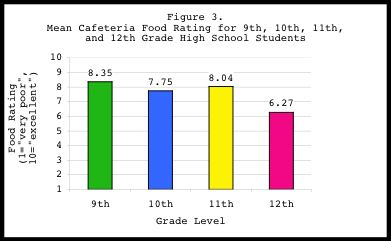
[Descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/) describes data (for example, a chart or graph) and **inferential statistics** allows you to make predictions (“inferences”) from that data. With inferential statistics, you take data from [samples](https://www.statisticshowto.datasciencecentral.com/sample/)and make generalizations about a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, you might stand in a mall and ask a sample of 100 people if they like shopping at [Sears](http://www.sears.com/). You could make a [bar chart](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/) of yes or no answers (that would be [descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/)) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (**all**shoppers in **all malls**) like shopping at Sears.

There are two main areas of inferential statistics:

1. **Estimating parameters**. This means taking a [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)from your sample data (for example the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)) and using it to say something about a population parameter (i.e. the population mean).
2. [**Hypothesis tests**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/). This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

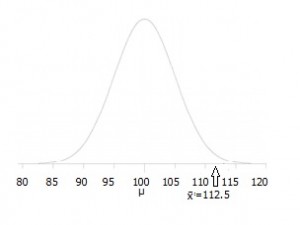
Let’s say you have some sample data about a potential new cancer drug. You could use descriptive statistics to describe your sample, including:

* Sample [mean](https://www.statisticshowto.datasciencecentral.com/mean/)
* Sample [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)
* Making a [bar chart](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/) or [boxplot](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/)
* Describing the shape of the sample [probability distribution](https://www.statisticshowto.datasciencecentral.com/probability-distribution/)



*A bar graph is one way to summarize data in descriptive statistics. Source: NIH.GOV.*

With inferential statistics you take that sample data from a small number of people and try to determine if the data can predict whether the drug will work for everyone (i.e. the population). There are various ways you can do this, from calculating a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) (z-scores are a way to show where your data would lie in a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) to [post-hoc](https://www.statisticshowto.datasciencecentral.com/post-hoc/) (advanced) testing.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/10/hypothesis-testing-example.jpg)

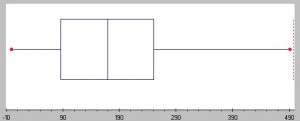
*A hypothesis test can show where your data is placed on a distribution like this one.*

Inferential statistics use statistical models to help you compare your sample data to other samples or to previous research. Most research uses statistical models called the Generalized Linear model and include [Student’s t-tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/), [ANOVA (Analysis of Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/anova/)), [regression](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/)analysis and various other models that result in straight-line (“linear”) probabilities and results.

## **Dispersion / Measures of Dispersion**

Dispersion in statistics is a way of describing how spread out a set of data is. When a data set has a large value, the values in the set are **widely scattered**; when it is small the items in the set are **tightly clustered**. Very basically, this set of data has a small value:  
1, 2, 2, 3, 3, 4  
…and this set has a wider one:  
0, 1, 20, 30, 40, 100

The spread of a data set can be **described by a range of descriptive statistics** including [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/), [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviationon/), and [interquartile range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/). Spread can also be shown in graphs: [dot plots](https://www.statisticshowto.datasciencecentral.com/what-is-a-dot-plot/), [boxplots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/), and [stem and leaf plots](https://www.statisticshowto.datasciencecentral.com/stemplot/) have a greater distance with samples that have a larger dispersion and vice versa.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/right-skewed-box-plot.jpg)

*The larger the box, the more dispersion in a set of data. Image: Seton Hall University*

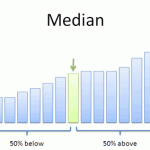
### **Measures of Dispersion**

* [**Coefficient of dispersion**](https://www.statisticshowto.datasciencecentral.com/coefficient-of-dispersion/): A “catch-all” term for a variety of formulas, including distance between [quartiles](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/).
* [**Standard deviation**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/): probably the most common measure. It tells you how spread out numbers are from the mean.
* [**Index of Dispersion**](https://www.statisticshowto.datasciencecentral.com/index-of-dispersion/): a measure of dispersion commonly used with [nominal variables](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/).
* [**Interquartile range (IQR)**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/): describes where the bulk of the data lies (the “[middle fifty](https://www.statisticshowto.datasciencecentral.com/middle-fifty/)” percent).
* [**Interdecile range**](https://www.statisticshowto.datasciencecentral.com/interdecile-range/): the difference between the first [decile](https://www.statisticshowto.datasciencecentral.com/decile/)(10%) and the last decile (90%).
* [**Range**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/): the difference between the smallest and largest number in a set of data.
* [**Mean difference**](https://www.statisticshowto.datasciencecentral.com/mean-difference/)**or difference in means:** measures the absolute difference between the mean value in two different groups in clinical trials.
* [**Median absolute deviation**](https://www.statisticshowto.datasciencecentral.com/median-absolute-deviation/)**(MAD)**: the median of the absolute deviations from a data set’s median.
* [**Quartiles**](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/): Numbers that split the data into four quarters (first, second, third, and fourth quartiles).

In some processes, like manufacturing or measurement, low dispersion is associated with high precision. High dispersion is associated with low precision.

### **Measures of Dispersion: Example**

Let’s say you were asked to compare measures of dispersion for two data sets. Data set A has the items 97,98,99,100,101,102,103 and data set B has items 70,80,90,100,110,120,130. By looking at the data sets you can probably tell that the means and medians are the same (100) which technically are called “measures of central tendency” in statistics.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/median.png)

However, the range (which gives you an idea of how spread out the entire set of data is) is much larger for data set B (60) when compared to data set A (6). In fact, nearly all measures of dispersion would be ten times greater for data set B, which makes sense as the range is ten times larger. For example, take a look at the standard deviations for the two data sets:

Standard deviation for A: 2.160246899469287.  
Standard deviation for B: 21.602468994692867.  
The figure for data set B is exactly ten times that of A.

## **Shapes of Distributions**

When a data set is graphed, each point is arranged to produce one of**dozens of different shapes.** The distribution shape can give you a visual which helps to show how the data is:

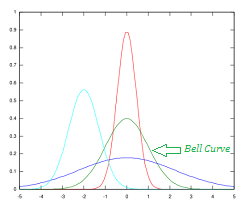
* Spread out (e.g. [dispersion](https://www.statisticshowto.datasciencecentral.com/dispersion/), [variability](https://www.statisticshowto.datasciencecentral.com/variability/), [scatter](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/)),
* Where the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)lies,
* What the [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/)of the data set is,

…and many other useful [statistics](https://www.statisticshowto.datasciencecentral.com/statistic/). Shapes of distributions are defined by several different factors:

### **1. Number of peaks**

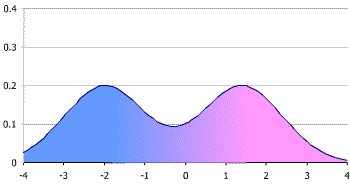
The peaks are usually called [modes](https://www.statisticshowto.datasciencecentral.com/mode/); Sometimes they are called “local maximums”, but that term is more commonly used in [calculus](https://calculushowto.com/). The mode tells you that the data count is higher in these areas than in any other areas on the graph.

* A [**unimodal**distribution](https://www.statisticshowto.datasciencecentral.com/unimodal-distribution-2/) has **one**mode. A single peak can take on many shapes (e.g. very tall and thin or very squat and fat). One of the most common types of unimodal distributions is the [***normal distribution***](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), sometimes called the “[bell curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)” because its shape looks like a bell.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/08/unimodal-small-2.png)

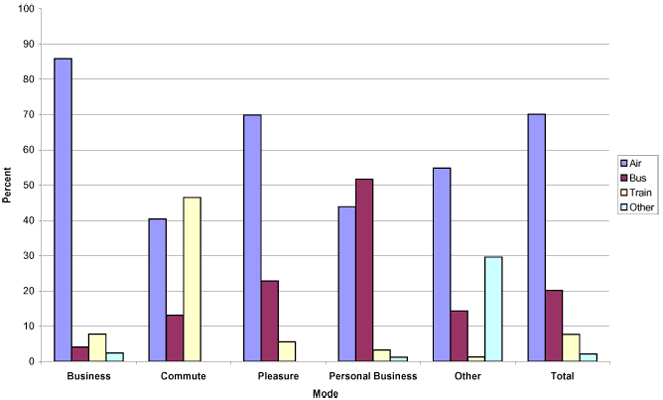
*Several unimodal distributions plotted on the same graph. The green “bell curve” is the normal distribution.*

* A [**bimodal**distribution](https://www.statisticshowto.datasciencecentral.com/what-is-a-bimodal-distribution/) has **two**modes.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/07/Bimodal.png)

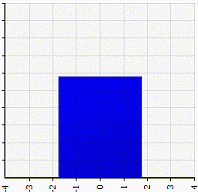
*Bimodal distribution.*

* A [**multimodal**distribution](https://www.statisticshowto.datasciencecentral.com/multimodal-distribution/) has **three or more** modes.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/multimodal-distribution.gif)

*Multimodal distribution with many modes, or peaks. Image credit: USDOT*

* If a data set has no clear peaks (i.e. the whole graph looks flat), it’s called a “[***uniform distribution****.*](https://www.statisticshowto.datasciencecentral.com/uniform-distribution/)”

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/shape_uniform.gif)

*A uniform distribution.*

### **Terminology Note**

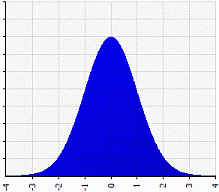
One of the most confusing aspects about statistics is the terminology, and the “mode” is no exception. It may seem at first glance that the word has two meanings:

* The most common number in a set. For example, the mode of 1, 2, 2, 3, 5 is the number “2”.
* A word to describe a peak on a graph.

However, the two terms actually mean the**same thing**. The most common number in the above list of numbers is “2”. If you were to plot that set of numbers on a graph, the peak would happen at x = 2, which means that the peak contains the most common number in the set. That said, if you are asked to “[Find the Mode](https://www.statisticshowto.datasciencecentral.com/mode/)“, that usually indicates you should look for the most frequent item in the set; it doesn’t mean you have to create a graph.

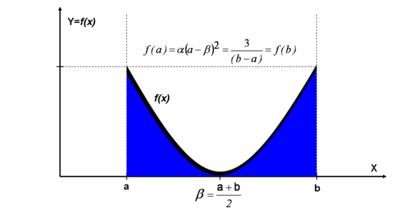
### **2. Symmetry**

A [symmetric graph](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/) has two sides that are mirror images of each other. The [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) is one example of a symmetric graph.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/shape_normal.gif)

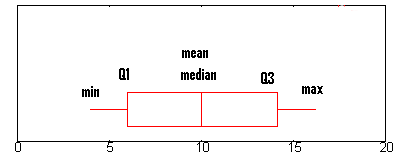
*The normal distribution.*

Another type of symmetric graph is the U-distribution, which—perhaps not surprisingly— looks like the letter “U”.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/u-dsit.jpg)

*Quadratic U-Distribution. Image credit: UCLA*

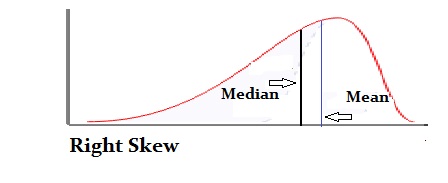
A symmetric box plot has the “**box**” in the **center** of the graph:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/03/normal-boxplot.png)

*A symmetric box plot.*

### **3. Skewness**

Shapes of distributions can differ in [skewness;](https://www.statisticshowto.datasciencecentral.com/skewness/) these distributions are not symmetrical. Instead, they have more points plotted on one side of the mean than on the other. This causes [long tails](https://www.statisticshowto.datasciencecentral.com/long-tail-distribution/) either in the negative direction on the number line (a negative, or left skew) or in the positive direction on the [number line](http://www.math.utah.edu/online/1010/line/) (a positive, or right skew). For more on how skewness affects shapes of distributions, see: [Skewed Distribution in Statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/left-skew-2.jpg)

*A Left-skewed, negative distribution with a long tail in the negative direction of the number line.*

**The tails of a distribution (i.e. how thin or fat they are) can also be described by**[**kurtosis**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/), which is measured against the [standard normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/#SNM). A positive value for kurtosis means you have a large peak and little data in the tails. A negative value means you have a flattened peak with lots of data in the tails.

## **[Mean, Median, Mode](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/):**

1. The **mean**is the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)of a data set.
2. The **mode**is the most common number in a data set.
3. The **median**is the middle of the set of numbers.

## **Mean vs Average:**

When you first started out in mathematics, you were probably taught that an [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)was a “middling” amount for a set of numbers. But in studying statistics and all of a sudden the “average” is now called the mean. What happened? The answer is that they are exactly the same word (they are synonyms). You added up the numbers, divided by the number of items you get the average. For example, the average of 10, 5 and 20 is:  
10 + 6 + 20 = 36 / 3 = 12.

That said, technically, the word mean is short for the [arithmetic mean](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/). We use different words in stats, because there are multiple different [types of means](#_Other_Types), and they all do different things.

When someone talks about the mean of a data set, they are usually talking about the arithmetic mean (most people just drop the word “arithmetic”). It’s called a different name to set it apart from other means found in math, including the [geometric mean](https://www.statisticshowto.datasciencecentral.com/geometric-mean-2/).

The mean is influenced by [outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/), so it isn’t always a good indicator of where the middle of a data set is. For data sets that have either a lot of low values or a lot of high values, the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)is often a better way to describe the “middle.”

## **Population vs. Sample Mean**

If your data is a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/), then the mean is called a [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/), represented by the letter μ. If the list is a [sample](https://www.statisticshowto.datasciencecentral.com/sample/), it’s called a [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)x̄.

## **Specific “Means” commonly used in Stats**

You’ll probably come across these in your stats class. They have very narrow meanings:

* [Mean of the sampling distribution](https://www.statisticshowto.datasciencecentral.com/sampling-distribution/#MeanSDM): used with [probability distributions](https://www.statisticshowto.datasciencecentral.com/probability-distribution/), especially with the [Central Limit Theorem](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/central-limit-theorem-definition-examples/). It’s an average of a set of distributions.
* [Sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/): the average value in a [sample](https://www.statisticshowto.datasciencecentral.com/sample/).
* [Population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/): the average value in a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/).

## **Other Types**

There are other types of means, and you’ll use them in various branches of math. Most have very narrow applications to fields like finance or physics; if you’re in elementary statistics you probably won’t work with them.

These are some of the most common types you’ll come across.

1. [Weighted mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#weighted)
2. [Harmonic mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#harmonic)
3. [Geometric mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#geometric)
4. [Arithmetic-Geometric mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#arithmetic)
5. [Root-Mean Square mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#root)
6. [Heronian mean.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#heronian)

## **Mean vs Median**

Both are measures of where the center of a data set lies, but they are usually different numbers. For example, take this list of numbers: 10,10,20,40,70.

* The mean (average) is found by adding all of the numbers together and dividing by the number of items in the set: 10 + 10 + 20 + 40 + 70 / 5 = 30.
* The median is found by ordering the set from lowest to highest and finding the exact middle. The median is just the middle number: 20.

Sometimes the two will be the same number. For example, the data set 1,2,4,6,7 has an average of 1 + 2 + 4 + 6 + 7 / 5 = 4 and a median (a middle) of 4.

## **Median**

The [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median) is the **middle number**in a data set. To find the median, list your data points in ascending order and then find the middle number. The middle number in this set is 28 as there are 4 numbers below it and 4 numbers above:  
23, 24, 26, 26, 28, 29, 30, 31, 33

**Note**: If you have an even set of numbers, average the middle two to find the mean. For example, the mean of this set of numbers is 28.5 (28 + 29 / 2).  
23, 24, 26, 26, 28, 29, 30, 31, 33, 34

## **Mode**

The[mode](https://www.statisticshowto.datasciencecentral.com/mode/) is the most **common number** in a set. For example, the mode in this set of numbers is 21:  
21, 21, 21, 23, 24, 26, 26, 28, 29, 30, 31, 33

## **SPSS Mean mode median**

In order to find the SPSS mean mode median, you’ll need to use the **Frequency tab**. It seems a little counter-intuitive, but the Descriptive Statistics tab does not give you the option to find the mode or the median.

SPSS has a very similar interface to Microsoft Excel. Therefore, if you’ve used Microsoft Excel before, you will quickly adapt to SPSS.

## **SPSS Mean Mode Median: Steps**

[Click to watch the video](https://youtu.be/-LJ2ymPbuVs) or read the steps below:

**Sample question:** Find the SPSS mean mode median for the following data set: 20,23,35,66,55,66

Step 1: **Open SPSS.** In the “What would you like to do?” dialog box, click the “type in data” radio button and then click “OK.” A new worksheet will open. Note: If you have opted out of the first help screen, you may not see this option. In that case, just start at Step 2.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/spss-mean-1.jpg)

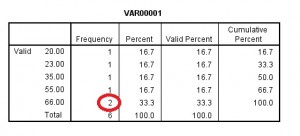
Step 2: **Type your data into the worksheet.** You can type the data into one column or multiple columns if you have multiple data sets. For this example, type 20, 23, 35, 66, 55, 66 into column 1. Do not leave spaces between the data (i.e. don’t leave any empty rows).

Step 3: **Click “Analyze,” hover over “Descriptive Statistics” and then click “Frequencies.”**

Step 3: **Click “Statistics” and then check the boxes “mean”, “mode” and “median.”** Click “Continue” twice (select “none” as the chart type in the second window).

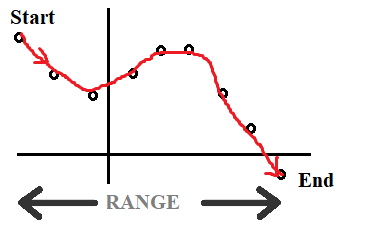
**Note**: In some versions of SPSS, you may only have to click “Continue” once and it may not give you an option for chart type.

The frequency results will appear as output. The top part of the output will display the mean, mode and median.

If you scroll down, the frequency table will also show you the mode. The mode is defined in statistics as the number with the highest frequency (for this sample data set, the number appearing the most is 66, with two results in the frequency column).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/spss-mean-6.jpg)

## **Range**

In statistics, the **range** is a [measure of spread](https://www.statisticshowto.datasciencecentral.com/measures-of-spread/): it’s the difference between the highest value and the lowest value in a data set.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2018/08/range-statistics.png)  
  
  
**Note**: In some areas of math, the [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/)can also mean the entire range of numbers — for example, the range of cell phone prices might be $40 to $550. In [calculus](https://calculushowto.com/), the range is defined differently. It is all of the output values of a function. See: [How to Find the Domain and Range of a Function](https://calculushowto.com/domain-and-range-of-a-function/).

### **How to Find a Range in Statistics**

Watch the video, or read the article below:

[](https://www.youtube.com/watch?v=La43tpiK0Z4)

The same two steps are used whether you are dealing with positive numbers, negative numbers, or time (e.g. seconds or minutes).

### **How to Find a Range**

**Example question 1**: What is the range for the following set of numbers? 10, 99, 87, 45, 67, 43, 45, 33, 21, 7, 65, 98?

Step 1: **Sort the numbers in order**, from smallest to largest:  
7, 10, 21, 33, 43, 45, 45, 65, 67, 87, 98, 99

Step 2: **Subtract the smallest number in the set from the largest number in the set**:  
99 – 7 = 92  
The range is 92

That’s it!

**Example question 2:** What is the range of these [integers](https://www.statisticshowto.datasciencecentral.com/integer/)?  
14, -12, 7, 0, -5, -8, 17, -11, 19

Step 1: **Sort the numbers in order, from smallest to largest**:  
-12, -11, -8, -5, 0, 7, 14, 17, 19

Step 2: **Subtract the smallest number in the set from the largest number in the set**:  
19 – -12 = 19 + 12 = 31  
The range is 31.

That’s it!

**Example question 3:** What is the range of the following times?  
2.7 hrs, 8.3 hrs, 3.5 hrs, 5.1 hrs, 4.9 hrs

Step 1: **Sort the numbers in order, from smallest to largest**:  
2.7, 3.5, 4.9, 5.1, 8.3

Step 2: **Subtract the smallest number in the set from the largest number in the set**:  
8.3 hr – 2.7 hr = 5.6 hr  
The range is 5.6 hr.

That’s how to find a range!

### **Another Example.**

**Problem**: You take 7 statistics tests over the course of a semester. You score 94, 88, 73, 84, 91, 87, and 79. What is the range of your scores?  
**Solution:**  
Step 1: Order your scores from smallest to largest:  
73, 79, 84, 87, 88, 91, 94.  
Step 2: Subtract the smallest number from the highest = 94 – 73 = 21.  
**Answer: 21.**

### **When it Might be Misleading**

The range only uses the smallest and the largest number in a set; The rest of the values are ignored. That could lead to a misleading result. Take the above test scores. Let’s say you had the flu one test day and scored a 10. Assuming your highest score on another test was 94, then:  
94 – 10 = 84!  
That’s not a good reflection of your overall test performance at all.

The score of 10 in the example above is what we call an [outlier](https://www.statisticshowto.datasciencecentral.com/find-outliers/). It’s an extremely high or low value that can throw off stats. That’s why other [measures of spread](https://www.statisticshowto.datasciencecentral.com/measures-of-spread/) are sometimes preferred, like the [mean](https://www.statisticshowto.datasciencecentral.com/mean/).

### **Rule of Thumb**

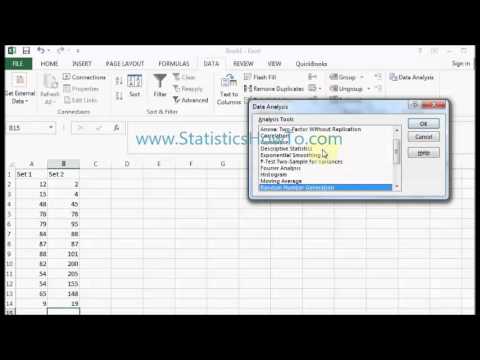
The rule of thumb says that the range is about four times the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). The standard deviation is another measure of spread in statistics. It tells you how your data is clustered around the [mean](https://www.statisticshowto.datasciencecentral.com/mean/). What the rule of thumb tells you in most cases is that the bulk of the data can be found pretty close to the mean (within a couple of standard deviations); The result is that those erroneous “outliers” should have very little effect on your final statistic.

Procedure for finding a standard deviation using the rule of thumb:  
Step 1: Find the range.  
Step 2: Divide Step 1 by four.

The rule of thumb doesn’t work that well for small data sets. And it doesn’t work at all if you don’t have data that fits a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/). That’s why you’ll rarely see it used in statistics. See: [Range rule of thumb.](https://www.statisticshowto.datasciencecentral.com/range-rule-of-thumb/)

### **Range in Excel 2013-2016**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=CeWwXk1XGuA)

To find a range in Excel, you have two options: you can use the MAX and MIN functions to find the largest and smallest numbers in a data set and then you can subtract the two. For example, if you had a data set in cells A1 to A10, you’d need three formulas in three blank cells. Lastly the format (assuming you put these formulas into cells B1:B3) would be:

B1 = MAX(A1:A10)  
B2 = MIN(A1:A10)  
B3 =(B1-B2)

A much easier way is to use Data Analysis, where in just a couple of clicks (with no entering formulas) you can display a variety of summary statistics, including the range (How to [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)).

### **Range in Excel: Data Analysis Steps**

1st: Click the “Data” tab and then click “Data Analysis.”

2nd: Click “Descriptive Statistics” and then click “OK.”

3rd: Click the Input Range box and then type the location for your data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box

4th: Click the radio button for Rows or Columns, depending on how your data is laid out.

5th: Click the “Labels in first row” box if your data has column headers.

6th: Click the “Descriptive Statistics” check box.

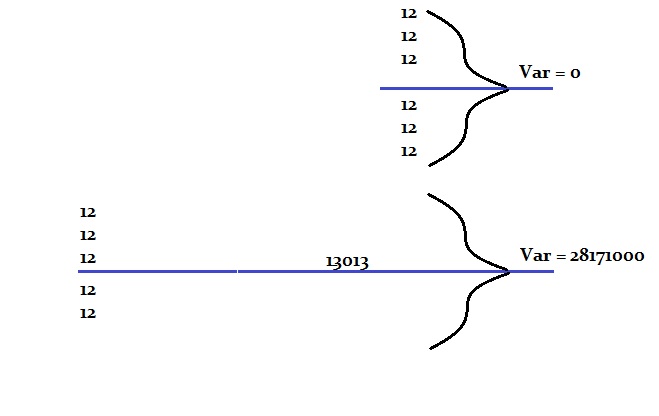
7th: Select a location for your output. For example, click the “New Worksheet” radio button.

8th: Click “OK.”

## **Variance**

Variance measures how far a data set is spread out. The technical definition is “The [*average*](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)of the squared differences from the [*mean*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean),” but all it really does is to give you a very **general idea of the spread of your data**. A value of zero means that there is no variability; All the numbers in the data set are the same.

* The data set 12, 12, 12, 12, 12 has a var. of zero (the numbers are identical).
* The data set 12, 12, 12, 12, **13** has a var. of 0.167; a small change in the numbers equals a very small var.
* The data set 12, 12, 12, 12, **13,013** has a var. of 28171000; a large change in the numbers equals a very large number.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/what-is-variance.jpg)

*How varied is your data set?*

### **How Much Can Data Vary?**

The smallest a variance gets is zero, but technically, it can be **infinite**with numbers in the millions or even billions and beyond.

### **How do I calculate it?**

[Watch the video](https://youtu.be/9rCLEEpyxHQ)

The variance for a **population** is calculated by:

1. Finding the mean(the average).
2. Subtracting the mean from each number in the data set and then squaring the result. The results are squared to make the negatives positive. Otherwise negative numbers would cancel out the positives in the next step. It’s the distance from the mean that’s important, not positive or negative numbers.
3. Averaging the squared differences.

However, it’s more usual in statistics to find the variance for a [**sample**](https://www.statisticshowto.datasciencecentral.com/sample/). When you calculate it for a sample, divide by the [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) minus one ([*Why use n-1?*](https://www.statisticshowto.datasciencecentral.com/bessels-correction/)) when calculating the average squared difference in Step 3 above. See: [Finding Sample Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/sample-variance/).

### [**Standard Deviation**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)

The square root of the variance is the[standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). While var. gives you a rough idea of spread, the standard deviation is more concrete, giving you exact distances from the mean.

### **Variance of a Binomial Distribution**

A binomial distribution is a simple experiment where there is “success” or “failure.” For example, choosing a winning lottery ticket could be a [binomial experiment](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-experiment/) (you either win or lose!). Tossing a coin to try and get heads is also binomial (with tossing a heads being a “success” and a tails a “failure”). The formula for the variance of binomial distribution is **n\*p (1-p) or n\*p\*q.** The two formulas are equivalent because q = (1-p).

**Sample problem:**If you flip a coin 50 times and try to get heads, what is the variance of binomial distribution?

Step 1: Find “p”. The first step to solving this problem is to realize that the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/)of getting a heads is 50 percent, or .5. Therefore, “p” (the probability) is .5.

Step 2: Find “q”, or 1-p. These two are equivalent. They are the probability of not getting a heads (in other words, the probability of getting a tails). 1 – 0.5 = 0.5. Therefore, “q” (or 1 – p) = 0.5.

Step 3: Multiply Step 1 (p) by Step 2 (q) by “n” (the number of trials). We are flipping the coin 50 times, so the number of trials is 50 (n = 50).

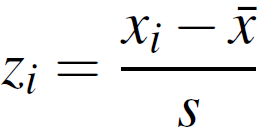
N \* p \* q = 50 \* .5 \* .5 = 12.5.

The var. of binomial distribution for flipping a coin 50 times is 12.5.

**OK, So what does the Binomial Variance mean?**

In essence, not a lot! The variance isn’t used for much at all, except for calculating standard deviation. For example, the standard deviation for this particular binomial distribution is:  
√12.5 = 3.54.

You’ll use the variance for things like calculating [z-scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)(this typically comes later in a stats class, after [normal distributions](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)), which has a standard deviation in the bottom of the formula:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/z-score.png)

*Alternate form of the z score.*

### **Population Variance**

The[population variance](https://www.statisticshowto.datasciencecentral.com/population-variance/)is a type of [**parameter**](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/). If you aren’t sure what a parameter is, you may want to review:  
[What is the Difference Between a Statistic and a Parameter](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/)?

The formula is:  
[population variance formula](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/03/population-variance-formula.jpg)  
[Watch the video](https://youtu.be/9rCLEEpyxHQ) to learn how to find the population variance or read the steps below:

**How to find the Population Variance**

Most of the time in statistics, you’ll want to find the [sample variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/sample-variance/), not the population variance. Why? Because statistics is usually all about making inferences from [samples](https://www.statisticshowto.datasciencecentral.com/sample/), not [populations](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). If you had all of the data from a population, there would be no need for statistics at all! That said, there really is very little difference between the formula for the population variance and the formula for the sample variance. If you have [sample](https://www.statisticshowto.datasciencecentral.com/sample/)data, you can still use this formula. You’d just need to insert your data into the columns instead of your population data. If you prefer to plug the numbers straight into the formula, just make sure you use the [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/) and not the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)([xbar](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/09/xbar.bmp)). In addition, the most common sample variance formula uses n-1 in the [denominator](http://www.merriam-webster.com/dictionary/denominator)instead of n.

**Sample problem:** Find the population variance for the following set of numbers: 28, 29, 30, 31, 32.

Step 1: Draw a table. Label the columns as shown and then write down your X values (the items in your [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)) in column 1:

|  |  |  |
| --- | --- | --- |
| X | X-μ | (X-μ)^2 |
| 28 |  |  |
| 29 |  |  |
| 30 |  |  |
| 31 |  |  |
| 32 |  |  |

Step 2: Find the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean). The mean for this set of data is (28 + 29 + 30 + 31 + 32) / 5 = 30.

Step 3: Fill in column 2. This column is your X value minus the mean. For example, the first entry is 28 – 30 = -2.

|  |  |  |
| --- | --- | --- |
| X | X-μ | (X-μ)^2 |
| 28 | -2 |  |
| 29 | -1 |  |
| 30 | 0 |  |
| 31 | 1 |  |
| 32 | 2 |  |

Step 4: Square the values from Step 3 and place those squares in the third column:

|  |  |  |
| --- | --- | --- |
| X | X-μ | (X-μ)^2 |
| 28 | -2 | 4 |
| 29 | -1 | 1 |
| 30 | 0 | 0 |
| 31 | 1 | 1 |
| 32 | 2 | 4 |

Step 5: Add up all of the numbers in column 3 (this is the summation Σ part of the formula):  
4 + 1 + 0 + 1 + 4 = 10

Step 6: Divide by the number of items in your data set:  
10 / 5 = 2  
The population variance for this set of data is 2.

## **Sample Variance**

The sample variance, s2, is used to calculate how varied a sample is. A sample is a select number of items taken from a population. For example, if you are measuring American people’s weights, it wouldn’t be feasible (from either a time or a monetary standpoint) for you to measure the weights of every person in the population. The solution is to take a sample of the population, say 1000 people, and use that [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) to estimate the actual weights of the whole population. The variance helps you to figure out how spread out your weights are.

[](https://www.statisticshowto.datasciencecentral.com/what-is-sample-variance/)

*Body types are varied — they come in all shapes and sizes.*

### **Definition of Sample Variance**

The variance is mathematically defined as **the average of the squared differences from the mean**. But what does that actually mean in English? In order to understand what you are calculating with the variance, break it down into steps:

* Step 1: Calculate the mean (the average weight).
* Step 2: Subtract the mean and square the result.
* Step 3: Work out the average of those differences.

### **What is the sample variance used for?**

While the variance is useful in a mathematical sense, it won’t actually give you any information that you can use. For example, if you take a sample population of weights, you might end up with a variance of 9801. That might leave you scratching your head about why you’re calculating it in the first place! The answer is, you can use the variance to figure out the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) — a much better measure of how spread out your weights are. In order to get the standard deviation, take the square root of the sample variance:  
√9801 = 99.

The standard deviation, in combination with the mean, will tell you what the majority of people weigh. For example, if your mean is 150 pounds and your standard deviation is 99 pounds, the majority of people weigh between 51 pounds (mean-99) and 249 pounds (mean+99).

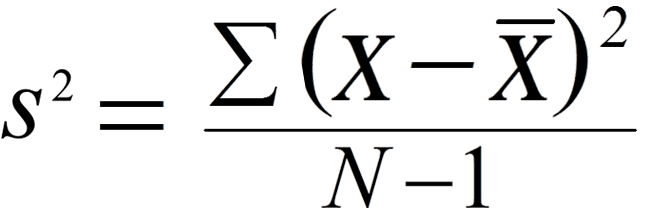
### **Calculating Sample Variance**

[Watch the video for an example](https://youtu.be/9rCLEEpyxHQ) or read on below for several more examples of how to find the sample variance:

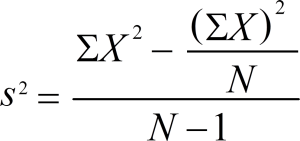
The variance formula can be tricky to use—especially if you are rusty on [order of operations](https://www.statisticshowto.datasciencecentral.com/statistics-basics/statistics-operations-pemdas-bedmas/).

### **How to Find the Sample Variance**

If you’re finding the sample variance by hand, the “usual” formula you’re given in textbooks is:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/usual.png)

However, if you’re working the formula by hand, it can be a bit cumbersome. An alternative version is the computational formula, which can be easier to work:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/computational.png)

### **How to Find the Sample Variance by Hand: Variance Example 1**

**Question**: Find the variance for the following set of data representing trees in California (heights in feet): 3, 21, 98, 203, 17, 9

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/tree.jpg)

Step 1: Add up the numbers in your given data set.

3 + 21 + 98 + 203 + 17 + 9 = 351

Step 2: Square your answer:

351 × 351 = 123,201

…and divide by the number of items. We have 6 items in our example so:

123,201 / 6 = 20,533.5

Set this number aside for a moment.

Step 3: Take your set of original numbers from Step 1, and square them individually this time:

3 × 3 + 21 × 21 + 98 × 98 + 203 × 203 + 17 × 17 + 9 × 9

Add those numbers (the squares) together:

9 + 441 + 9604 + 41209 + 289 + 81 = 51,633

Step 4: Subtract the amount in Step 2 from the amount in Step 3.

51,633 – 20,533.5 = 31,099.5

Set this number aside for a moment.

Step 5: Subtract 1 from the number of items in your data set\*. For our example:

6 – 1 = 5

Step 6: Divide the number in Step 4 by the number in Step 5. This gives you the **variance**:

31,099.5 / 5 = 6,219.9

### **How to find the sample variance: Standard Deviation Example 1**

Step 7: Take the square root of your answer from Step 6. This gives you the **standard deviation**:

√6,219.9 = 78.86634

That’s it!

**\*Important note:**The [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) formula is **slightly different**for [populations](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)and [samples](https://www.statisticshowto.datasciencecentral.com/sample/)(a portion of the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)). If you have a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/), you’ll be dividing by “n” (the number of items in your data set). However, if you have a sample (which is the case for most statistics questions you’ll get in class!) you’ll need to divide by n-1. For why n-1 is used, see: [Bessel’s Correction](https://www.statisticshowto.datasciencecentral.com/bessels-correction/).

### **How to Find the Sample Variance: Example 2**

Your paychecks for the last few weeks are: $600, $470, $430, $300 and $170. What is the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)?

Step 1: Add up all of the numbers:  
170 + 300 + 430 + 470 + 600 = 1970

Step 2: Square the total, and then divide by the number of items in the data set  
1970 x 1970 = 3880900  
3880900 / 5 = 776180

Step 3: Take your set of original numbers from step 1, and square them individually this time. Then add them all up:  
(170 x 170) + (300 x 300) + (430 x 430) + (470 x 470) + (600 x 600) = 884700

Step 4: Subtract the amount in step 2 from the amount in step 3:  
884700 – 776180 = 108520

Step 5: I subtracted 1 from the number of items in my data set:  
5 – 1 = 4

Step 6: Divide the number in step 4 by the number in step 5:  
108520 / 4 = 27130  
This is my [Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)!

Step 7: Take the square root of the number from step 6 (the [Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)),  
√(27130) = 164.7118696390761  
This is my [Standard Deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)!

### **How to Find the Sample Variance: Example 3**

This example uses the same formula, it’s just a slightly different way of working it.

You survey households in your area to find the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)rent they are paying. Find the standard deviation from the following data:  
$1550, $1700, $900, $850, $1000, $950.

Step 1: Find the [mean](https://www.statisticshowto.datasciencecentral.com/mean):  
($1550 + $1700 + $900 + $850 + $1000 + $950)/6 = $1158.33

Step 2: Subtract the [mean](https://www.statisticshowto.datasciencecentral.com/mean)from each value. This gives you the differences:  
$1550 – $1158.33 = $391.67  
$1700 – $1158.33 = $541.67  
$900 – $1158.33 = -$258.33  
$850 – $1158.33 = -$308.33  
$1000 – $1158.33 = $158.33  
$950 – $1158.33 = $208.33

Step 3: Square the differences you found in Step 3:  
$391.672 = 153405.3889  
$541.672 = 293406.3889  
-$258.332 = 66734.3889  
-$308.332 = 95067.3889  
$158.332 = 25068.3889  
$208.332 = 43401.3889

Step 4: Add up all of the squares you found in Step 3 and divide by 5 (which is 6 – 1):  
(153405.3889 + 293406.3889 + 66734.3889 + 95067.3889 + 25068.3889 + 43401.3889) / 5 = 135416.66668

Step 5: Find the square root of the number you found in Step 4 (the [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)):  
√135416.66668 = 367.99  
The [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) is 367.99.

### **How to Find Sample Variance Example 4: Steps**

**Sample Question:**Find sample variance / standard deviation for the following data set: 1245, 1255, 1654, 1547, 1787, 1989, 1878, 2011, 2145, 2545, 2656.

Step 1: **Add up all of the numbers in your data set**:  
1245 + 1255 + 1547 + 1654 + 1787 + 1878 + 1989 + 2011 + 2145 + 2545 + 2656 = 20712

Step 2: **Square the number you found in Step 1:**  
20712 x 20712 = 428986944  
…and then divide by the number of items on your data set.  
428986944 / 11 = 38998813.09090909  
Set this number aside for a moment.

Step 3: **Square all of the numbers in your data set and then add them together**.  
(1245 x 1245) + (1255 x 1255) + (1547 x 1547) + (1654 x 1654) + (1787 x 1787) + (1878 x 1878) + (1989 x 1989) + (2011 x 2011) + (2145 x 2145) + (2545 x 2545) + (2656 x 2656) = 41106856

Step 4: **Subtract the number you calculated in Step 2 from the number you calculated in Step 3**:  
41106856 – 38998813.09090909 = 2108042.9090909064

Step 5: **Subtract 1 from the number of items in your data set**:  
11 – 1 = 10.

Step 6: **Divide the number you calculated in step 4 by the number you calculated in step 5**:  
2108042.9090909064 / 10 = 210804.29090909063  
This is the Variance.

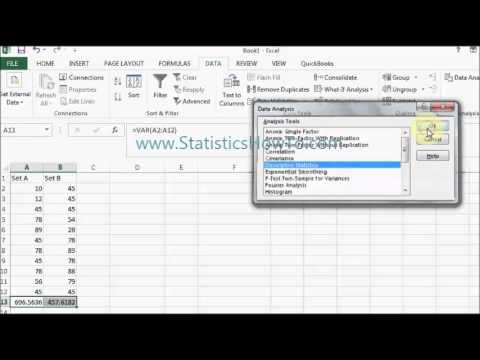
Step 7: Take the square root of Step 6 to find the standard deviation:  
√ 210804.29090909063 = 459.13.

That’s it!

### **How to Find the Sample Variance in Excel 2007-2013**

### **Sample Variance in Excel 2013**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=S_Trr0JAb7U)

[Variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)is a tool to tell you how much a data set varies. Its major use in stats is as a way to find the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/), which is a more useful measure of spread and in fact is much more widely used than the sample [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/). The equations for [finding the sample variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/sample-variance/) are quite ugly. Technology is the best way to find it without the chance of math errors creeping in.

Excel gives you two options. If you haven’t already, be sure to load the Data analysis Toolpak (How to [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)). It’s a powerful tool that you’ll use over and over again in stats. If you don’t have the Toolpak (or don’t want to install it), your second option is to use the VAR function.

### **Sample Variance Excel 2013: Data Analysis Toolpak**

Step 1: Click the “Data” tab and then click “Data Analysis.”

Step 2: Click “**Descriptive Statistics**” and then click “OK.”

Step 3: Click the Input Range box and then type the location for your data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box

Step 4: Click the radio button for Rows or Columns, depending on how your data is laid out.

Step 5: Click the “Labels in first row” box if your data has column headers.

Step 6: Click the “Descriptive Statistics” check box.

Step 7: Select a location for your output. For example, click the “New Worksheet” radio button.

Step 8: Click “OK.”

### **Sample Variance Excel 2013: VAR Function**

Step 1: Type your data into a single column.

Step 2: Click a blank cell.

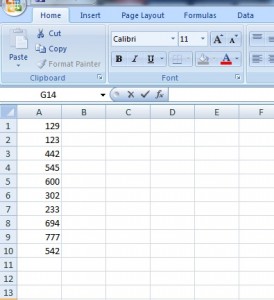
Step 3: Type “=VAR(A1:A100)” where A1:A100 is the location of your data set (i.e. in cells A1 to A100). Press the “Enter” key to get the sample [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/).

### **Sample Variance in Excel 2010**

Sample variance in Excel 2007-2010 is calculated using the “Var” function. [Watch this one-minute video](https://youtu.be/2evEY6MTxZQ) on how to calculate it, or read the steps below.

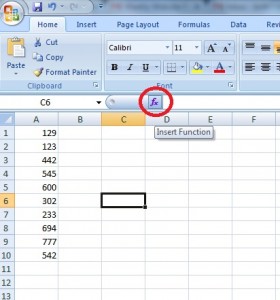
**Sample question:** Find the sample variance in Excel 2007-2010 for the following sample data: 123, 129, 233, 302, 442, 542, 545, 600, 694, 777

Step 1: **Type your data into a single column**in an Excel worksheet. For this example, I typed “123, 129, 233, 302, 442, 542, 545, 600, 694, 777” into column A. Don’t leave any blank cells between your data.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/variance-in-excel-1.jpg)

Step 2: **Click any empty cell**.

Step 3: **Click the “Insert function” button on the toolbar.**This opens the Insert Function dialog box.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/variance-in-excel-2.jpg)

Step 4: **Type “Var” into the Search for a Function text box**and then click “Go.” VAR should be highlighted in the function list.

Step 5: **Click “OK.”**

Step 6: **Type the location of the sample data into the Number1 text box**. This sample data was typed into cells A1 to A10, so I typed “A1:A10” into the text box. Make sure to separate the first and last cells by a colon (A1**:**A10).

Step 7: **Click “OK.”** Excel will return the [sample variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/sample-variance/) in the cell you chose in Step 2. For this question, the [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)of 123, 129, 233, 302, 442, 542, 545, 600, 694, 777 is 53800.46.

**Tip:** You can also access the VAR function from the “Formulas” tab in Excel. Click the “Formulas” tab and then click the “Insert Function” button on the far left of the toolbar. Continue from Step 4 above to calculate the [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/).

**Tip:** You don’t have to type your sample data into a worksheet. Technically, you could open the VAR function dialog box and then type your data into the Number1, Number2 etc. boxes. However, the advantage of typing the data directly into the worksheet is that you can perform more functions on your data (like [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)) if you need to.

## **Standard Deviation**

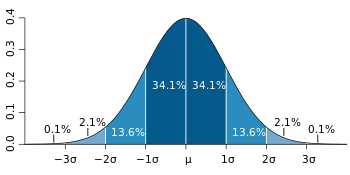
Standard deviation is a measure of dispersement in statistics. “Dispersement” tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) or [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/). For example, are all your scores close to the average? Or are lots of scores way above (or way below) the average score?

[](https://www.youtube.com/watch?v=heN3uvJ99Vo)

### **What Does it Look Like on a Graph?**

The [bell curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) (what statisticians call a “[normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)“) is commonly seen in statistics as a tool to understand standard deviation.

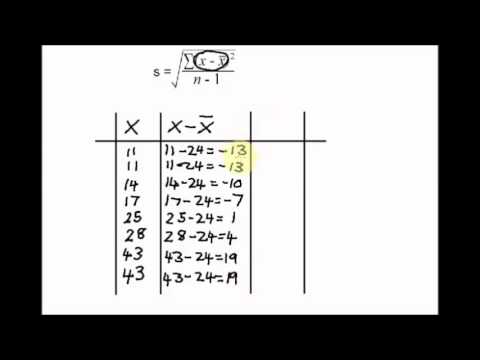
The following graph of a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) represents a great deal of data in real life. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean), or average, is represented by the Greek letter μ, in the center. Each segment (colored in dark blue to light blue) represents one standard deviation away from the mean. For example, 2σ means two standard deviations from the mean.

[](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)

### **Real Life Example**

A [normal distribution curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) can represent hundreds of situations in real life. Have you ever noticed in class that most students get Cs while a few get As or Fs? That can be modelled with a bell curve. People’s weights, heights, nutrition habits and exercise regimens can also be modelled with graphs similar to this one. That knowledge enables companies, schools and governments to make predictions about future behaviour. For behaviours that fit this type of bell curve (like performance on the [SAT](http://sat.collegeboard.org/home)), you’ll be able to predict that 34.1 + 34.1 = 68.2% of students will score very close to the average score, or one standard deviation away from the [*mean.*](https://www.statisticshowto.datasciencecentral.com/mean)

### **How to Find the Sample Standard Deviation by Hand**

[](https://www.youtube.com/watch?v=arzaMpDxYSQ)

### **Standard Deviation for a Binomial**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/09/coin.jpg)

*A coin toss can be a binomial experiment.*

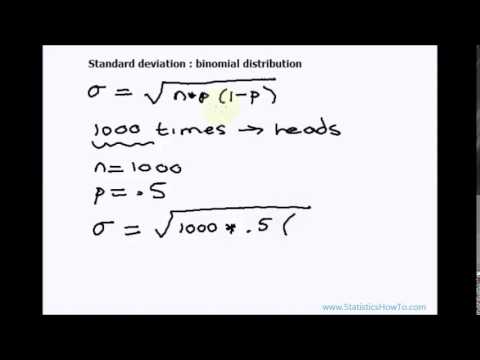
A [binomial distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/) is one of the simplest types of distributions in statistics. It’s a type of distribution where there is either success, or failure. For example, winning the lottery: or not winning the lottery. You can find the standard deviation for a binomial distribution in two ways:

1. With a formula
2. With a [probability distribution table](https://www.statisticshowto.datasciencecentral.com/what-is-a-probability-distribution-table/) (scroll down for the steps)

The formula to find the standard deviation for a binomial distribution is:

[standard deviation binomial distribution](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/standard-deviation-binomial-distribution.gif)

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=dsOdS4EJbZY)

### **Example question:**

Find the standard deviation for the following binomial distribution: flip a coin 1000 times to see how many heads you get.

Step 1: Identify n and p from the question. N is the number of trials (given as 1000) and p is the probability, which is .5 (you have a 50% chance of getting a heads in any coin flip).

At this point you can insert those numbers into the formula and solve. If formulas aren’t your forte, follow these additional steps:

Step 2: Multiply n by p:  
1000 \* .5 = 500.

Step 3: Subtract “p” from 1:  
1 – .5 = .5.

Step 4: Multiply Step 2 by Step 3: 500 \* .5 = 250.

Step 5: Take the square root of Step 4:  
√ 250 = 15.81.

That’s it!

### **Standard Deviation of Discrete Random Variables**

With discrete [random variables](https://www.statisticshowto.datasciencecentral.com/random-variable/), sometimes you’re given a [probability distribution table](https://www.statisticshowto.datasciencecentral.com/what-is-a-probability-distribution-table/) instead of “p” and “n”. As long as you have a table you can calculate the standard deviation of discrete random variables with this formula:

[standard deviation discrete random variable](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/standard-deviation-discrete-random-variable.png)

**Example question:**Find the standard deviation of the discrete random variables shown in the following table, which represents flipping three coins:

[standard deviation discrete random variable](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/std-dev-discrete-random.jpg)

Step 1: Find the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)(this is also called the [expected value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/expected-value/)) by multiplying the probabilities by x in each column and adding them all up:  
μ = (0 \* 0.125) + (1 \* 0.375) + (2 \* 0.375) + (3 \* 0.125) = 1.5

Step 2: work the inner part of the above equation, without the square root:

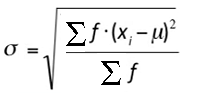
* ((0 – 1.5)2 \* 0.125 ) +
* ((1 – 1.5)2 \* 0.375 ) +
* ((2 – 1.5)2 \* 0.375 ) +
* ((3 – 1.5)2 \* 0.125 ) +
* = 0.75

Step 3: Take the square root of Step 2:  
σ = √ 0.75 = 0.8660254.

That’s it!

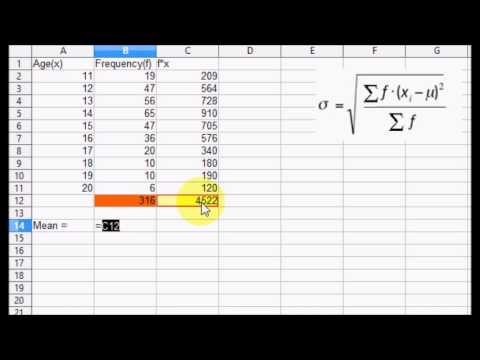
### **Standard Deviation for a**Frequency Distribution

The formula to find the standard deviation for a frequency distribution is:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/11/standard-deviation-frequency-distribution.png)  
  
Where:

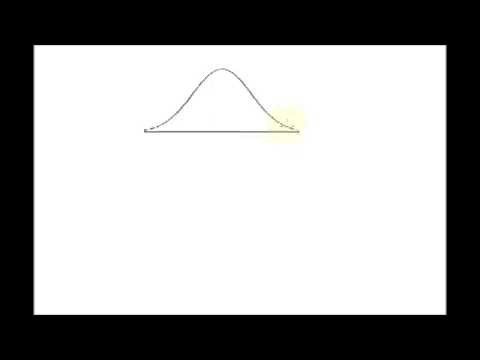
* μ is the mean for the frequency distribution,
* f is the individual frequency counts,
* x is the value associated with the frequencies.

If formulas aren’t your forte, watch this short video, which shows you how to work the formula:

[](https://www.youtube.com/watch?v=swzThKDKFoM)

## **Skewed Distribution**

Watch the video or read the article below:

[](https://www.youtube.com/watch?v=24YCewQ8s3M)

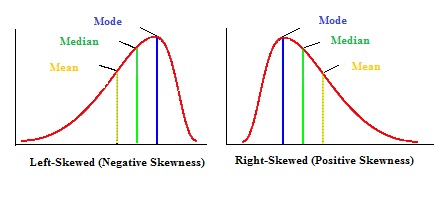
If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don’t show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half. For example, the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) is a [symmetric distribution](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/)with no skew. The tails are exactly the same.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/normal-distribution-probability.jpg)

*A normal curve.*

A [**left-skewed distribution**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewLeft) has a long left tail. Left-skewed distributions are also called negatively-skewed distributions. That’s because there is a [long tail](https://www.statisticshowto.datasciencecentral.com/long-tail-distribution/) in the negative direction on the number line. The mean is also to the left of the peak.

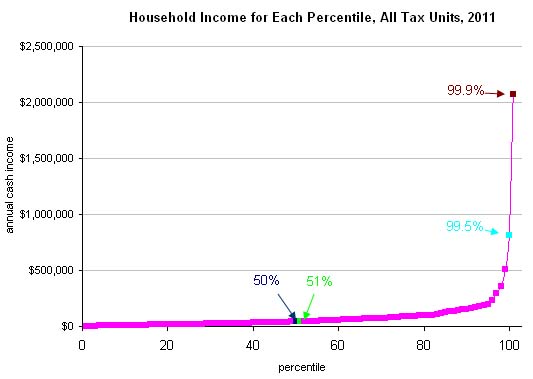
A [**right-skewed distribution**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewRight) has a long right tail. Right-skewed distributions are also called positive-skew distributions. That’s because there is a long tail in the positive direction on the number line. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)is also to the right of the peak.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg)

The [mean, mode and median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/) can be used to figure out if you have a positively or negatively skewed distribution.

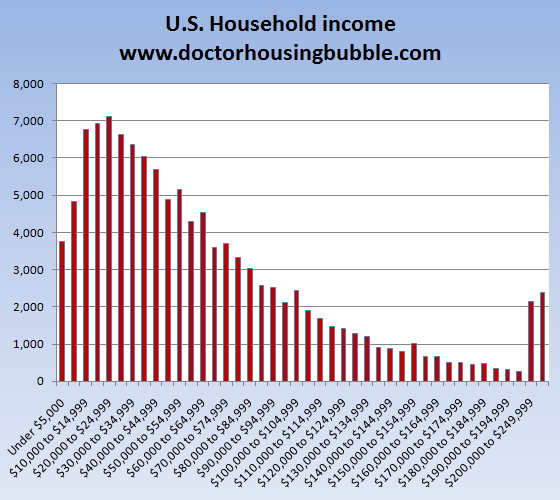
* If the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)is greater than the mode, the distribution is positively skewed.
* If the mean is less than the [mode](https://www.statisticshowto.datasciencecentral.com/mode/), the distribution is [negatively skewed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewLeft).
* If the mean is greater than the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median), the distribution is positively skewed.
* If the mean is less than the median, the distribution is negatively skewed.

The normal distribution is the most common distribution you’ll come across. Next, you’ll see a fair amount of negatively skewed distributions. For example, household income in the U.S. is negatively skewed with a very long left tail.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/positively-skewed.jpg)

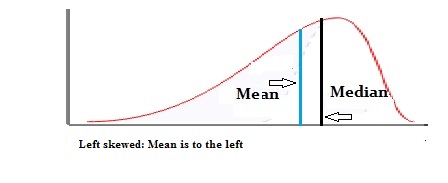
*Income in the U.S. Image: NY Times.*

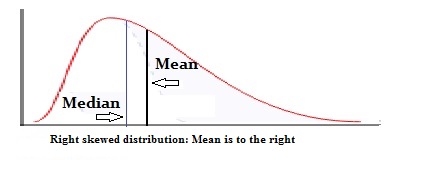
Interestingly, you can take the **same data** and make it a right-skewed distribution. This positively-skewed graph plots number of household’s income brackets:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/income-left-skew.png)

### **Mean and Median in Skewed Distributions**

In a normal distribution, the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)and the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)are the same number while the mean and median in a skewed distribution become different numbers:

A left-skewed, negative distribution will have the mean to the **left** of the median.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/left-skewed.jpg)

A right-skewed distribution will have the mean to the **right**of the median.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/right-skewed.jpg)

### **Effects on Statistics**

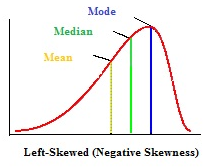
The normal distribution is the easiest distribution to work with in order to gain an understanding about statistics. Real life distributions are usually skewed. Too much skewness, and many statistical techniques don’t work. As a result, advanced mathematical techniques including [logarithms](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/logarithms/)and quantile regression techniques are used. Read more about quantile regression [here](http://www.econ.uiuc.edu/~roger/research/rq/rq.html).

### **Skewed Left (Negative Skew)**

A left skewed distribution is sometimes called a negatively skewed distribution because it’s long tail is on the negative direction on a number line.

A common misconception is that the peak of distribution is what defines “peakness.” In other words, a peak that tends to the left is left skewed distribution. This is incorrect. There are two main things that make a distribution skewed left:

1. The mean is to the left of the peak. This is the main definition behind “skewness”, which is technically a measure of the distribution of values around the mean.
2. The tail is longer on the left.
3. In most cases, the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)is to the left of the median. This isn’t a reliable test for skewness though, as some distributions (i.e. many [multimodal distributions](https://www.statisticshowto.datasciencecentral.com/multimodal-distribution/)) violate this rule. You should think of this as a “general idea” kind of rule, and not a set-in-stone one.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/skewed-left.png)

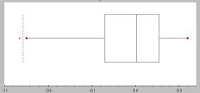
*In a left skewed distribution, the mean is to the left of the peak.*

### **Left Skewed and Numerical Values**

Skewness can be shown with a list of numbers as well as on a graph. For example, take the numbers 1,2, and 3. They are evenly spaced, with 2 as the mean (1 + 2 + 3 / 3 = 6 / 3 = 2). If you add a number to the far left (think in terms of adding a value to the number line), the distribution becomes left skewed:  
-10, 1, 2, 3.  
Similarly, if you add a value to the far right, the set of numbers becomes right skewed:  
1, 2, 3, 10.

### **Left Skewed Boxplot**

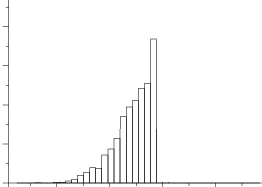
If the bulk of observations are on the high end of the scale, a [boxplot](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/) is left skewed. Consequently, the left whisker is longer than the right whisker.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/left-skewed-boxplot.png)

*A left skewed boxplot, showing a long left whisker. Image:*[*SHU.EDU*](http://pirate.shu.edu/~wachsmut/Teaching/MATH1101/Descriptives/box.html)

### **Left Skewed Histogram**

Left skewed histograms are [Histograms](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/histogram-make-chart/)with long tails on the left.

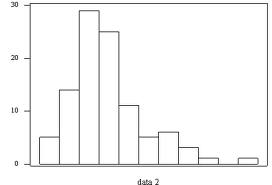
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/left-skewed-histogram.png)

### **Skewed Right / Positive Skew**

A right skewed distribution is sometimes called a positive skew distribution. That’s because the tail is longer on the positive direction of the [number line](http://www.math.utah.edu/online/1010/line/).

### **Right Skewed Histogram**

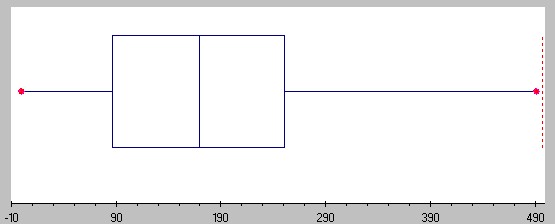
A [histogram](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/histogram-make-chart/)is right skewed if the peak of the histogram veers to the left. Therefore, the histogram’s tail has a positive skew to the right.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/right-skewed-histogram.jpg)

*A skewed to the right histogram. Image: SUNY Oswego*

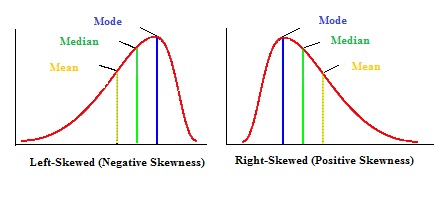
### **Right Skewed Box Plot**

If a [box plot](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/) is skewed to the right, the box shifts to the left and the right whisker gets longer. As a result, the mean is greater than the median

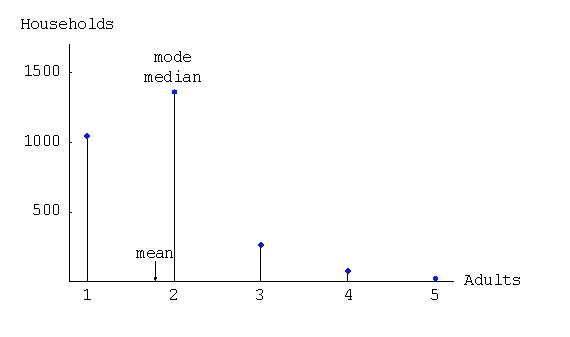
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/right-skewed-box-plot.jpg)

*Image: Seton Hall University*

### **Right Skewed Mean and Median**

The rule of thumb is that in a right skewed distribution, the mean is usually to the right of the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg)  
  
  
However, like most rules of thumb, there are exceptions. Most right skewed distributions you come across in elementary statistics will have the mean to the right of the median. The [Journal of Statistics Education](https://www.amstat.org/publications/jse/v13n2/vonhippel.html) points out an exception to the rule:

In a [data analysis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/) course, a third moment formula calculates the skew (see: [What is a Moment?](https://www.statisticshowto.datasciencecentral.com/what-is-a-moment/)). Consequently, some distributions can break the rule of thumb. The following distribution was made from a 2002 [General Social Survey](http://www3.norc.org/GSS+Website/). Respondents stated how many people older than 18 lived in their household. This is a right-skewed graph, but the mean is clearly to the left of the median.

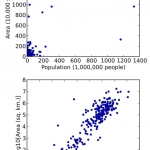
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/skew-mean-median.jpg)

*Image: Journal of Statistics Education*

There are other exceptions which most involve theoretical mathematics and [calculus](https://calculushowto.com/). The important point to note is that although the mean is generally to the right of the median in a right skewed distribution, it isn’t an absolute fact.

### **Log Transformation of a Skewed Distribution.**

Log transformation means taking a data set and taking the natural logarithm of variables. Sometimes your data may not quite fit the model you are looking for, and a log transformation can help to fit a very skewed distribution into a more normal model (a “[bell curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)“). As a result, you can more easily see patterns in your data. Log transformation does not “normalize” your data; it’s purpose is to reduce skew.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/03/log-transformation.png)

*“Population vs area” by Skbkekas.*[*Wikimedia Commons*](https://commons.wikimedia.org/wiki/File:Population_vs_area.svg#mediaviewer/File:Population_vs_area.svg)

In the image above, it’s practically impossible to see any pattern in the above image. However, in the second image, the data has had a log transformation. Consequently, the pattern becomes apparent.

### **Log Transformations and Statistical Tests**

If you are running a [parametric statistical test](https://www.statisticshowto.datasciencecentral.com/parametric-statistics) on your data (for example, an [ANOVA](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/anova/)), using data that’s highly skewed to the right or left can lead to misleading test results. Therefore, if you want to perform a test on this kind of data, run a log transformation and then run the test on the transformed numbers.

### **When Should I Use Log Transformation?**

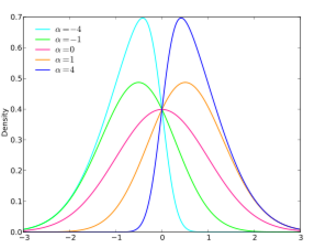
Many possible [transformations](https://www.statisticshowto.datasciencecentral.com/transformation-statistics/)exist. However, you should only use a log transformation if:

* Your data is highly skewed to the right (i.e. in the positive direction).
* The residual’s [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) is proportional to your fitted values
* The data’s relationship is close to exponential.
* You think the [residuals](https://www.statisticshowto.datasciencecentral.com/residual/)reflect multiplicative errors that have accumulated during each step of the computation.

### **Log transformation in Software**

* In SPSS: [IBM’s instructions can be found here.](https://www-01.ibm.com/support/docview.wss?uid=swg21481524)

### **Skew Normal Distribution**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/08/skew-normal.png)

*The probability density function for the skew normal, showing various alphas. Image: skbkekas|Wikimedia Commons.*

The skew normal distribution is a normal distribution with an extra [shape parameter](https://www.statisticshowto.datasciencecentral.com/shape-parameter/), α. The shape parameter [skews](https://www.statisticshowto.datasciencecentral.com/skewness/)the normal distribution to the left or right. As it is only the skew of the normal distribution that’s being changed, the skew normal family has many of the same properties of the normal distribution:

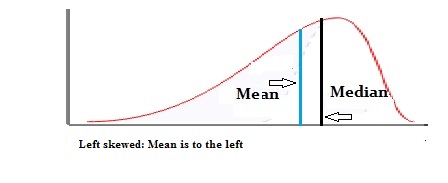
* It’s defined over the real number line.
* The square of a [random variable](https://www.statisticshowto.datasciencecentral.com/random-variable/) is a [chi-square](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/)variable (from a chi-square distribution) with one [degree of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/).
* The distribution is [unimodal](https://www.statisticshowto.datasciencecentral.com/unimodal-distribution-2/)(one peak).
* The [location parameter](https://www.statisticshowto.datasciencecentral.com/location-parameter/), μ(i.e. the mean), defines where the peak is and the [scale parameter](https://www.statisticshowto.datasciencecentral.com/scale-parameter/), σ(i.e. the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)) determines the distribution’s spread.

The skew normal has a number of interesting properties related to alpha:

* If the skew normal has a skew of zero, then it becomes the normal distribution.
* If the sign of alpha changes, the distribution will flip over the y-axis.
* As alpha increases (in [absolute value](https://www.statisticshowto.datasciencecentral.com/real-numbers/#absolute)), the skew also increases.
* As alpha tends towards infinity, the series converges to the [folded normal density function.](https://www.statisticshowto.datasciencecentral.com/folded-normal-distribution/)

Therefore, the normal distribution can be seen as a special case of the skew normal distribution.

## **Skewness**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/left-skewed.jpg)

Skewness is a measure of symmetry in a distribution. Actually, it’s more correct to describe it as a measure of **lack** of symmetry. A standard normal distribution is perfectly symmetrical and has zero skew. Other examples of zero-skewed distributions are the [T Distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/), the [uniform distribution](https://www.statisticshowto.datasciencecentral.com/uniform-distribution/) and the [Laplace distribution](https://www.statisticshowto.datasciencecentral.com/laplace-distribution-double-exponential/). However, other distributions don’t have zero skew. Therefore, we need a way to calculate how much the distribution is skewed.

### **Equations**

The following table shows the equations for calculating the skew for various distributions you’re likely to come across in [elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/). Note that the normal, Student’s T, uniform and Laplace distributions are not shown on the table as they are not skewed.

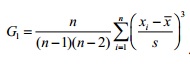
|  |  |
| --- | --- |
| DISTRIBUTION. | EQUATION. |
| [Bernoulli distribution](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/). | [bernoulli](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/bernoulli.gif) |
| [Beta distribution](http://www.math.uah.edu/stat/special/Beta.html). | [beta](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/beta.gif) |
| [Binomial distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/). | [binomial](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/binomial.gif) |
| [Chi square distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/#chisquaredist). | [chi square](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/chi-square.gif) |
| [F distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/f-statistic-value-test/#Fdist). | [f dist](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/f-dist.gif) |
| [Negative binomial](https://www.statisticshowto.datasciencecentral.com/negative-binomial-experiment/). | [negative binomial](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/negative-binomial.gif) |
| [Poisson Distribution](https://www.statisticshowto.datasciencecentral.com/poisson-distribution/). | [poisson](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/poisson.gif) |

### **Calculation**

The formula given in most textbooks is Skew = 3 \* (Mean – Median) / Standard Deviation. This is known as an [alternative Pearson Mode Skewness](https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/).

You could calculate skew by hand. But why bother? Most stats software can calculate it for you.

[Skewness in Excel 2013](https://www.statisticshowto.datasciencecentral.com/skewness/): “=SKEW(xx:yy)” where xx:yy is the cell location of your data (for example, C1:C25).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/adjusted-pearson.jpg)

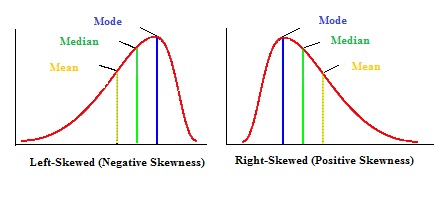
You probably won’t actually see or use this formula; It generally doesn’t even appear in elementary stats textbooks. But it is the one used behind the scenes by most software. The results of the calculation tell you:

* The direction of the skew (positive or negative).
* How the sample compares with a normal (symmetric) distribution. The further the skew result is from zero, the greater the skew.

### **Different Ways to Measure a Skew:**

1. [Pearson Mode](https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/).
2. [Bowley](https://www.statisticshowto.datasciencecentral.com/bowley-skewness/).
3. [Kelly’s Measure.](https://www.statisticshowto.datasciencecentral.com/kellys-measure-of-skewness/)
4. [Momental.](https://www.statisticshowto.datasciencecentral.com/skewness/#momental)

### **What is Momental Skewness?**

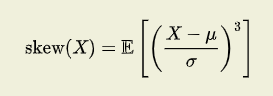
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg)  
Momental skewness is one of four ways you can calculate the skew of a distribution. It’s called “Momental” because the first moment in statistics is the [mean](https://www.statisticshowto.datasciencecentral.com/mean/).

The formula for calculating momental skewness (γ) is:  
α(m) = 1/2 γ1 = μ3/2 σ3

α is the shape parameter

Where μ is the mean and σ is the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) and γ is the [Fisher Skewness](http://archive.lib.msu.edu/crcmath/math/math/f/f185.htm).

The skewness of X is the third moment of the [standard score](http://www.randomservices.org/random/expect/Variance.html#std) of X:

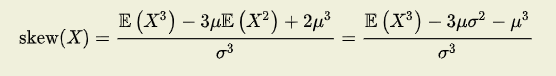


The distribution of X is said to be positively skewed, negatively skewed or unskewed depending on whether skew(X) is positive, negative, or 0.

Suppose that the distribution of X is symmetric about a. Then

1. E(X)=a
2. skew(X)=0.

Skew(x)  can be expressed in terms of the first three moments of X.



### **Why use Momental Skewness?**

Which technique you use depends on what you know about your data. For example, if you know the [mean](https://www.statisticshowto.datasciencecentral.com/mean/), [mode](https://www.statisticshowto.datasciencecentral.com/mode/)(or [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)) and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) you can use Pearson’s. Momental skewness could be an option if you only know the mean and standard deviation for your set of data.

### **What do the Results Mean?**

A [symmetrical distribution](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/) has a skew of zero. A positive result means that your data is positively skewed. A negative results means that your data is negatively skewed. Momental skewness will give you a result that is half of typical skewness.

### **Skewness in Excel 2013**

When you calculate skewness in Excel, Excel will return a positive or a negative number. A [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), which has perfectly symmetrical tails, has a skewness of zero.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/skew.jpg)

*Negative skewness (left) and positive skewness (right)*

You have two different options for calculating skewness in Excel 2013: the SKEW function or Data Analysis (How to [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)). If you’re taking a statistics class, it would be worth your while to install the Toolpak as it gives you access to many stats functions that aren’t in the base installation of Excel.

### **SKEW Function**

Step 1: Type your data into columns.

Step 2: Click an empty cell on the worksheet.

Step 3: Type “=SKEW(xx:yy)” where xx:yy is the cell location of your data (for example, C1:C25).

### **Data Analysis**

Step 1: Click the “Data” tab and then click “Data Analysis.”

Step 2: Click “Descriptive Statistics” and then click “OK.”

Step 3: Click the Input Range box and then type the location for your data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box

Step 4: Click the radio button for Rows or Columns, depending on how your data is laid out.

Step 5: Click the “Labels in first row” box if your data has column headers.

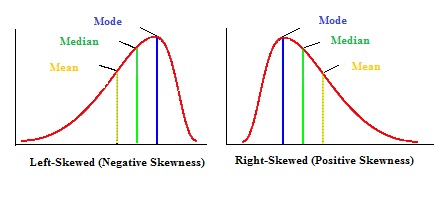
Step 6: Click the “Descriptive Statistics” check box.

Step 7: Select a location for your output. For example, click the “New Worksheet” radio button.

Step 8: Click “OK.”

## **Pearson Mode Skewness**

Pearson mode skewness, also called Pearson’s first coefficient of skewness, is a way to figure out the [skewness](https://www.statisticshowto.datasciencecentral.com/skewness/)of a distribution. This is one of the most popular ways to find skewness is the [Pearson Mode Skewness](https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/) formula

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg)

The [mean, mode and median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/) can be used to figure out if you have a positively or negatively skewed distribution.

* If the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)is greater than the mode, the distribution is positively skewed.
* If the mean is less than the [mode](https://www.statisticshowto.datasciencecentral.com/mode/), the distribution is [negatively skewed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/#SkewLeft).
* If the mean is greater than the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median), the distribution is positively skewed.
* If the mean is less than the median, the distribution is negatively skewed.

### **Pearson Mode skewness**

Pearson mode skewness uses the above facts to help you find out if you have positive or negative skewness. If you have a distribution and you know the mean, mode, and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ), then the Pearson mode skewness formula is:

**(mean-mode)/σ**

Also called as Pearson’s first coefficient of skewness

**Sample problem:**You have data with a mean of 19, a mode of 20 and a standard deviation of 25. What does Pearson Mode Skewness tell you about the distribution?  
(mean-mode)/σ = (19-20)/25 = -0.04.  
There is a very slight negative skewness (-0.04).

**Note**: For most intents and purposes, this would count as a [symmetric distribution](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/) as the skewness is so small.

### **Pearson Mode Skewness: Alternative Formula.**

If you don’t know the mode, you won’t be able to use Pearson mode skewness. However, the direction of skewness can be also figured out by finding where the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)and the [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)are. According to [Business Statistics](http://books.google.com/books?id=tWmoP49v1cIC&pg=PA159&lpg=PA159&dq=pearson+mean-mode&source=bl&ots=29GD9TU7gF&sig=riP4tclXBnX8GRKygrxbgyrk_zM&hl=en&sa=X&ei=qoTuUrDPH6SsyAGF24D4BQ&ved=0CEkQ6AEwAw#v=onepage&q=pearson%20mean-mode&f=false), this leads to a second, equivalent formula:

**3(Mean – Median) / σ**

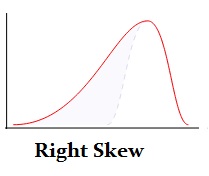
This formula is also called Pearson’s second coefficient of skewness.

### **Pearson Mode Skewness: What the Results mean**

The difference between the mean and mode, or mean and median, will tell you **how far the distribution departs from symmetry**. A symmetric distribution (for example, the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)) has a skewness of zero.

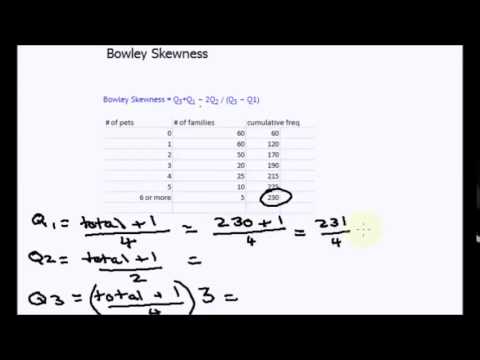
Both equations give you results in [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/), which are dimensionless units of measurement from the mean.

## **Bowley Skewness**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/right-skew.jpg)

Bowley [skewness](https://www.statisticshowto.datasciencecentral.com/skewness/)is a way to figure out if you have a positively-skewed or negatively skewed distribution. One of the most popular ways to find skewness is the [Pearson Mode Skewness](https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/) formula. However, in order to use it you must know the [mean, mode (or median)](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/) and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) for your data. Sometimes you might not have that information; Instead you might have information about your [quartiles](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/). If that’s the case, you can use Bowley Skewness as an alternative to find out more about the asymmetry of your distribution. It’s very useful if you have extreme data values ([outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/))or if you have an [open-ended distribution](http://stats.oecd.org/glossary/detail.asp?ID=3770).

### **Bowley Skewness Formula**

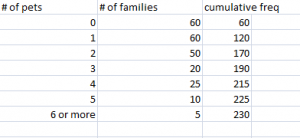
[](https://www.youtube.com/watch?v=34U6Wg1c_U8)

Bowley Skewness = Q3+Q1 – 2Q2 / (Q3 – Q1)

* Skewness = 0 means that the curve is symmetrical.
* Skewness > 0 means the curve is positively skewed.
* Skewness < 0 means the curve is negatively skewed.

### **Bowley Skewness Worked Example**

Q. Find the Bowley Skewness for the following set of data:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/bowley-skewness-example1.png)

Step 1: Find the [Quartiles](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/)for the data set. You’ll want to look for the “nth” observation using the following formulas:

Q1 = (total cum freq + 1) / 4th observation = (230 + 1 / 4 ) = 57.75  
Q2 = (total cum freq + 1 )/ 2th observation = (230 + 1 / 2 ) = 115.5  
Q3 = (total cum freq + 1) 3/ 4th observation = 3(230 + 1 / 4) = 173.25

Step 2: Look in your table to find the nth observations you calculated in Step 1:  
Q1 = 57.75th observation = 0  
Q2 = 115.5th observation = 1  
Q3 = 173.25th observation = 3

Step 3: Plug the above values into the formula:  
Skq = Q3 + Q1 – 2Q2 / Q3 – Q1  
Skq = 3 + 0 – 2 / 3 – 0 = 1/3

Skq = + 1/3, so the distribution is positively skewed.

### **Why Bowley Skewness works**

In a [symmetric distribution](https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/), like the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), the first (Q1) and third (Q3) quartiles are at equal distances from the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)(Q2). In other words, (Q3-Q2) and (Q2-Q1) will be equal. If you have a [skewed distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/) then there will be a difference between those two values.

### **Limitations of Bowley Skewness.**

Bowley Skewness is an **absolute measure of skewness**. In other words, it’s going to give you a result in the units that your distribution is in. That’s compared to the Pearson Mode Skewness, which gives you results in a dimensionless unit — the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). This means that you cannot compare the skewness of different distributions with different units using Bowley Skewness.

### **Alternative Bowley Skewness formula**

According to [Business Statistics](http://books.google.com/books?id=tWmoP49v1cIC&pg=PA159&lpg=PA159&dq=pearson+mean-mode&source=bl&ots=29GD9TU7gF&sig=riP4tclXBnX8GRKygrxbgyrk_zM&hl=en&sa=X&ei=qoTuUrDPH6SsyAGF24D4BQ&ved=0CEkQ6AEwAw#v=onepage&q=pearson%20mean-mode&f=false), Bowley recognized that the Bowley Skewness formula could not be used to compare different distributions with different units. For example, you can’t compare a distribution measured in heights in cm with one of weights in pounds. He offered an alternative formula. You should use this formula if you want to compare different distributions with different units:

Relative Skewness = (Q3 + Q1) – (2 \* [Median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)/ Q3 – Q1).

## **Kelly’s Measure**

Kelly’s Measure of Skewness is one of several ways to measure [skewness](https://www.statisticshowto.datasciencecentral.com/skewness/)in a data distribution. [Bowley’s skewness](https://www.statisticshowto.datasciencecentral.com/bowley-skewness/) is based on the middle 50 percent of the observations in a data set. It leaves 25 percent of the observations in each tail of the distribution. Kelly suggested that leaving out fifty percent of data to calculate [skewness](https://www.statisticshowto.datasciencecentral.com/skewness/)was too extreme. He created a measure to find skewness with more data. Kelly’s measure is based on P90 (the 90th [percentile](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/)) and P10 (the 10th percentile). Only twenty percent of observations (ten percent in each tail) are excluded from the measure.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/income-left-skew.png)

*A right-skewed distribution.*

### **Kelly’s Measure Formula**

Kelley’s measure of skewness is given in terms of percentiles and [deciles](http://www.merriam-webster.com/dictionary/decile)(D). Kelley’s absolute measure of skewness (Sk)is:

Sk=P90 + P10 – 2\*P50 = D9 + D1-2\*D5.

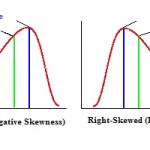
This formula is **not practically used**. In fact, Kelly’s measure of skewness is **rarely used** at all, even in it’s more common form, which is measured as coefficient of skewness:

SP = P90 – 2\*P50 + P10 / P90 – P10

### **What does Kelly’s Measure of Skewness Tell us?**

Kelly’s Measure of Skewness gives you the same information about skewness as the other three types of skewness measures ([Bowley skewness](https://www.statisticshowto.datasciencecentral.com/bowley-skewness/), [Pearson Mode Skewness](https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/) and [Momental Skewness.](https://www.statisticshowto.datasciencecentral.com/kellys-measure-of-skewness/momental)).

A measure of skewness = 0 means that the distribution is symmetrical.  
A measure of skewness > 0 means a positive skewness.  
A measure of skewness < means a negative skewness.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/pearson-mode-skewness.jpg)

## **Kurtosis: Definition, Leptokurtic, Platykurtic**

* A **positive value** tells you that you have [heavy-tails](https://www.statisticshowto.datasciencecentral.com/heavy-tailed-distribution/) (i.e. a lot of data in your tails).
* A**negative value**means that you have [light-tails](https://www.statisticshowto.datasciencecentral.com/heavy-tailed-distribution/) (i.e. little data in your tails).

This heaviness or lightness in the tails usually means that your data looks flatter (or less flat) compared to the normal distribution. The standard normal distribution has a kurtosis of 3, so if your values are close to that then your graph’s tails are nearly normal. These distributions are called **mesokurtic**.

Kurtosis is the fourth [moment](https://www.statisticshowto.datasciencecentral.com/what-is-a-moment/)in statistics.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/kurtosis-excel-2013.jpg)

*The distribution on the left has a very negative kurtosis (no tails); the one on the right has positive kurtosis (heavier tails compared to the normal distribution).*

### **Mesokurtic**

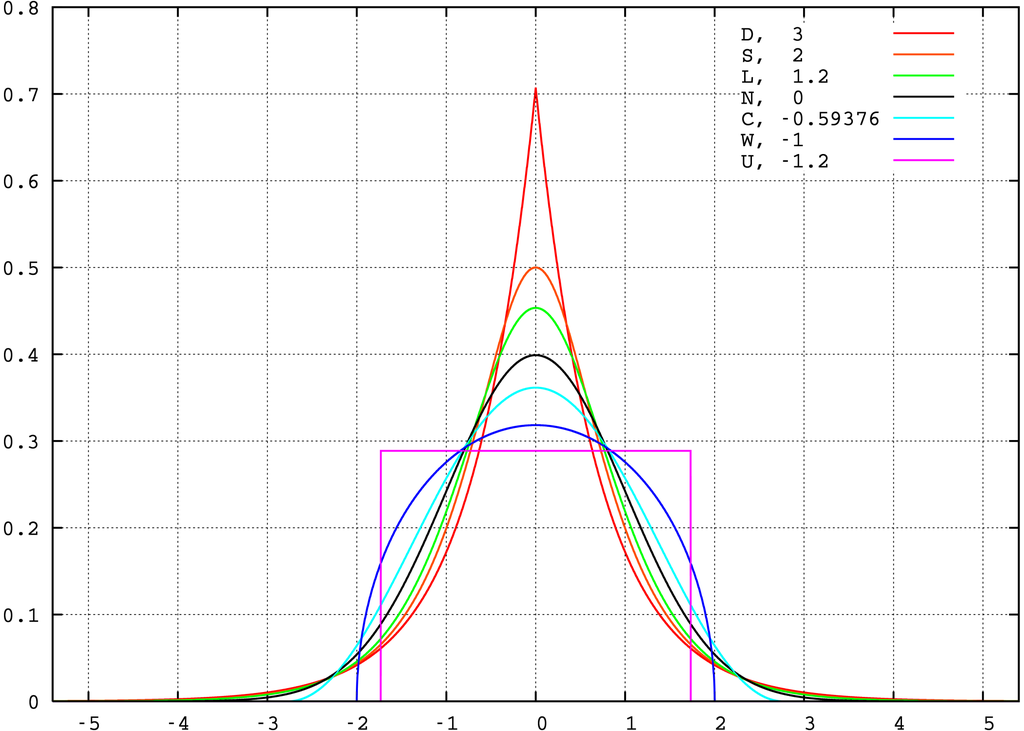
Mesokurtic distributions are technically defined as having a kurtosis of zero, although the distribution doesn’t have to be exactly zero in order for it to be classified as mesokurtic. The most common mesokurtic distributions are:

* The normal distribution.
* Any distribution with a Gaussian (normal) shape and zero probability at other places on the real line.
* The [binomial distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/) is mesokurtic for some values (i.e. for p = 1/2±√(1/12).

### **What is Excess kurtosis?**

Excess kurtosis is usually defined as kurt – 3 (see [Important note about equations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/#note)). It is a measure of how the distribution’s tails compare to the normal (Aldrich, E, 2014).

* Excess kurt for the normal distribution is 0 (i.e. 3 -3 = 0).
* Negative excess equals lighter tails than a normal distribution.
* Positive excess equals heavier tails than the normal.

The following graph shows a variety of distributions. Note how the tails are fatter or thinner than the normal (black):  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2017/12/1024px-Standard_symmetric_pdfs.png)  
  
  
Key:  
Red, kurt 3, Laplace (D)ouble exponential distribution;  
Orange, kurt 2, hyperbolic (S)ecant distribution;  
Green, kurt 1.2, [(L)ogistic distribution;](https://www.statisticshowto.datasciencecentral.com/logistic-distribution/)  
Black, kurt 0, (N)ormal distribution;  
Cyan, kurt −0.593762…, raised (C)osine distribution;  
Blue, kurt −1, (W)igner semicircle distribution;  
Magenta, kurt −1.2,[(U)niform distribution.](https://www.statisticshowto.datasciencecentral.com/uniform-distribution/)

### **Calculating Kurtosis**

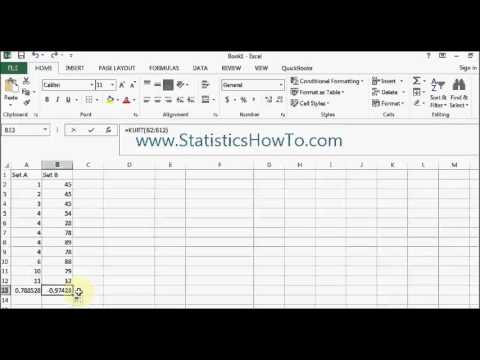
**Important Note about formulas:** **There’s no real consensus for exactly what the correct equation is for calculating kurt**. Which definition/equation you use is a matter of convention in your field, the particular software you’re working with, and sometimes the preference of the author. Therefore, it’s a good idea to check which formula you’re working with. [This Cross Validated thread](https://stats.stackexchange.com/questions/61740/differences-in-kurtosis-definition-and-their-interpretation) has an excellent rundown of the different equations and which software uses which equation.

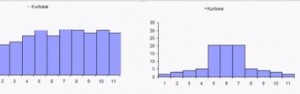
For Minitab and SPSS, you can find the option in the “Descriptive Statistics” tab.

### **Kurtosis in Excel 2013**

**Note**: The “KURT” reported by Excel is actually the **excess**kurtosis. See the [note on formulas](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/#note) above.

Watch the video or read the steps below:



[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/kurtosis-excel-2013.jpg)

*Negative kurt(left) and positive kurt (right)*

There are two options in Excel for finding kurtosis: the KURT Function and the Data Analysis Toolpak (How to [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)).

### **Kurtosis Excel 2013: KURT function**

Step 1: Type your data into columns in an Excel worksheet.

Step 2: Click a blank cell.

Step 3: Type “=KURT(A1:A99)” where A1:99 is the cell locations for your data.

### **Kurtosis Excel 2013: Data Analysis**

Step 1: Click the “Data” tab and then click “Data Analysis.”

Step 2: Click “Descriptive Statistics” and then click “OK.”

Step 3: Click the Input Range box and then type the location for your data. For example, if you typed your data into cells A1 to A10, type “A1:A10” into that box

Step 4: Click the radio button for Rows or Columns, depending on how your data is laid out.

Step 5: Click the “Labels in first row” box if your data has column headers.

Step 6: Click the “Descriptive Statistics” check box.

Step 7: Select a location for your output. For example, click the “New Worksheet” radio button.

Step 8: Click “OK.”

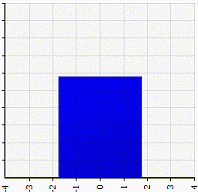
### **Platykurtic**

Platykurtic distributions have negative kurtosis. The tails are very thin compared to the normal distribution, or — as in the case of the uniform distribution— non-existent.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/kurtosis-excel-2013.jpg)

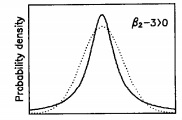
*Platykurtic (left) and leptokurtic (right).*

An example of a very platykurtic distribution is the [uniform distribution](https://www.statisticshowto.datasciencecentral.com/uniform-distribution/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/shape_uniform.gif)

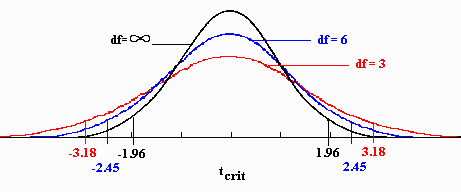
*A uniform distribution.*

### **Leptokurtic**

A leptokurtic distribution has excess positive [kurtosis](https://www.statisticshowto.datasciencecentral.com/kurtosis/), where the [kurtosis](https://www.statisticshowto.datasciencecentral.com/kurtosis/)is greater than 3. The tails are fatter than the normal distribution. The following illustration shows a leptokurtic distribution along with a normal distribution (dotted line).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/04/lepto.jpg)

### **The Leptokurtic T-Test**

The [T distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/) is an example of a leptokurtic distribution. It has fatter tails than the normal (you can also look at the first image above to see the fatter tails). Therefore, the [critical values](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/) in a [Student’s t-test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/) will be larger than the critical values from a [z-test](https://www.statisticshowto.datasciencecentral.com/z-test/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/11/t-distribution.gif)

*The t-distribution.*

### **Financial Markets**

[Kurtosis](https://www.statisticshowto.datasciencecentral.com/kurtosis/)isn’t just a theory confined to mathematical textbooks; it has real life applications, especially in the world of economics. Fund managers usually focus on risks and returns, [kurtosis](https://www.statisticshowto.datasciencecentral.com/kurtosis/)(in particular if an investment is lepto- or platy-kurtic). According to stock trader and analyst Michael Harris, a leptokurtic return means that risks are coming from [outlier](https://www.statisticshowto.datasciencecentral.com/find-outliers/)events. This would be a stock for investors willing to take extreme risks. For example, real estate (with a kurt of 8.75) and High Yield US bonds (8.63) are high risk investments while Investment grade US bonds (1.06) and Small cap US stocks (1.08) would be considered safer investments.

## **Interquartile Range (IQR)**

The **interquartile range** is a measure of where the “[middle fifty](https://www.statisticshowto.datasciencecentral.com/middle-fifty/)” is in a data set. Where a [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/)is a measure of where the beginning and end are in a set, **an interquartile range is a measure of where the bulk of the values lie.** That’s why it’s preferred over many other [measures of spread](https://www.statisticshowto.datasciencecentral.com/measures-of-spread/) (i.e. the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)or [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)) when reporting things like school performance or SAT scores.

The interquartile range formula is the first [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/)subtracted from the third [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/):

**IQR = Q3 – Q1.**

### **Solve the formula by hand**

[Watch the video](https://youtu.be/R6VDj7pEG30) or read the steps below:

**Steps:**

* Step 1: **Put the numbers in order.**  
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
* Step 2: **Find the**[**median**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median).  
  1, 2, 5, 6, 7**, 9**, 12, 15, 18, 19, 27.
* Step 3: **Place parentheses around the numbers above and below the median.**  
  Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.  
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
* Step 4: **Find Q1 and Q3**  
  Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.  
  (1, 2, **5**, 6, 7)**,  9**, ( 12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
* Step 5: **Subtract Q1 from Q3 to find the interquartile range**.  
  18 – 5 = 13

### **What if I Have an Even Set of Numbers?**

**Sample question**: Find the IQR for the following data set: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21.

* Step 1: **Put the numbers in order**.  
  3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
* Step 2: **Make a mark in the center of the data**:  
  3, 5, 7, 8, 9, **|** 11, 15, 16, 20, 21
* Step 3: **Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.**  
  (3, 5, 7, 8, 9), **|** (11, 15, 16, 20, 21)
* Step 4: **Find Q1 and Q3**  
  Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.  
  (3, 5, **7**, 8, 9), **|** (11, 15, **16,** 20, 21). Q1 = 7 and Q3 = 16.
* Step 5: **Subtract Q1 from Q3**.  
  16 – 7 = 9.

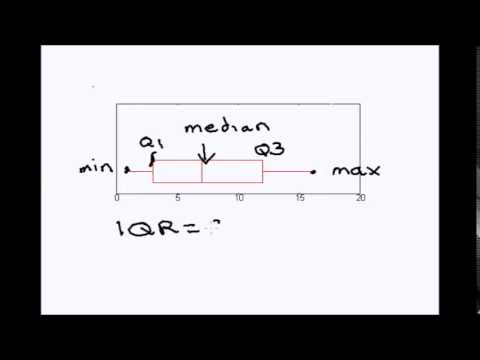
### **Find an interquartile range for an odd set of numbers: Alternate Method**

As you may already know, nothing is “set in stone” in statistics: when some statisticians find an interquartile range for a set of odd numbers, they include the median in both  quartiles. For example, in the following set of numbers: 1,2,5,6,7,9,12,15,18,19,27 some statisticians would break it into two halves, including the median (9) in both halves:  
(1,2,5,6,7,9),(9,12,15,18,19,27)

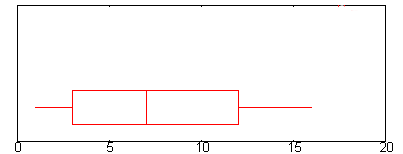
This leads to two halves with an even set of numbers, so you can follow the steps above to find the IQR.

## **Box Plot interquartile range: How to find it**

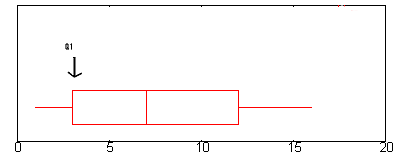
Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=oI0qDG5ZqZg)

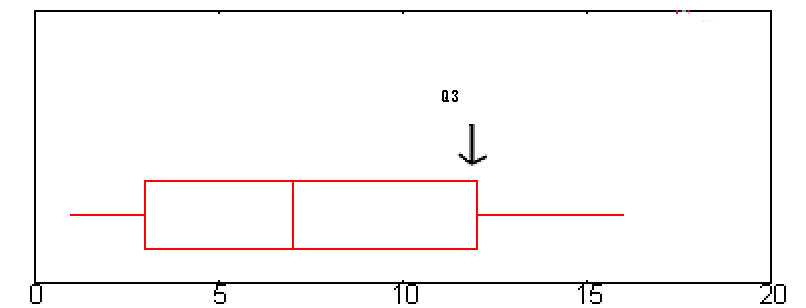
### **Box Plot interquartile range: How to find it**



**Sample question:** Find the interquartile range for the above box plot.

* **Step 1:** Find Q1.Q1 is represented by the left hand edge of the “box” (at the point where the whisker stops).  
  

In the above graph, Q1 is approximately at **2.6**. (A complete explanation of Q1 is here: [The five number summary](https://www.statisticshowto.datasciencecentral.com/how-to-find-a-five-number-summary-in-statistics/).)

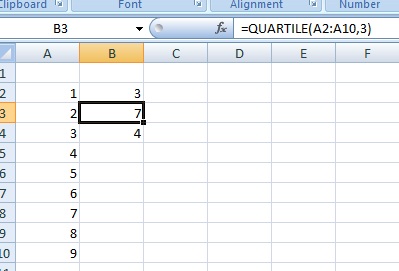
* **Step 2:** Find Q3.  
  Q3 is represented on a boxplot by the right hand edge of the “box”.  
    
  Q3 is approximately **12** in this graph.
* **Step 3:** Subtract the number you found in step 1 from the number you found in step 3.  
  This will give you the interquartile range. 12 – 2.6 = **9.4**.

That’s it!

### **Interquartile Range in Excel 2007**

### **How to Find an Interquartile Range Excel 2007**

[Watch the video](https://youtu.be/wda-jwGHNSg) or read the steps below to find an interquartile range in Excel 2007:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/04/iqr.jpg)  
  
  
**Steps:**Step 1: Enter your data into a single Excel column on a worksheet. For example, type your data in cells A2 to A10. Don’t leave any gaps in your data.

Step 2: Click a blank cell (for example, click cell B2) and then type **=QUARTILE(A2:A10,1)**. You’ll need to replace A2:A10 with the actual values from your data set. For example, if you typed your data into B2 to B50, the equation is **=QUARTILE(B2:B50,1)**. The “1” in this Excel formula(A2:A10,**1**) represents the first quartile (i.e the point lying at 25% of the data set).

Step 3: Click a second blank cell (for example, click cell B3) and then type **=QUARTILE(A2:A10,3)**. Replace A2:A10 with the actual values from your data set. The “3” in this Excel formula (A2:A10,**3**) represents the third quartile (i.e. the point lying at 75% of the data set).

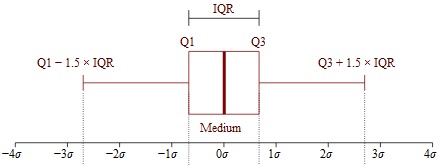
Step 4: Click a third blank cell (for example, click cell B4) and then type **=B3-B2**. If your quartile functions from Step 2 and 3 are in different locations, change the cell references.

Step 5: Press the “Enter” key. Excel will return the IQR in the cell you clicked in Step 4

That’s it!

### **What is an Interquartile Range?**

Imagine all the data in a set as points on a [number line](http://www.math.utah.edu/online/1010/line/). For example, if you have 3, 7 and 28 in your set of data, imagine them as points on a number line that is centered on 0 but stretches both infinitely below zero and infinitely above zero. Once plotted on that number line, the smallest data point and the biggest data point in the set of data create the boundaries of an interval of space on the number line that contains all data points in the set. The **interquartile range** (IQR) is the length of the middle 50% of that interval of space.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/03/iqr.jpg)

*The interquartile range is the middle 50% of a data set. Box and whiskers image by Jhguch at en.wikipedia*

If you want to know that the IQR is in formal terms, the IQR is calculated as: The difference between the third or **upper quartile** and the first or **lower quartile**. Quartile is a term used to describe how to divide the set of data into four equal portions (think quarter).

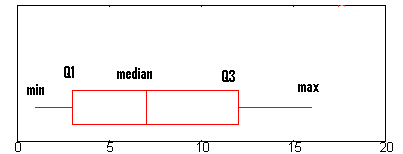
### **IQR Example**

If you have a set containing the data points 1, 3, 5, 7, 8, 10, 11 and 13, the first quartile is 4, the second quartile is 7.5 and the third quartile is 10.5. Draw these points on a number line and you’ll see that those three numbers divide the number line in quarters from 1 to 13. As such, the **IQR**of that data set is 6.5, calculated as 10.5 minus 4. The first and third quartiles are also sometimes called the 25th and 75th [percentiles](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/)because those are the equivalent figures when the data set is divided into percents rather than quarters.

### **What is The Interquartile Range Formula?**

The IQR formula is:  
IQR = Q3 – Q1

Where Q3 is the upper quartile and Q1 is the lower quartile.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/boxplot4.bmp)

### **IQR as a test for normal distribution**

Use the interquartile range formula with the mean and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) to test whether or not a population has a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/). The formula to determine whether or not a population is normally distributed are:

Q1 – (σ z1) + X  
Q3 – (σ z3) + X

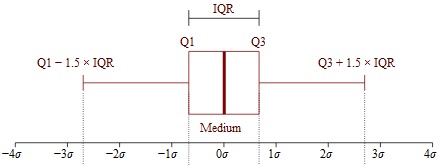
Where Q1 is the first quartile, Q3 is the third quartile, σ is the standard deviation, z is the standard score (“[z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)“) and X is the mean. In order to tell whether a population is normally distributed, solve both equations and then compare the results. If there is a significant difference between the results and the first or third quartiles, then the population is not normally distributed.

### **What is an Interquartile Range Used For?**

The IQR is used to measure how spread out the data points in a set are from the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)of the data set. The higher the IQR, the more spread out the data points; in contrast, the smaller the IQR, the more bunched up the data points are around the mean. The IQR range is one of many measurements used to measure how spread out the data points in a data set are. It is best used with other measurements such as the median and total range to build a complete picture of a data set’s tendency to cluster around its mean.

## **Interquartile Mean (IQM) / Midmean**

The interquartile mean (IQM) is the mean of the middle 50 percent of data in a data set. Unlike the “regular” [arithmetic mean](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/), it is [resistant](https://www.statisticshowto.datasciencecentral.com/resistance-resistant-measures/)to [outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/03/iqr.jpg)

*The IQM is the mean of the interquartile range (IQR).*

### **How to Find the Interquartile Mean**

The calculation is different depending on if your data is divisible by 4 or not.

### **Data is Divisible by Four**

**Example question:**Find the IQM for the following data set:  
5 6 17 30 44 55 56 8 9 11 13 15 1 3 16 65

Step 1: Sort the data from smallest to largest:  
1 3 5 6 8 9 11 13 15 16 17 30 44 55 56 65

Step 2: Discard the bottom 25% and top 25% of numbers. In other words, split the data set into quarters and remove the top and bottom quarters:  
| 8 9 11 13 | 15 16 17 30 |

Step 3: [Find the mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#video) of the remaining numbers:  
(8 + 9 + 11 + 13 + 15 + 16 + 17 + 30 ) / 8 = 14.875

That’s it!

### **Data is NOT Divisible by Four**

**Example question:**Find the IQM for the following data set:  
6 17 30 44 55 56 8 9 11 13 15 1 3 16 65

Step 1: Sort the data from smallest to largest:  
1 3 6 8 9 11 13 15 16 17 30 44 55 56 65

Step 2: Divide the number of items in the set by four. The set has 15 items, so 15/4 = 3.75.

Step 3: Remove the [whole number](https://www.statisticshowto.datasciencecentral.com/whole-number/) (Step 2) from the bottom and the top of the set. For this example, the whole number is 3 (from 3.75):  
 8 9 11 13 15 16 17 30 44   
which leaves:  
8 9 11 13 15 16 17 30 44

Step 4: Figure out how many items are in the [interquartile range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/). The IQR is the middle two quarters, so there would be 3.75 \* 2 = 7.5 numbers.

Step 5: Place parentheses around the middle set of numbers using the whole number from Step 4. In this example, the whole number is 7:  
8 (9 11 13 15 16 17 30) 44

Step 6: Take the fractional part from Step 4 (.5 in this case) and divide it by two (because there are two numbers on the outside of the parentheses):  
.5/2 = .25  
This means that the numbers 8 and 44 will each contribute 25% to the IQM.

Step 7: Multiply the two “outside” numbers (8 and 44 in this case) by the fraction in Step 6:  
8 \* .25 = 2  
44 \* .25 = 11

Step 8: Replace the two outside numbers by the fractional numbers (Step 7) and find the mean. **When dividing by “n”, use the number of items in the IQR from Step 4 (7.5 in this case), not the actual number count (9 in this example):**  
2 (9 11 13 15 16 17 30) 11 =  
(2 + 9 + 11 + 13 + 15 + 16 + 17 + 30 + 11)/7.5 = **16.53.**

## **Semi Interquartile Range / Quartile Deviation**

The **semi interquartile range (SIR)** (also called the quartile deviation) is a[measure of spread](https://www.statisticshowto.datasciencecentral.com/measures-of-spread/). It tells you something about how data is dispersed around a central point (usually the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)). The SIR is half of the [interquartile range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/).

### **How to Calculate the Semi Interquartile Range / Quartile Deviation**

As the SIR is half of the Interquartile Range, all you need to do is[find the IQR](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/#interquartile%20range%20by%20hand)and then divide your answer by 2.

Another way is to use the **quartile deviation formula**:  
[semi interquartile range](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2017/03/semi-interquartile-range.png)  
  
**Note**: You might see the formula QD = 1/2(Q3 – Q1). Algebraically they are the same.

**Breaking down the above formula:**

Step 1: Find the **first**quartile, Q1. If you’re given Q1 in the question, great. If not, you’ve got several options, including:

1. Use a calculator, like [this one](https://www.statisticshowto.datasciencecentral.com/calculators/interquartile-range-calculator/). Plug in your numbers and click the blue button. Q1 is equal to the 25th percentile listed in the results.
2. Follow [*these instructions to find the interquartile range by hand*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/#interquartile%20range%20by%20hand) (part of the process is to find [*quartiles*](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/)).

Step 2: Find the **third**quartile, Q3. If you’re given Q3 in the question, great. If not, use one of the options listed in Step 1. If you choose to use the calculator, Q3 is equal to the 75th percentile.

Step 3: Subtract Step 1 from Step 2.

Step 4: Divide by 2.

## **Example**

**Question**: Find the Quartile Deviation for the following set of data:  
{490, 540, 590, 600, 620, 650, 680, 770, 830, 840, 890, 900}

Step 1: Find the **first**quartile, Q1.  
This is the median of the lower half of the set {490, 540, 590, 600, 620, 650}.  
Q1 = (590 + 600) / 2 = 595.

Step 2: Find the **third** quartile, Q3.  
This is the median of the upper half of the set {680, 770, 830, 840, 890, 900}.  
Q3 = (830 + 840) / 2 = 835.

Step 3: Subtract Step 1 from Step 2.  
835 – 595 = 240.

Step 4: Divide by 2. 240 / 2 = 120

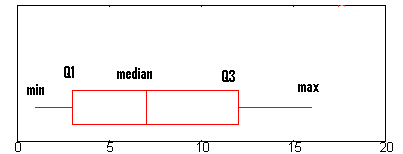
The quartile deviation for this set of data is 12.

### **Coefficient of Quartile Deviation**

The coefficient of quartile deviation (sometimes called the quartile coefficient of dispersion) allows you to compare dispersion for two or more sets of data. The formula is:  
[https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2017/03/quartile-coefficient-of-dispersion.png](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2017/03/quartile-coefficient-of-dispersion.png)

Q3+Q1 = Range  
  
If one set of data has a larger coefficient of quartile deviation than another set, then that data set’s interquartile dispersion is greater.

## **Find a Five-Number Summary in Statistics**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/boxplot4.bmp)

The five number summary includes 5 items:

* The minimum.
* Q1 (the first [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/), or the 25% mark).
* The [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median).
* Q3 (the third [quartile](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/), or the 75% mark).
* The maximum.

The five number summary gives you a rough idea about what your data set looks like. for example, you’ll have your lowest value (the minimum) and the highest value (the maximum). **Although it’s useful in itself, the main reason you’ll want to find a five-number summary is to find more useful statistics, like the**[**interquartile range**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/)**, sometimes called the**[**middle fifty**](https://www.statisticshowto.datasciencecentral.com/middle-fifty/).

[Watch the video](https://youtu.be/omOSu7_Z22o) or read the steps below:

### **How to Find a Five-Number Summary: Steps**

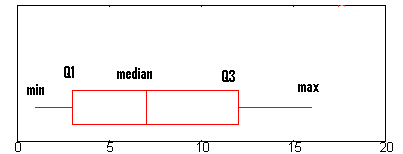
* **Step 1:**Put your numbers in ascending order (from smallest to largest). For this particular data set, the order is:  
  Example: 1,2,5,6,7,9,12,15,18,19,27.
* **Step 2:**Find the minimum and maximum for your data set. Now that your numbers are in order, this should be easy to spot.  
  In the example in step 1, the minimum (the smallest number) is 1 and the maximum (the largest number) is 27.
* **Step 3:**Find the [*median*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median). The median is the middle number. If you aren’t sure how to find the median, see: [How to find the mean mode and median.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/)
* **Step 4:**Place parentheses around the numbers ***above and below*** the median.  
  (This is not technically necessary, but it makes Q1 and Q3 easier to find).  
  (1,2,5,6,7),9,(12,15,18,19,27).
* **Step 5:**Find Q1 and Q3. Q1 can be thought of as a median in the lower half of the data, and Q3 can be thought of as a median for the upper half of data.  
  (1,2,**5**,6,7)**,  9**, ( 12,15,**18**,19,27).
* **Step 6:**Write down your summary found in the above steps.  
  minimum=1,  Q1 =5, median=9,  Q3=18, and maximum=27.

### **When the Summary doesn’t exist**

Sometimes, it’s impossible to find a five-number summary. In order for the five numbers to exist, your data set must meet these two requirements:

* Your data must be [**univariate**](https://www.statisticshowto.datasciencecentral.com/univariate/). In other words, the data must be a single [variable](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/types-of-variables/). For example, this list of weights is one variable: 120, 100, 130, 145. If you have a list of ages and you want to compare the ages to weights, it becomes [bivariate data](https://www.statisticshowto.datasciencecentral.com/bivariate-analysis/) (two variables). For example: age 1 (25 pounds), 5 (60 pounds), 15 (129 pounds). The matching pairs makes it impossible to find a five number summary.
* Your data must be [ordinal](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/), [interval](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/#interval), or [ratio](https://www.statisticshowto.datasciencecentral.com/nominal-ordinal-interval-ratio/#ratio).

### **Box and whisker chart**

A [box and whiskers chart](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/#definition) is a visual representation of the summary.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/boxplot4.bmp)

## **Measures of Position**

**Measures of position** give us a way to see where a certain data point or value falls in a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)or [distribution](https://www.statisticshowto.datasciencecentral.com/probability-distribution/). A measure can tell us whether a value is about the [average](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/average/), or whether it’s unusually high or low. Measures of position are used for [quantitative data](https://www.statisticshowto.datasciencecentral.com/quantitative-variables-data/) that falls on some numerical scale. Sometimes, measures can be applied to [ordinal variables](https://www.statisticshowto.datasciencecentral.com/ordinal-numbers/)— those variables that have an order, like first, second…fiftieth.

Measures of position can also show how two values from different distributions or [measurement scales](https://www.statisticshowto.datasciencecentral.com/scales-of-measurement/) compare. For example, a person’s height (measured in feet) and weight (measured in pounds) can be compared by converting the measurements to [z-scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).

### **Common Measures of Position**

1. Box and Whiskers Plot,
2. Deciles,
3. Five Number Summary,
4. Interquartile Range (IQR),
5. Outliers,
6. Percentiles,
7. Quartiles,
8. Standard scores (i.e. z-scores),
9. Tukey’s upper hinge and lower hinge.

## **1. Box and Whiskers Plot**

A [box and whiskers plot](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/box-plot/) shows the spread and center of data. It is a graphical representation of the [five number summary](https://www.statisticshowto.datasciencecentral.com/how-to-find-a-five-number-summary-in-statistics/): minimum, maximum, [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median), and the first and third [quartiles](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/).

## **2. Deciles**

[Deciles](https://www.statisticshowto.datasciencecentral.com/decile/)are similar to quartiles. But where quartiles split the data in four equal parts, deciles split the data into ten parts: The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th and 100th percentiles.

## **3. Five Number Summary**

The [five number summary](https://www.statisticshowto.datasciencecentral.com/how-to-find-a-five-number-summary-in-statistics/) is an overview of your data. The statistics in the summary are the smallest value (minimum), the largest (maximum), the middle (median) and the first and third quartiles.

## **4. Interquartile Range (IQR)**

The [interquartile range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/interquartile-range/)tells you where the “middle fifty” is in a data set. While the [range](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/)tells you where the beginning and end are in a set, the IQR shows you where the bulk of the “middling” values lie.

## **5. Outliers**

[Outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/)are unusual values that fall outside of an expected range of values. For example, if you’re measuring IQ values of children, your statistics would be thrown off if Einstein and Stephen Hawking were in your class: their IQs would be outliers.

## **6. Percentiles**

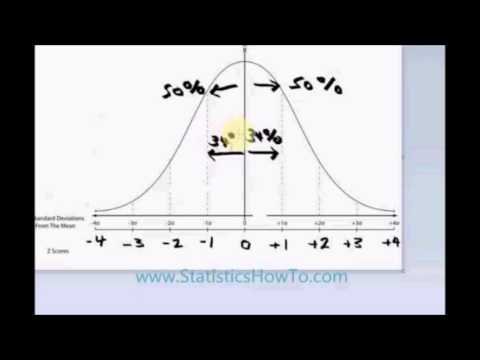
A [percentile](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/)is a number where a certain percentage of scores fall below that number. For example, a 90th percentile marks the spot where 90% of values fall below that cut-off point.

## **7. Quartiles**

Simply put, [quartiles](https://www.statisticshowto.datasciencecentral.com/what-are-quartiles/)divide your data into quarters: the lowest quarter, two middle quarters, and a highest quarter.

## **8. Standard scores (i.e. z-scores)**

[Z-scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)are a way to compare results from a test to a “normal” population.

[](https://www.youtube.com/watch?v=Gp4s5Ouk1gM)

## **9. Tukey’s upper hinge and lower hinge**

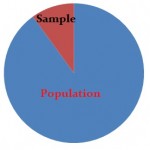
[Tukey’s upper hinge and lower hinge](https://www.statisticshowto.datasciencecentral.com/upper-hinge-lower-hinge/) are created when you split a data set into four pieces (with three hinges). As the median is included in this “splitting,” Tukey’s hinges are sometimes called inclusive quartiles.

## **Population in Statistics**

In stats, a sample is a **part of a population**. A population is a whole, it’s every member of a group. A population is the opposite to a sample, which is a fraction or percentage of a group. Sometimes it’s possible to survey every member of a group. A classic example is the [Census](https://www.statisticshowto.datasciencecentral.com/what-is-a-census/), where it’s the law that you have to respond. Note: if you do manage to survey everyone, it actually is called a census: The U.S. Census is just one example of a census.

In most cases, it’s impractical to survey everyone. Imagine how long it would take you to call every dog owner in the U.S. to find out what their preferred brand of dog food was. In addition, sometimes people either don’t want to respond or forget to respond, leading to incomplete censuses. Incomplete censuses become samples by definition.

**Sample vs. Population Example**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/10-percent-condition.jpg)

If you go into a candy store, the owner might have **samples of their products** on display. It wouldn’t be possible for you to sample everything in the store; Financially the owner wouldn’t want you to taste everything for free. And you probably wouldn’t want to eat a sample of candy from a couple hundred jars or you might get sick to your stomach. So, you might base your opinion about the entire store’s candy line based on the samples they have to offer. The same logic holds true for most surveys in stats; You’re only going to want to take a sample of the whole population (“population” in this example would be the entire candy line). The result is a **statistic about that population.**

### **[Statistic vs. Parameter](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/" \t "_blank).**

A parameter is **data about an entire population.** For example, if you want to find out which classes freshmen at a certain college were taking, you could ask everyone (perhaps via email) and it would be possible to get a parameter. Statistics are when you base your data from samples. For example, you might ask 20 percent of the freshman class what classes they are taking and use that data to make assumptions about what everyone is taking. Obviously, if you base your results from a bit of the population, your results aren’t going to be perfect. That’s where we talk about [**margins of error**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/) and [**confidence intervals**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) in stats. In the candy store, you might be able to get a good feel for the candy line if you taste a few samples, but how confident are you that you can accurately say if your sampling wasn’t skewed? Perhaps the candy that day was extra fresh and tasted wonderful, or perhaps the flavours offered were ones that you didn’t care for. If you had the opportunity to taste test everything, you could offer an excellent opinion about the parameters of the candy line, but with sampling, all you have is a statistic.

**Sample in Statistics**

In statistics, you’ll be working with samples. A sample is just a part of a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, if you want to find out how much the average American earns, you aren’t going to want to survey everyone in the population (over 300 million people), so you would choose a small number of people in the population. For example, you might select 10,000 people In statistics, you’ll be working with samples. A sample is just a part of a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, if you want to find out how much the average American earns, you aren’t going to want to survey everyone in the population (over 300 million people), so you would choose a small number of people in the population. For example, you might select 10,000 people.

* [**Probability Sampling**](https://www.statisticshowto.datasciencecentral.com/probability-sampling/) uses [randomization](#Randomization)to select sample members. You know the probability of each potential member’s inclusion in the sample. For example, 1/100. However, it isn’t necessary for the odds to be equal. Some members might have a 1/100 chance of being chosen, others might have 1/50.
* [**Non-probability sampling**](https://www.statisticshowto.datasciencecentral.com/non-probability-sampling/) uses non-random techniques (i.e. the judgment of the researcher). You can’t calculate the odds of any particular item, person or thing being included in your sample.

**Finding a Sample**

Technically, you can’t just choose 10,000 people. In order for it to be **statistical**(i.e. one that you can use in statistics), the actual size **must be found using a statistical method**. Ten thousand people might not be the optimal amount for valid survey results: you may need more, or less. There are many, many ways to find sample sizes, including using data from prior experiments or using a [size calculator](http://www.nss.gov.au/nss/home.nsf/NSS/0A4A642C712719DCCA2571AB00243DC6?opendocument). How you find a sample size can be quite complex, depending on what you want to do with your data. You can find out more about how to find them here: [Sample size: How to find it](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/).

**Common Sampling Types**

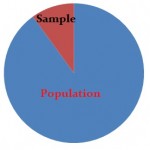
The most common techniques you’ll likely meet in [elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/) or AP statistics include [taking a sample with and without replacement](https://www.statisticshowto.datasciencecentral.com/sampling-with-replacement-without/). Specific techniques include:

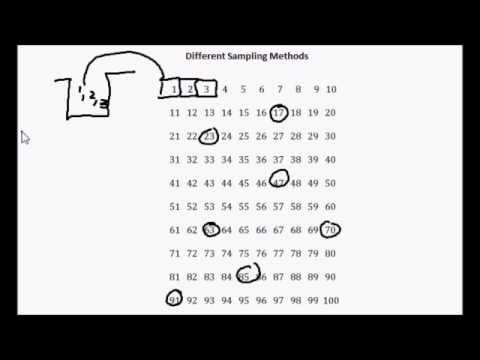
* [**Bernoulli samples**](https://www.statisticshowto.datasciencecentral.com/bernoulli-sampling/) have independent [Bernoulli trials](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/#trial) on population elements. The trials decide whether the element becomes part of the [sample](https://www.statisticshowto.datasciencecentral.com/sample/). All population elements have an equal chance of being included in each choice of a single sample. The sample sizes in Bernoulli samples follow a [binomial distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/)(unequal samples in size). **Poisson samples**(less common): An independent Bernoulli trial decides if each population element makes it to the sample.
* [**Cluster samples**](https://www.statisticshowto.datasciencecentral.com/what-is-cluster-sampling/) divide the population into groups (clusters). Then a [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) is chosen from the clusters. It’s used when researchers don’t know the individuals in a population but do know the population subsets or groups.
* In [**systematic sampling**](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/), you select sample elements from an ordered frame. A [sampling frame](https://www.statisticshowto.datasciencecentral.com/sampling-frame/) is just a list of participants that you want to get a sample from. For example, in the equal-probability method, choose an element from a list and then choose every kth element using the equation k = N\n. Small “n” denotes the sample size and capital “N” equals the size of the population.
* [**SRS**](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/)**(Simple Random Sampling)**: Select items completely randomly, so that each element has the same probability of being chosen as any other element. Each subset of elements has the same probability of being chosen as any other subset of k elements.
* In [**stratified sampling**](https://www.statisticshowto.datasciencecentral.com/stratified-random-sample/), sample each subpopulation independently. First, divide the population into [homogeneous](http://www.merriam-webster.com/dictionary/homogeneous) (very similar) subgroups before getting the sample. Each population member only belongs to one group. Then apply simple random or a systematic method within each group to choose the sample. [**Stratified Randomization**:](https://www.statisticshowto.datasciencecentral.com/stratified-randomization/) a sub-type of stratified used in clinical trials. First, divide patients into strata, then randomize with [permuted block randomization](https://www.statisticshowto.datasciencecentral.com/permuted-block-randomization/).

**Methods**

If you’ve decided to assemble your sample from scratch (for example, you aren’t using prior data), then you need to **choose a sampling method.** Which sampling method you use depends on what resources and information you have available. For example, [the draft](https://en.wikipedia.org/wiki/Conscription_in_the_United_States) worked by drawing random birth dates, a method called [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/). In order for that to work, the government needed a list of every potential draftee’s name and date of birth. The draft could also have used [systematic sampling](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/), drawing the nth name from a list (for example, every 100th name). For that to have worked, all the names must first have been compiled on a list. For more about all the different types of sampling methods, see: [Different Sampling Methods.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/sampling-in-statistics/#diff)

**Different Sampling Methods: How to Tell the Difference**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/10-percent-condition.jpg)  
You’ll come across many terms in statistics that define **different sampling methods**: [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/), [systematic sampling](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/), [stratified random sampling](https://www.statisticshowto.datasciencecentral.com/stratified-random-sample/) and [cluster sampling](https://www.statisticshowto.datasciencecentral.com/what-is-cluster-sampling/). How to tell the difference between the different sampling methods can be a challenge.

[](https://www.youtube.com/watch?v=A7fcdRhSp8k)

**Different Sampling Methods: How to Tell the Difference: Steps**

**Step 1:** Find out if the study sampled from individuals (for example, picked from a pool of people). You’ll find [**simple random sampling**](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) in a school lottery, where individual names are picked out of a hat. But a more “systematic” way of choosing people can be found in “[systematic sampling](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/),” where every nth individual is chosen from a population. For example, every 100th customer at a certain store might receive a “[doorbuster](http://www.wisegeek.com/what-is-a-door-buster.htm)” gift.

**Step 2:** Find out if the study picked groups of participants. For large numbers of people (like the number of potential draftees in the Vietnam war), it’s much simpler to pick people by groups ([**simple random sampling**](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/)). In the case of the draft, draftees were chosen by birth date, “simplifying” the procedure.

**Step 3:** Determine if your study contained data from more than one carefully defined group (“strata” or “cluster”). Some examples of strata could be: Democrats and Republics, Renters and Homeowners, Country Folk vs. City Dwellers, Jacksonville Jaguars fans and San Francisco 49ers fans. If there are two or more very distinct, clear groups, you have a [**stratified sample**](https://www.statisticshowto.datasciencecentral.com/stratified-random-sample/) or a “[cluster sample](https://www.statisticshowto.datasciencecentral.com/what-is-cluster-sampling/).”

* If you have data about the individuals in the groups, that’s a stratified sample. In order to perform stratified sampling on this sample, you could perform random sampling of each strata independently.
* If you only have data about the groups themselves (you may only know the location of the individuals), then that’s a **cluster sample**.

**Step 4:** Find out if the sample was easy to get. **Convenience samples** are like convenience stores: why go out of your way to get samples, when you can nip out to the corner store? A classic example of convenience sampling is standing at a shopping mall, asking passers by for their opinion.

**Sampling Error**

Errors happen when you take a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)from the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)rather than using the entire population. In other words, it’s the difference between the [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)you measure and the [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)you would find if you took a [census](https://www.statisticshowto.datasciencecentral.com/what-is-a-census/)of the entire population.

If you were to survey the entire population (like the [US Census](https://www.census.gov/)), there would be no error. It’s nearly impossible to calculate the error margin. However, when you take samples at random, you estimate the error and call it the [margin of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/#WhatMofE).

For example, if you wanted to figure out how many people out of a thousand were under 18, and you came up with the figure 19.357%. If the actual percentage equals 19.300%, the difference (19.357 – 19.300) of 0.57 or 3% = the margin of error. If you continued to take samples of 1,000 people, you’d probably get slightly different statistics, 19.1%, 18.9%, 19.5% etc, but they would all be around the same figure. This is one of the reasons that you’ll often see sample sizes of 1,000 or 1,500 in surveys: they produce a very acceptable margin of error of about 3%.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/census.jpg)

*A well planned survey can reduce error.*

***Formula:****the formula for the margin of error is 1/√n, where n is the size of the sample. For example, a*[*random sample*](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/)*of 1,000 has about a 1/√n; = 3.2% error.*

Sample error can only be reduced, this is because it is considered to be an acceptable trade-off to avoid measuring the entire population. In general, the larger the sample, the smaller the margin of error. There is a notable exception: if you use [cluster sampling](https://www.statisticshowto.datasciencecentral.com/what-is-cluster-sampling/), this may increase the error because of the similarities between cluster members. A carefully designed experiment or survey can also reduce error.

**Another Type of Error**

The **non-sampling** error could be one reason as to why there’s a difference between the sample and the population. This is due to poor [data collection methods](#Data_Collection_Methods) like faulty instruments or inaccurate data recording, [selection bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/), [non response bias](https://www.statisticshowto.datasciencecentral.com/non-response-bias/) (where individuals don’t want to or can’t respond to a survey), or other mistakes in collecting the data. Increasing the sample size will not reduce these errors. They key is to avoid making the errors in the first place with a well-planned design for the survey or experiment.

## **Sample Size**

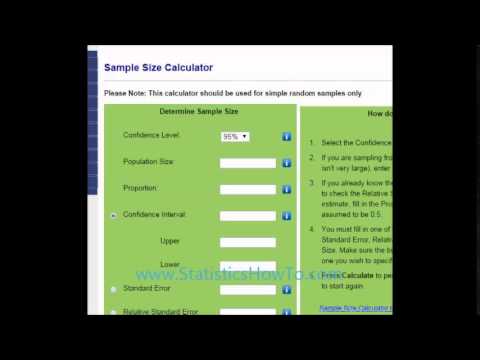
A sample size is a **part of the**[**population**](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/) chosen for a survey or experiment. For example, you might take a survey of dog owner’s brand preferences. You won’t want to survey all the millions of dog owners in the country (either because it’s too expensive or time consuming), so you take a sample size. That may be several thousand owners. The sample size is a representation of all dog owner’s brand preferences. If you choose your sample wisely, it will be a good representation.

### **When Error can Creep in**

When you only survey a small sample of the population,[**uncertainty**](https://www.statisticshowto.datasciencecentral.com/uncertainty-in-statistics/)**creeps in to your statistics**. If you can only survey a certain percentage of the true population, you can never be 100% sure that your statistics are a complete and accurate representation of the population. This uncertainty is called [sampling error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/sampling-in-statistics/#Serror) and is usually measured by a [**confidence interval**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/). For example, you might state that your results are at a 90% confidence level. That means if you were to repeat your survey over and over, 90% of the time your would get the same results.

## **How to Find a Sample Size in Statistics**

Watch the video or read on below:

**[](https://www.youtube.com/watch?v=kmbIobVN1pg)**

A sample is a percentage of the total [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/) in statistics. You can use the data from a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)to make inferences about a population as a whole. For example, the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of a sample can be used to approximate the standard deviation of a population. Finding a sample size can be one of the most challenging tasks in statistics and depends upon many factors including the size of your original population.

## **General Tips**

Step 1: **Conduct a census** if you have a small population. A “small” population will depend on your budget and time constraints. For example, it may take a day to take a census of a student body at a small private university of 1,000 students but you may not have the time to survey 10,000 students at a large state university.

Step 2: **Use a sample size from a similar study.** Chances are, your type of study has already been undertaken by someone else. You’ll need access to academic databases to search for a study (usually your school or college will have access). A pitfall: you’ll be relying on someone else correctly calculating the sample size. Any errors they have made in their calculations will transfer over to your study.

Step 3: **Use a table**to find your sample size. If you have a fairly generic study, then there is probably a table for it. For example, if you have a clinical study, you may be able to use a table published in [Machin et. al’s Sample Size Tables for Clinical Studies, Third Edition](http://onlinelibrary.wiley.com/book/10.1002/9781444300710).

Step 4: **Use a**[**sample size calculator**](http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator), like this one.

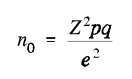
Step 5: **Use a formula**. There are many different formulas you can use, depending on what you know (or don’t know) about your population. If you know some parameters about your population (like a known standard deviation), you can use the techniques below. If you don’t know much about your population, use [Slovin’s formula.](https://www.statisticshowto.datasciencecentral.com/how-to-use-slovins-formula/" \t "_blank).

## **Cochran’s Sample Size Formula**

The Cochran formula allows you to calculate an ideal sample size given a desired level of precision, desired [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/), and the estimated proportion of the attribute present in the population.

Cochran’s formula is considered especially appropriate in situations with**large populations.** A sample of any given size provides more information about a smaller population than a larger one, so there’s a ‘correction’ through which the number given by Cochran’s formula can be reduced if the whole population is relatively small.

The Cochran formula is:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2018/01/cochran-1.jpeg)  
  
  
Where:

* e is the desired level of precision (i.e. the [margin of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/)),
* p is the (estimated) proportion of the population which has the attribute in question,
* q is 1 – p.

## The[z-value](https://www.statisticshowto.datasciencecentral.com/percentile-z-score/)is found in a [Z table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/).

## **Cochran’s Formula Example**

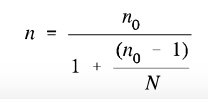
Suppose we are doing a study on the inhabitants of a large town, and want to find out how many households serve breakfast in the mornings. We don’t have much information on the subject to begin with, so we’re going to assume that half of the families serve breakfast: this gives us maximum variability. So p = 0.5. Now let’s say we want 95% confidence, and at least 5 percent—plus or minus—precision. A 95 % confidence level gives us Z values of 1.96, per the normal tables, so we get

((1.96)2 (0.5) (0.5)) / (0.05)2 = 385.

So a random sample of 385 households in our target population should be enough to give us the confidence levels we need.

## **Modification for the Cochran Formula for Sample Size Calculation In Smaller Populations**

If the population we’re studying is small, we can modify the sample size we calculated in the above formula by using this equation:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2018/01/cochran-2.jpeg)  
  
  
Here n0 is Cochran’s sample size recommendation, N is the population size, and n is the new, adjusted sample size. In our earlier example, if there were just 1000 households in the target population, we would calculate

385 / (1+( 384 / 1000 )) = 278

So for this smaller population, all we need are 278 households in our sample; a substantially smaller sample size

## **How to Find a Sample Size Given a Confidence Interval and Width (unknown population standard deviation)**

Part two shows you how to find a sample size for a given **confidence interval and width** (e.g. 95% interval, 6% wide) for an ***unknown population standard deviation***.

Sample question: 41% of Jacksonville residents said that they had been in a hurricane. How many adults should be surveyed to estimate the true proportion of adults who have been in a hurricane, with a 95% confidence interval 6% wide?

**Step 1:** Using the data given in the question, figure out the following variables:

* **za/2**: Divide the confidence interval by two, and look that area up in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/):  
  .95 / 2 = 0.475

The closest [z-score](#Z_score2percentile) for [0.475 is 1.96](https://www.google.com/search?q=0.475+z+table&rlz=1C1CHBF_enIN811IN811&sxsrf=ACYBGNQ3_uPBjHEb0d-GRKaIHOy2tmSlsg:1567789651244&tbm=isch&source=iu&ictx=1&fir=vdtHwtOpQH5ozM%253A%252CWQMy8Mch9hRbxM%252C_&vet=1&usg=AI4_-kTFhYFTncMJytU5wc69QVU_LrYwgA&sa=X&ved=2ahUKEwjjtPrJ17zkAhVu4HMBHVYmBTYQ9QEwAXoECAgQCQ#imgrc=vdtHwtOpQH5ozM:&vet=1)

* **E** ([margin of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/#WhatMofE)):  Divide the given width by 2.  
  6% / 2  
  = 0.06 / 2  
  = **0.03**
* phat: use the given percentage.  41% = **0.41**. If you aren’t given phat, use 50%.
* qhat :  subtract phatfrom 1.  
  1 – 0.41 = **0.59**

**Step 2:**Multiply *phat*by *qhat*. Set this number aside for a moment.  
0.41 × 0.59 =  **0.2419**

**Step 3:** Divide ***Za/2*** by ***E***.  
1.96 / .03 = **65.3333333**

**Step 4:** Square Step 3:  
65.3333333 × 65.3333333 = **4268.44444**

**Step 5:** Multiply Step 2 by Step 4:  
0.2419 × 4268.44444 = **1,032.53671**  
= **1,033 people to survey**.

## **How to Find a Sample Size Given a Confidence Interval and Width (known population standard deviation)**

Part 3 shows you how to determine the appropriate sample size for a given **confidence interval and width**, given that you know the population **standard deviation**.

Sample question: Suppose we want to know the average age of an Florida State College student, plus or minus 0.5 years. We’d like to be 99% confident about our result. From a previous study, we know that the standard deviation for the population is 2.9.

**Step 1:**Find ***z a/2***by dividing the [*confidence interval*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) by two, and looking that area up in the[*z-table*](https://www.statisticshowto.datasciencecentral.com/tables/z-table/):  
.99/2 = 0.495.  The closest z-score for 0.495 is 2.58**.**

**Step 2:**Multiply step 1 by the standard deviation.  
2.58 \* 2.9 = 7.482

**Step 3:**Divide Step 2 by the margin of error. Our margin of error (from the question), is 0.5.7.482/0.5 = 14.96

**Step 4:**Square Step 3.  
14.96 \* 14.96 = 223.8016

That’s it!

## **How to find a sample size in Excel.**

Watch the video or read the steps below:

## 

## **How to use Excel Sampling to find a Sample**

If you have a set of data and you know your sample size, you can use Excel’s [Data Analysis toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/) to select either a periodic sample or a random sample. A [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) is just that — randomly selected from your data set. A periodic sample (also called a [systematic sample](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/)) is where Excel chooses the nth data item to include in your sample. For example, if you wanted to choose every 5th number from the following list: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, Excel would return 8 and 13 (the 5th and 10th numbers in order).

If you don’t know what sample size you need, calculate it before using the Data Analysis tool (using the methods outlined at the top of this article). The Data Analysis tool can help you extract a sample, but **it can’t help you decide on the size**. Why? There are many “human” factors that go into selecting a sample size including budget, prior research (you can use a sample size from previous research) and tables constructed from previous research.

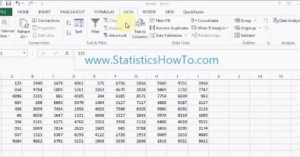
## **How to use Excel Sampling to find a Sample: Steps**

Step 1: Enter your data items into Excel. You can enter your data into rows or columns. Ensure the rows and columns are even; for example, enter data into column A to cell 12 and column B to cell 12.

Step 2: Click “Data” and then click “Data Analysis.” If you don’t see Data Analysis on your toolbar, [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)

.

Step 3: Click “Sampling” and then click “OK.”

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/sampling-in-excel.jpg)

*Data entered into a worksheet for Excel sampling: the rows and columns are even.*

Step 4: Click in the Input Range box and then select your entire data set.

Step 5: Click either “Periodic Sampling” or “Random Sampling.” If you choose periodic, enter the nth number (i.e. every 5) and if you choose random sampling, enter the sample size.

Step 6: Choose an output range. For example, click the “New Worksheet” button and Excel will return the sample in a new worksheet.

Step 7: Click “OK.”

That’s it!

## **Data Collection Methods**

## **Data Collection Methods: General Categories**

Data collection methods fall into four general categories:

1. A [**census**](https://www.statisticshowto.datasciencecentral.com/what-is-a-census/) is a survey of a whole [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, the [U.S. Census](https://www.census.gov/). Censuses can be very expensive and time-consuming, if the population is large.
2. A [**sample survey**](https://www.statisticshowto.datasciencecentral.com/survey-sampling/) takes a fraction of the population. Sample surveys are cheaper than censuses, but are not as accurate. [Bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)can also be an issue. Survey sampling is selecting members from a target population to be in a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)for a sample survey. Usually the survey is some type of questionnaire (i.e. in-person, phone or internet survey).
3. An **experiment**is a controlled study of a group. Experiments are very common in the medical fields. The researcher controls how members are placed study groups and which treatment each group receives. [Bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/) can be a major issue with experiments.
4. An [**observational study**](https://www.statisticshowto.datasciencecentral.com/experimental-design/#Observational)is about the same as an experiment. However, the researcher does not use control groups or assign treatments.

## **What is the best of the Data Collection Methods?**

There is no one “best” data collection method. Each method has its pros and cons. Which one you choose depends on what kind of data you have (i.e. [qualitative data or quantitative data](https://www.statisticshowto.datasciencecentral.com/quantitative-variables-data/)) and which pros/cons are important for your study.

In general, the following data collection methods work for [qualitative research](https://www.statisticshowto.datasciencecentral.com/research-methods-qualitative-research-and-quantitative-research/):

* Document review.
* In depth interviews.
* Observation methods.

[Quantitative research](https://www.statisticshowto.datasciencecentral.com/research-methods-qualitative-research-and-quantitative-research/)data collection methods, which tends to rely on [random samples](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/), include:

* Surveys with closed-ended questions.
* Clinical trials/experiments.
* Extracting data from computer and information systems.
* Observing, counting and recording events that are well defined (for example, counting the number of people who come into a restaurant).

## **Types of Data Collection Methods**

* **Interviews**are very structured for quantitative research. They are less structured for qualitative research. In a structured interview, you would asks= a series of standard questions; in less-structured interview you might ask [open-ended questions](http://www.oucom.ohiou.edu/fd/Open%20Ended%20Questions.htm).
  + **Face-to-face interviews** can be time-consuming and expensive. However, you can establish rapport with an interviewee and may have the opportunity to clarify responses. Face-to-face interviews have a very high response rate.
  + **Telephone Interviews** are cost-effective and take less time than face-to-face interviews. However, the response rate isn’t as high. Another con is that survey bias may creep in. For example, people without phones can’t be surveyed.
  + **Computer Assisted Personal Interviewing** ([*CAPI*](http://sru.soc.surrey.ac.uk/SRU3.html)) is like a personal interview, but you bring a laptop or tablet instead of paper questionnaires to enter the responses. CAPI has the same drawbacks as face-to-face interviews
* **Questionnaires**
  + **Paper and pencil questionnaires** can be sent to a large group of people. They have the advantage of respondents tending to be more truthful with responses. A disadvantage is a very low response rate.
  + **Internet questionnaires** are cheaper and faster than paper questionnaires. But they exclude people who do not have access to computers. In addition, people tend to hurry to complete internet surveys so the responses may not be valid.

## **Experimental Design**

Experimental design is **a way to carefully plan experiments in advance** so that your results are both objective and [valid](https://www.statisticshowto.datasciencecentral.com/reliability-validity-definitions-examples/). The terms **“Experimental Design”** and**“Design of Experiments”** are used interchangeably and mean the same thing. However, the medical and social sciences tend to use the term “Experimental Design” while engineering, industrial and computer sciences favor the term “Design of experiments.”

**Ideally, your experimental design should:**

* Describe how participants are allocated to[experimental groups.](https://www.statisticshowto.datasciencecentral.com/experimental-group/) A common method is completely randomized design, where participants are assigned to groups at random. A second method is randomized block design, where participants are divided into homogeneous blocks (for example, age groups) before being randomly assigned to groups.
* Minimize or eliminate [confounding variables](https://www.statisticshowto.datasciencecentral.com/confounding-variable/), which can offer alternative explanations for the experimental results.
* Allow you to make inferences about the relationship between [independent variables](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/) and[dependent variables](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/).
* Reduce [variability](https://www.statisticshowto.datasciencecentral.com/variability/), to make it easier for you to find differences in [treatment](https://www.statisticshowto.datasciencecentral.com/statistical-treatment/)outcomes.

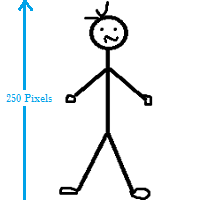
**Design of experiments involves:**

* The systematic collection of data
* A focus on the design itself, rather than the results
* Planning changes to [independent (input) variables](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/) and the effect on [dependent variables](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/) or response variables
* Ensuring results are valid, easily interpreted, and definitive.

**The most important principles** are:

* [Randomization](https://www.statisticshowto.datasciencecentral.com/randomization-experimental-design/): the assignment of study components by a completely random method, like [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/). Randomization eliminates [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)from the results
* [Replication](https://www.statisticshowto.datasciencecentral.com/replication-replicability/): the experiment must be replicable by other researchers. This is usually achieved with the use of statistics like the [standard error of the sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/#SESM) or [confidence intervals](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/).
* Blocking: controlling sources of variation in the experimental results.

## **Uncertainty in Statistics**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/10/greatest-possible-error.png)

*This stick man is about 250 pixels tall. A tiny amount of error (perhaps one half of a pixel) is built in.*

In the real word, uncertainty (sometimes called error or bias) is a part of everyday life, but in statistics **we try to quantify just how much uncertainty is in our experiment, survey or test results**. The two main types are epistemic (things we don’t known because of a lack of data or experience) and aleatoric (things that are simply unknown, like what number a die will show on the next roll).

It is measured through a variety of ways.

## **Measures of Uncertainty**

The [**confidence interval (CI)**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) shows what the uncertainty is with a certain [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)(e.g. a [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/)). The [**margin of error**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/) is a range of values above and below a confidence interval’s sample statistic. For example, a survey might report a 95% confidence level of between 4.88 and 5.26. That means if the survey is repeated using the same methods, 95% of the time the true population [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)will fall between 4.88 and 5.26, 98% of the time.

The [mean error](https://www.statisticshowto.datasciencecentral.com/mean-error/)refers to the mean (average) of all errors. “Error” in this context is the difference between a measured and true value.

For**what happens to**[**measurement errors**](https://www.statisticshowto.datasciencecentral.com/measurement-error/)when you use uncertain measurements to calculate something else (For example, using length to calculate area), see: [Propagation of Uncertainty](https://www.statisticshowto.datasciencecentral.com/error-propogation/). In general terms, [relative precision](https://www.statisticshowto.datasciencecentral.com/relative-precision/) shows uncertainty as a fraction of a quantity. It is the ratio of a measurement’s precision and the measurement itself.

For the **entropy of a distribution** where a row variable X explains a column variable Y, see: [Uncertainty Coefficient](https://www.statisticshowto.datasciencecentral.com/uncertainty-coefficient/).

## **Sources of Uncertainty**

* **Interpolation errors** happen because of a lack of data, and may be compounded by your choice of interpolation method.
* **Model bias** happens because any model is an approximation, or a best guess at what a true distribution might look like.
* **Numerical errors** are human errors that creep in when translating mathematical models into a computer.
* **Observational error** is due to the variability of measurements in an experiment.
* **Parameter uncertainty** happens because we don’t know the exact, or “best” values in a population—we can only take a good guess with sampling.

## **Margin of Error**

The **margin of error**is the range of values below and above the [sample statistic](https://www.statisticshowto.datasciencecentral.com/sample-statistic-definition-examples/) in a [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/). The confidence interval is a way to show what the [**uncertainty**](https://www.statisticshowto.datasciencecentral.com/uncertainty-in-statistics/) is with a certain [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)(i.e. from a poll or survey). For example, a poll might state that there is a 98% confidence interval of 4.88 and 5.26. That means if the poll is repeated using the same techniques, 98% of the time the true population parameter ([parameter vs. statistic](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/)) will fall within the interval estimates (i.e. 4.88 and 5.26) 98% of the time.

## **Margin of Error Percentage**

A margin of error percentage tells you **how many percentage points your results will differ**from the real population value. For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.

The Margin of Error can be calculated in two ways:

1. Margin of error = [Critical value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/) x [**Standard deviation**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)
2. Margin of error = Critical value x [**Standard error**](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/) of the statistic

## **Statistics Aren’t Always Right!**

The idea behind confidence levels and margins of error is that any survey or poll will differ from the true [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)by a certain amount. However, [confidence intervals](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/) and margins of error reflect the fact that there is room for error, so although 95% or 98% confidence with a 2 percent Margin of Error might sound like a very good statistic, room for error is built in, which means sometimes statistics are wrong. For example, [a Gallup poll](http://www.gallup.com/poll/158519/romney-obama-gallup-final-election-survey.aspx) in 2012 (incorrectly) stated that Romney would win the 2012 election with Romney at 49% and Obama at 48%. The stated confidence level was 95% with a margin of error of +/- 2, which means that the results were calculated to be accurate to within 2 percentages points 95% of the time.

The [real results from the election](http://uselectionatlas.org/RESULTS/national.php)were: Obama 51%, Romney 47%, which was actually even outside the range of the Gallup poll’s margin of error (2 percent), showing that not only can statistics be wrong, but polls can be too.

## **How to Calculate Margin of Error**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=MV_BwPHWGh8)

The margin of error tells you the **range of values** above and below a [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/vote.jpg)

*Margins of error are commonly used in election polls.*

A poll might report that a certain candidate is going to win an election with 51 percent of the vote; The [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/) is 95 percent and the error is 4 percent. Let’s say the poll was repeated using the same techniques. The pollsters would expect the results to be within 4 percent of the stated result (51 percent) 95 percent of the time. In other words, 95 percent of the time they would expect the results to be between:  
51 – 4 = 47 percent and  
51 + 4 = 55 percent.  
**The margin of error can be calculated in two ways, depending on whether you have**[**parameters**](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)**from a population or**[**statistics**](https://www.statisticshowto.datasciencecentral.com/statistic/)**from a sample**:

1. Margin of error = Critical value x [Standard deviation](https://www.statisticshowto.datasciencecentral.com/relative-standard-deviation/) for the population.
2. Margin of error = Critical value x [Standard error](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/) of the sample.

## **How to Calculate Margin of Error: Steps**

Step 1: **Find the critical value**. The critical value is either a [**t-score**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/) or a **z-score**. If you aren’t sure, see: [T-score vs z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/t-score-vs-z-score/). In general, for small [sample sizes](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) (under 30) or when you don’t know the population [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/), use a [t-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/). Otherwise, use a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).

[Click here for a minute video that shows you how to find a critical value.](https://www.youtube.com/watch?v=RAnFyF_6zHk)

Step 2: **Find the**[**Standard Deviation**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)**or the**[**Standard Error**](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/)**.** These are essentially the same thing, only you must know your population [parameters](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)in order to calculate standard deviation. Otherwise, calculate the standard error (see: [What is the Standard Error?](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/)).  
[Click here](https://www.youtube.com/watch?v=aBXJnvQ6KFk)for a short video on how to calculate the standard error.

Step 3: **Multiply the**[**critical value**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/)from Step 1**by the**[**standard deviation**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) or [standard error](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/) from Step 2. For example, if your CV is 1.95 and your SE is 0.019, then:  
1.95 \* 0.019 = 0.03705

**Sample question:** 900 students were surveyed and had an average GPA of 2.7 with a standard deviation of 0.4. Calculate the margin of error for a 90% confidence level:

1. The critical value is 1.645 (see [this video](https://www.youtube.com/watch?v=RAnFyF_6zHk) for the calculation)
2. The standard deviation is 0.4 (from the question), but as this is a sample, we need the standard error for the mean. The formula for the SE of the mean is standard deviation / √(sample size), so: 0.4 / √(900)=0.013.
3. 1.645 \* 0.013 = 0.021385

That’s how to calculate margin of error!

**Second example**:[Click here](https://youtu.be/nZjwlCboeV0) to view a second video on YouTube showing calculations for a 95% and 99% Confidence Interval.

**Tip**: You can use the[t-distribution calculator](https://www.statisticshowto.datasciencecentral.com/calculators/tdist-calculator/) on this site to find the t-score and the [variance and standard deviation calculator](https://www.statisticshowto.datasciencecentral.com/calculators/variance-and-standard-deviation-calculator/) will calculate the standard deviation from a sample.

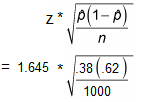
## **Margin of Error for a Proportion**

The formula is a little different for proportions:  
[moe](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/moe.png)  
  
  
Where:  
[phat](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/phat.bmp) = sample proportion (“P-hat”).  
n = sample size  
z = z-score

**Example question:** 1000 people were surveyed and 380 thought that climate change was not caused by human pollution. Find the MoE for a 90% confidence interval.

Step 1: **Find P-hat** by dividing the number of people who responded positively. “Positively” in this sense doesn’t mean that they gave a “Yes” answer; It means that they answered according to the statement in the question. In this case, 380/1000 people (38%) responded positively.

Step 2: **Find the z-score that goes with the given confidence interval.** You’ll need to reference [this chart of common critical values.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/#CommonCI) A 90% confidence interval has a z-score (a critical value) of 1.645.

Step 3: **Insert the values into the formula and solve:**  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/moep-2.png)

= 1.645 \* 0.0153

= 0.0252

Step 4: **Turn Step 3 into a percentage**:  
0.0252 = 2.52%  
The margin of error is 2.52%.

## **Standard error**

The standard error(SE) is very similar to [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). Both are measures of spread. The higher the number, the more spread out your data is. To put it simply, the two terms are essentially equal — but there is one important difference. While the standard error uses **statistics** (sample data) standard deviations use **parameters** (population data). (What is the [difference between a statistic and a parameter?](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/)).

In statistics, you’ll come across terms like “the standard error of the mean” or “the standard error of the median.” The SE tells you how far your [sample statistic](https://www.statisticshowto.datasciencecentral.com/sample-statistic-definition-examples/) (like the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)) deviates from the actual population mean. The larger your sample size, the smaller the SE. In other words, the larger your sample size, the closer your sample mean is to the actual population mean.

## **What is the SE Calculation?**

How you find the standard error depends on what stat you need. For example, the calculation is different for the mean or proportion. When you are asked to find the sample error, you’re probably finding the standard error. That uses the following formula: s/√n. You might be asked to find standard errors for other stats like the mean or proportion.

## **What is the Standard Error Formula?**

The following tables show how to find the standard deviation (first table) and SE (second table). That assumes you know the right population parameters. If you don’t know the population parameters, you can find the standard error:

* Sample mean.
* Sample proportion.
* Difference between means.
* Difference between proportions.

|  |  |
| --- | --- |
| **Parameter (Population)** | **Formula for Standard Deviation.** |
| Sample mean, [xbar](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/09/xbar.bmp) | = σ / sqrt (n) |
| Sample proportion, p | = sqrt [P (1-P) / n) |
| Difference between means. [difference between means](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/difference-between-means.jpg) | = sqrt [σ21/n1 + σ22/n2] |
| Difference between proportions. [difference between proportions](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/difference-between-proportions.jpg) | = sqrt [P1(1-P1)/n1 + P2(1-P2)/n2] |

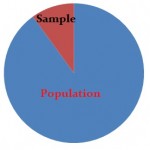
|  |  |
| --- | --- |
| **Statistic (Sample)** | **Formula for Standard Error.** |
| Sample mean, [what is the standard error](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/09/xbar.bmp) | = s / sqrt (n) |
| Sample proportion, p | = sqrt [p (1-p) / n) |
| Difference between means. [difference between means](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/difference-between-means.jpg) | = sqrt [s21/n1 + s22/n2] |
| Difference between proportions. [difference between proportions](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/difference-between-proportions.jpg) | = sqrt [p1(1-p1)/n1 + p2(1-p2)/n2] |

**Key for above tables**:  
P = Proportion of successes. Population.  
p = Proportion of successes. Sample.  
n = Number of observations. Sample.  
n2 = Number of observations. Sample 1.  
n2 = Number of observations. Sample 2.  
σ21 = Variance. Sample 1.  
σ22 = Variance. Sample 2.

## **Bias in Statistics**

Bias is the tendency of a statistic to [**overestimate**](http://www.merriam-webster.com/dictionary/overestimate) or [**underestimate**](http://www.merriam-webster.com/dictionary/underestimate) a parameter. Bias can seep into your results for a slew of reasons including sampling or measurement errors, or [unrepresentative samples](https://www.statisticshowto.datasciencecentral.com/representative-sample/).

## **Sampling Errors**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/10-percent-condition.jpg)

Sampling error is the tendency for a statistic not to **exactly match** the population. Error doesn’t necessarily mean that a mistake was made in your sampling; [**Sampling Variability**](https://www.statisticshowto.datasciencecentral.com/sampling-variability/) could be a more accurate name. For example, let’s say you have a population in the United States with an average height of 5 feet 9 inches. If you take a sample, even a fairly sizable sample of say, 10,000 people, it’s unlikely that you’ll get exactly 5 feet 9 inches. You might get **very close**, perhaps to within a fraction of an inch. If you repeat the experiment, you might get another very close result. For example, in experiment 1 you might get 5 feet 8.9 inches and in experiment 2 you might get 5 feet 9.1 inches. The tendency for statistics to get very close, but not exactly right, is called sampling error. **Note:** If the statistic is unbiased, the average of all statistics from all samples will average the true population parameter.

**Measurement Errors**

[Measurement errors](https://www.statisticshowto.datasciencecentral.com/measurement-error/) are where a provided response is**different from the real value.** For example, you might survey to find out if a person voted for President Obama. A person may have voted for him, but they are confused by the wording of the questionnaire and mistakenly respond that they did not vote for him. Several factors may cause measurement error, including:

* The way the interviewer poses the question.
* The wording on the questionnaire.
* The way the data is collected.
* The respondent’s record-keeping system.

## **Biased Estimator**

In statistics, an[**estimator**](https://www.statisticshowto.datasciencecentral.com/estimator/) is a rule for calculating an estimate of a quantity based on observed data. For example, you might have a rule to calculate a population mean. The result of using the rule is an **estimate** (a statistic) that hopefully is a true reflection of the population. The bias of an estimator is the difference between the statistic’s [expected value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/expected-value/) and the true value of the population parameter. If the statistic is a true reflection of a population parameter it is an [**unbiased**](https://www.statisticshowto.datasciencecentral.com/unbiased/)estimator. If it is not a true reflection of a population parameter it is a **biased estimator**.

The word **bias** in the regular English language implies that you have a personal reason to misrepresent a piece of information. However, in statistics, it doesn’t mean that the interviewer, the researcher or even the respondent in an interview is biased in some way. It just means that the estimator being used doesn’t produce a good estimate.

### **Example of a Biased Estimator**

[](https://www.statisticshowto.datasciencecentral.com/what-is-biased-estimator)

*Photo credit: Bell & Jeff|Flickr.com*

You are playing the party game “[Pin the tail on the donkey](https://en.wikipedia.org/wiki/Pin_the_tail_on_the_donkey).” (If you aren’t familiar with the game, a picture of a donkey is placed on the wall and you are given a paper tail to pin on the donkey while you are blindfolded. The person who pins the tail closest to the actual spot where the real tail should go wins the game). You try six times to pin the tail in the right place and each time you pin the tail in the wrong place, at the bottom or to the front of the donkey. Your estimation for the actual spot where the tail should have gone is a**biased estimator**because you put the tails in the wrong place.

## **Selection Bias**

Ideally, you should randomly select every participant in a survey. But, sometimes biases creep in, whether intentional or unintentional. Selection bias takes away from the “randomness” you are hoping to achieve. It’s usually a result of not using the correct procedures to choose your participants. Types of selection bias include: the healthy worker effect, non-response bias, [undercoverage](https://www.statisticshowto.datasciencecentral.com/undercoverage-definition/), and voluntary response bias.

## **Healthy Worker Effect**

In general, people who are working are healthier than people who are unemployed. The healthy worker effect, a type of [membership bias](https://www.statisticshowto.datasciencecentral.com/membership-bias/), is a particular type of selection bias that happens when you are studying the effects of occupational exposure to a compound, like asbestos, and only include employed persons in your study. These people are less likely to suffer from the effects of exposure than people who are currently unemployed, including people who are disabled due to the exposure.

## **Hospital Patient Bias**

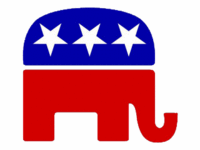
This type of selection bias (also called [Berkson’s Paradox](https://www.statisticshowto.datasciencecentral.com/berksons-paradox-definition/)) is when a [case-control study](https://www.statisticshowto.datasciencecentral.com/case-control-study/) uses hospitalized patients as controls. If those patients are hospitalized due to a connection with the disease being studied, then any measure of the drug or procedure’s effect may weaken.

## **Non-response bias**

[Non-response bias](https://www.statisticshowto.datasciencecentral.com/non-response-bias/) is a type of bias that happens when some people fail to respond to a survey. People may refuse to answer, or lack the time or inclination to answer. for example, you may have a survey about cheating on tax returns; the people most likely to **not** answer are the very people you are trying to reach: cheaters on tax returns. This type of selection bias can also creep in with other types of sensitive information, like questions about prostitution, alcoholism, or illegal drug use. Non-response bias can also become a factor if you haven’t constructed your survey properly. For example, if you have a snail mail survey for young adults or a smartphone survey for older adults; both these scenarios are likely to lead to a lower response rate for your targeted population.

Missing data can be filled in (“imputed”) with procedures like [Multiple Imputation.](https://www.statisticshowto.datasciencecentral.com/multiple-imputation/)

## [**Undercoverage**](https://www.statisticshowto.datasciencecentral.com/undercoverage-definition/)

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/02/selection-bias.gif)

Similar to non-response bias, undercoverage is when your respondents aren’t from the population you hoped for. A classic example is the [Literary digest voter survey,](http://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html) which predicted Franklin Roosevelt would beat by Alfred Landon in the 1936 presidential election. The survey had an undercoverage of low income voters, who were more likely to be Democrats. Ironically, the survey was one of the largest and most expensive surveys undertaken, with a sample of about 2.4 million people; despite this huge number of people, the [sampling error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/sampling-in-statistics/#Serror) was a massive 19 percent.

## **Voluntary response bias**

Some surveys — like call in radio shows — tend to attract very opinionated people. These types of voluntary responses lead to an under-representation of the general population in favor of strong opinions.

## **Volunteer Bias**

This crops up frequently in clinical trials; the people who volunteer for the trials may not represent the population you are trying to target. For example, if your study is for a new drug to treat diabetes and you offer significant compensation, people with low socio-economic background may make up the bulk of volunteers.

## **Survivorship Bias**

Survivorship bias is a type of **selection bias**, which results in a [sample](https://www.statisticshowto.datasciencecentral.com/sample/) that isn’t reflective of the actual [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). With survivorship bias, you concentrate on the “survivors” of a particular process. The concept sounds simple, but in reality it’s tricky to implement.

## **Examples**

Take a simple example. The [probability of getting a full house in poker](http://www.math.hawaii.edu/~ramsey/Probability/PokerHands.html) is 0.001441, or about .01%. But the odds of getting a full house after the cards have been dealt is 1. Imagine for a moment that you know nothing about card games (in real life, we often know little about the phenomenon we’re studying). If you were studying “full house in poker” and only looked at the survivors (the successes). You would conclude, erroneously, that the probability of getting a full house is 1 because you didn’t take into account all of the failures.  
[survivorship bias](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/01/graphics-cards-778788.gif)

### **In Business**

Want to succeed in business? Create a Fortune 100 company? Be the next Bill Gates? If mass-market paperbacks are all true, all you have to do is study how businesses are successful. Some of the [New York Times best sellers](https://www.nytimes.com/best-sellers-books/2014-01-05/business-books/list.html) from recent years include:

* Outliers by Malcolm Gladwell. Why some people succeed.
* Steve Jobs by Walter Isaacson. A biography of the Entrepreneur.
* Various other “How to be a business success” stories.

The fact is success is more likely if you study failures — those businesses that dropped out along the way — rather than successes. The success stories are few and far between and have a lot to do with luck in addition to business know how and a good product. Want to become successful like Steve Jobs? Dropping out of college, lying to get your first job, and experimenting with psychedelics probably have little to do with it.

### **Abraham Wald’s Naval Work**

In World War II, statistician Abraham Wald tried to determine how to minimize bomber losses. Prior to Wald’s work, researchers from the Center for Naval Analysis analyzed bombers that came back with damage and recommended reinforcement of damaged areas on all bombers. However, they didn’t take into account that only the surviving aircraft came back; the bombers that did not survive likely had damage to other, more critical areas. Wald recognized that survivorship bias played a part in the Center for Naval Analysis’s decision and recommended basically the opposite–the reinforcement of areas that had not been hit on the surviving aircraft. You can find an article about Wald’s work [here](http://people.ucsc.edu/~msmangel/Wald.pdf).  
Reference: [UNC.EDU.](http://cphp.sph.unc.edu/trainingpackages/ERIC/eric_notebook_8.pdf)

## **Social Desirability Bias**

Social desirability bias is the tendency to answer questionnaires or surveys according to what is socially acceptable. People tend to report inaccurately on sensitive topics like abortion, drug use, or prostitution. This is usually attributed to embarrassment or lack of comfort in revealing true feelings or attitudes.

Indirect questions (i.e. general, non personal questions) are advisable when dealing with sensitive issues as they tend to make people more honest about their true feelings. Forced choice items and use of proxy subjects can also reduce or prevent this type of bias.

## **Acquiescence Bias**

This type of bias usually happens because people want to be polite or to be agreeable, although it can also happen because people want to skim through a survey quickly. See: [Acquiescence Bias](https://www.statisticshowto.datasciencecentral.com/acquiescence-bias/).

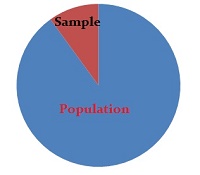
## **Availability Bias**

This type of bias is where you make a probability calculation based on the first thing that comes to mind. Advertisers use it to their advantage. See: [Availability Bias.](https://www.statisticshowto.datasciencecentral.com/availability-bias/)

## **Other Types of Bias**

* [Accidental Bias](https://www.statisticshowto.datasciencecentral.com/accidental-bias/)
* [Aggregation Bias](https://www.statisticshowto.datasciencecentral.com/aggregation-bias/)
* [Allocation Bias](https://www.statisticshowto.datasciencecentral.com/allocation-bias/)
* [Ascertainment Bias](https://www.statisticshowto.datasciencecentral.com/ascertainment-bias/)
* [Assignment Bias](https://www.statisticshowto.datasciencecentral.com/assignment-bias/)
* [Attrition Bias](https://www.statisticshowto.datasciencecentral.com/attrition-bias/).
* [Central Tendency Bias](https://www.statisticshowto.datasciencecentral.com/central-tendency-bias/).
* [Diagnostic Bias](https://www.statisticshowto.datasciencecentral.com/diagnostic-bias/)
* [Funding Bias](https://www.statisticshowto.datasciencecentral.com/funding-bias/)
* [Information Bias](https://www.statisticshowto.datasciencecentral.com/information-bias-observation/)
* [Misclassification Bias](https://www.statisticshowto.datasciencecentral.com/non-differential-misclassification/)
* [Neyman Bias](https://www.statisticshowto.datasciencecentral.com/neyman-bias/)
* [Observer Bias](https://www.statisticshowto.datasciencecentral.com/observer-bias/)
* [Performance Bias](https://www.statisticshowto.datasciencecentral.com/performance-bias/)
* [Publication Bias](https://www.statisticshowto.datasciencecentral.com/publication-bias/)
* [Referral Bias](https://www.statisticshowto.datasciencecentral.com/referral-bias/)
* [Reporting Bias](https://www.statisticshowto.datasciencecentral.com/reporting-bias/)
* [Self-selection bias](https://www.statisticshowto.datasciencecentral.com/self-selection-bias/)
* [Spectrum Bias](https://www.statisticshowto.datasciencecentral.com/spectrum-bias/)
* [Verification Bias](https://www.statisticshowto.datasciencecentral.com/verification-bias/)
* [Wording Bias](https://www.statisticshowto.datasciencecentral.com/wording-bias/)

## **Representative Sample**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/sample-mean-small.jpg)

*Does your sample match the population?*

A **representative sample**is where your [sample](https://www.statisticshowto.datasciencecentral.com/sample/)matches some characteristic of your [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/), usually the characteristic you’re targeting with your research. For example, if you’re conducting a survey about how women’s experience as a data scientist compare to men, you would want your sample to reflect the percentage of women in the data science workforce.

## **Unrepresentative Samples**

The purpose of sampling is to obtain a [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)that tells you something about a population. A statistic is **representative**if it represents the attributes of a known [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)in the population. When the statistic **does not represent the population parameter,** it is called **unrepresentative**. The type of bias that occurs in statistics when there is an unrepresentative sample is called [*selection bias*](https://www.statisticshowto.datasciencecentral.com/what-is-bias/).

**Representative Doesn’t Mean**[**Replication**](https://www.statisticshowto.datasciencecentral.com/replication-replicability/)

Even if the sample is labelled as “representative”, it doesn’t mean that every aspect of the population is included. For example, in [quota sampling](https://www.statisticshowto.datasciencecentral.com/quota-sampling/), you maintain correct proportions present in the population. For example, if your original population is 45% female and 55% male, your quota sample should reflect those percentages. However, your sample tells you nothing about age distribution, income distribution, or other metrics; it’s merely representative of proportions of female and male.

## **Obtaining Representative Samples**

In general [random selection](https://www.statisticshowto.datasciencecentral.com/random-selection-and-assignment/) results in a representative sample; you can make generalizations and predictions about the population as long as you have used a [probability sampling method](https://www.statisticshowto.datasciencecentral.com/probability-sampling/).

In general, you should strive to avoid [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)in your survey, trial or experiment. Although it may seem like a simple task to randomly select from a population to get a representative sample, the reality is the practice is full of pitfalls. For example, the results of your study might become skewed due to factors you didn’t account for, like knowledge of which patients are getting which treatments in clinical trials, or poor [data collection methods](https://www.statisticshowto.datasciencecentral.com/data-collection-methods/). As a researcher, you may delegate responsibility for parts of your study to other people or even to outside sources; you should make sure everyone involved is following your carefully planned procedure to the letter.

One way to avoid unrepresentative samples is to make sure you haven’t excluded certain population members, like minorities or people who work two jobs. If it isn’t possible to get a representative sample (perhaps because of availability of participants), then you should tweak your results to reflect the population. One way to do this is with [**weighting factors**](https://www.statisticshowto.datasciencecentral.com/weighting-factor/), which are used to make samples match the population. For example, let’s say your sample ends up with 95% male and 5% female. [Industry data](https://www.kdnuggets.com/2018/09/growing-participation-women-data-science-community.html)tells you that the percentage of females may be as high as 25%. In order to make sure that you have a representative sample, you could add a little more “weight” to data from females.

On the other hand, some sampling methods are designed in such a way that they can’t result in representative samples. However, they are useful for very specific situations—as long as you understand their limitations. For example, [**convenience sampling**](https://www.statisticshowto.datasciencecentral.com/convenience-sampling/) involves drawing a sample from a convenient, readily available population. Although the sample isn’t representative of the population it can be useful for pilot testing.

## **Sampling variability**

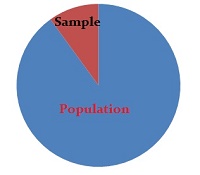
Sampling variability is **how much an estimate varies between samples**. “Variability” is another name for [range](https://www.statisticshowto.datasciencecentral.com/mean/); Variability between samples indicates the range of values differs between samples.

Sampling variability is often written in terms of a [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/). The [variance](https://www.statisticshowto.datasciencecentral.com/statistic/)(σ2) and [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviationion/) (σ) are common measures of variability. You might also see reference to the variability of the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/) (x&772;), which is just another way of saying the sample mean differs from sample to sample. **Sampling variability only refers to a statistic (i.e. a number generated from a sample) — never a population.**

## **Variability and Sampling Error**

A closely related term (almost a synonym) is sampling error. An error in sampling isn’t a mistake — it’s a measure of how much a value differs from the “true” value. Let’s say the true weight of a population is 150 lbs. You take a sample and find the mean weight for the sample is 151 lbs. The 1 lb difference is an “error.” If you sample again, you might get different mean weights of 148 lbs, or 150.5 lbs, or 153 lbs. The different errors — 1/2 lb, 1 lb, 2 lbs, 3 lbs — are a reflection of the variability between your samples, or **sampling variability.**

## **Variability and Sample Sizes**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/sample-mean-small.jpg)

*The “perfect” sample size is practically impossible to find.*

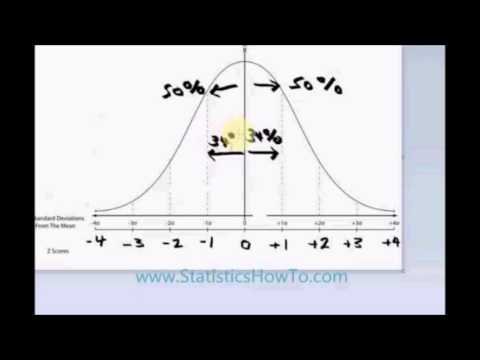
Increasing or decreasing sample sizes leads to changes in the variability of samples. For example, a sample size of 10 people taken from the same population of 1,000 will very likely give you a very different result than a sample size of 100.

There is no “perfect” sample size that will give you accurate estimates for the sample mean, variance and other statistics. Instead, you take your best “guess” — using standardized statistical procedures (see: [Finding the sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/)). In general, estimates will change from sample to sample and will probably never exactly match the population parameter.

## **Z-Score**

**Simply put, a z-score is the number of**[**standard deviations**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)**from the mean a data point is.** But more technically it’s a measure of how many standard deviations below or above the population [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) a [raw score](https://www.statisticshowto.datasciencecentral.com/raw-score/) is. A z-score is also known as a **standard score** and it can be placed on a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) μ and also the population standard deviation σ.

Z-scores are a way to compare results from a test to a “normal” population. Results from tests or surveys have thousands of possible results and units. However, those results can often seem meaningless. For example, knowing that someone’s weight is 150 pounds might be good information, but if you want to compare it to the “[average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)” person’s weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you **where**that person’s weight is **compared to the average population’s** mean weight.

[](https://www.youtube.com/watch?v=Gp4s5Ouk1gM)

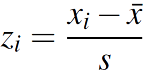
## **Z Score Formulas**

## **The Z Score Formula: One Sample**

The basic z score formula for a sample is:  
**z = (x – μ) / σ**

For example, let’s say you have a test score of 190. The test has a mean (μ) of 150 and a [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ) of 25. Assuming a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), your z score would be:  
z = (x – μ) / σ  
= 190 – 150 / 25 = 1.6.

The z score tells you how many standard deviations from the mean your score is. In this example, your score is 1.6 standard deviations above the mean.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/11/alternate-z-score.png)

You may also see the z score formula shown to the left. This is**exactly the same formula** as z = x – μ / σ, except that x̄ (the[sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)) is used instead of μ (the population mean) and s (the [sample standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/#HFSSD)) is used instead of σ (the population standard deviation). However, the steps for solving it are exactly the same.

## **Z Score Formula: Standard Error of the Mean**

When you have **multiple samples**and want to describe the standard deviation of those sample means ([the standard error](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/)), you would use this z score formula:

**z = (x – μ) / (σ / √n)**

This z-score will tell you how many standard errors there are between the sample mean and the population mean.

**Sample problem:** In general, the mean height of women is 65″ with a standard deviation of 3.5″. What is the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/)of finding a [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) of 50 women with a mean height of 70″, assuming the heights are normally distributed?

z = (x – μ) / (σ / √n)  
= (70 – 65) / (3.5/√50) = 5 / 0.495 = 10.1

The key here is that we’re dealing with a sampling distribution of means, so we know we have to include the standard error in the formula. We also know that 99% of values fall within 3 standard deviations from the mean in a normal probability distribution (see [68 95 99.7 rule](https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/)). Therefore, there’s less than 1% probability that any sample of women will have a mean height of 70″.

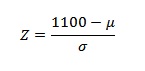
Confused about when to use σ and when to use σ √n? See: [Sigma / sqrt (n) — why is it used?](https://www.statisticshowto.datasciencecentral.com/sigma-sqrt-n-used/)

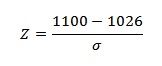
## **How to Calculate a Z-Score**

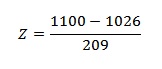
You can easily calculate a [z-score on a TI-83 calculator](https://www.statisticshowto.datasciencecentral.com/critical-z-value-ti-83/) or in [Excel](https://office.microsoft.com/en-us/excel/). However, if you don’t have either, you can calculate it by hand.

[Watch the video](https://www.youtube.com/watch?v=QgCXKzyqtMQ) or read below steps

**Sample question:**You take the SAT and score 1100. The mean score for the SAT is 1026 and the standard deviation is 209. How well did you score on the test compared to the average test taker?

Step 1: **Write your X-value into the z-score equation**. For this sample question the X-value is your SAT score, 1100.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/CALCULATE-A-Z-SCORE-1.jpg)

Step 2: **Put the mean, μ, into the z-score equation**.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/CALCULATE-A-Z-SCORE-2.jpg)

Step 3: **Write the standard deviation, σ into the z-score equation**.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/CALCULATE-A-Z-SCORE-3.jpg)

Step 4: **Calculate the answer using a calculator**:  
(1100 – 1026) / 209 = .354. This means that your score was .354 std devs above the mean.

Step 5: (**Optional**) Look up your z-value in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to see what percentage of test-takers scored below you. A z-score of .354 is .1368 + .5000\* = .6368 or 63.68%.

\*Why add .500 to the result? The z-table shown has scores for the RIGHT of the mean. Therefore, we have to add .500 for all of the area LEFT of the mean. For more examples of when to add (or subtract) .500, see several examples in: [Area under a normal distribution curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/find-the-area-under-a-normal-curve/).

## **Z scores and Standard Deviations**

Technically, a z-score is the number of [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the mean value of the reference population (a population whose known values have been recorded, like in [these charts](http://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf) the CDC compiles about people’s weights). For example:

* A z-score of 1 is 1 [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) above the mean.
* A score of 2 is 2 [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) above the mean.
* A score of -1.8 is -1.8 [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) below the mean.

A z-score tells you where the score lies on a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)curve. A z-score of **zero**tells you the values is **exactly average**while a score of +3 tells you that the value is much higher than average.

## **How is it Used in Real Life?**

You can use the z-table and the normal distribution graph to give you a visual about how a z-score of 2.0 means “higher than average”. Let’s say you have a person’s weight (240 pounds), and you know their z-score is 2.0. You know that 2.0 is above average (because of the high placement on the normal distribution curve), but you want to know how much above average is this weight?

The z-score in the center of the curve is zero. The z-scores to the **right of the mean are positive**and the z-scores to the **left of the mean are negative**. If you look up the score in the[z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/), you can tell what percentage of the population is above or below your score. The table below shows a z-score of 2.0 highlighted, showing .9772 (which converts to 97.72%). If you look at the same score (2.0) of the normal distribution curve above, you’ll see it corresponds with 97.72%.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/The_Normal_Distribution.jpg)

That tells you 97.72% of the population’s scores lie below that particular score and 100% – 97.72% = 2.28% of the scores lie above that score. A mere 2.28 of the population is above this person’s weight….probably a good indication they need to go on a diet!

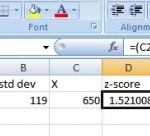
## **How to Find a Z-Score in Excel**

## **Z-Score in Excel: Overview**

A z-score tells you how many [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) a certain piece of data (“X”) is above or below the mean. A z-score in Excel can quickly be calculated using a basic formula. The formula for calculating a z-score is z=(x-μ)/σ, where μ is the [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/) and σ is the population standard deviation (note: if you don’t know the population standard deviation or the [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) is below 6, you should use a[t-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/) instead of a z-score). In other words, the formula is z = (data point – mean) / standard deviation.

## **Z-Score in Excel: Steps**

[Watch the video](https://www.youtube.com/watch?v=20K5YOUQwb4) or read the steps below:

**Sample question:** You take the [GRE](https://www.ets.org/gre)and scored 650 in the verbal section of the test. Calculate a z-score in Excel to determine how good your score is compared to the general population of test takers. The population mean (μ) is 469 and the population standard deviation (σ) is 119.  
  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/11/z-in-excel.png)

Step 1: **Type the population mean into a blank cell.** For this example, type “469” into cell A2. Optional: type the word “mean” as a column header in cell A1 so you remember what the value in cell A2 stands for.

Step 2: **Enter the population standard deviation into a blank cell**. For this example, type “119” into cell B2. Optional: type the words “standard deviation” as a column header in cell B1 so you remember what the value in cell B2 stands for.

Step 3: **Type the X-value (in this sample problem, X is your GRE score) into a blank cell.** For this example, type “650” into cell C2. Optional: type the words “X” as a column header in cell B1 so you remember what the value in cell B2 stands for.

Step 4: **Enter the following formula into an empty cell:**  
=(C2-A2)/B2

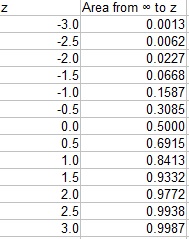
Step 5: **Press “Enter.”** The z-score will appear in cell D2: The z-score of 1.521008 in this sample problem indicates your GRE score was 1.521008.

## **Z score to Percentile Calculator and Manual Methods**

## **How do I Convert a Z Score to Percentile?**

Use a calculator, like the one below, or use a table (see instructions further down) and calculate the [percentile](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/)by hand. In either case, you will be more easily able to convert a z score to a percentiles if you know some basics about [normal distributions](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), like the [68 95 99.7 rule.](https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/) This rule states that 68 percent of the area under a [bell curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) lies between -1 and 1 [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) either side of the [mean](https://www.statisticshowto.datasciencecentral.com/mean/), 94 percent lies within -2 and 2 standard deviations and 99.7 percent lies within -3 and 3 standard deviations; these standard deviations are the “[z scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).”

## **Basic Z Table**

A [Z Table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/)has z scores and their associated areas. Once you have found the area, convert that to a [percentile](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/). This mini table shows the area for z scores in .5 increments:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/03/z-to-percentile.jpg)

For example, let’s say you wanted to convert a z score of -2 to a percentile. The area listed in the table is .0227. To convert this decimal to a percentile, move the decimal point two places to the right and then add a percentage sign:  
.0227 becomes 2.27%

Things become a little more tricky when you want a z score that’s not listed in the table above. In that case, you should use a more comprehensive z-table.

## **Z Score to Percentile Example**

**Z Score of 0.33**  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/03/z-score-to-percentile-3.jpg)

## **Z-table (Right of Curve or Left)**

## **What is a Z Table: Overview**

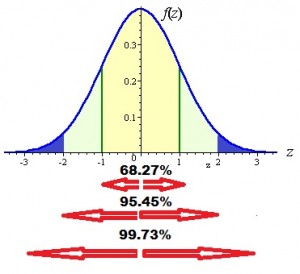
[Watch the video](https://youtu.be/hYGZqfs8078)

The **z-table**is short for the “Standard Normal z-table”. The [Standard Normal model](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) is used in [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/), including tests on proportions and on the difference between two means. The area under the whole of a normal distribution curve is 1, or 100 percent. The z-table helps by telling us what percentage is under the curve at any particular point.

## **What is a Z Table: Standard Normal Probability**

Every set of data has a different set of values. For example, heights of people might range from eighteen inches to eight feet and weights can range from one pound (for a preemie) to five hundred pounds or more. Those wide ranges make it difficult to analyze data, so we “standardize” the normal curve, setting it to have a [mean](https://www.statisticshowto.datasciencecentral.com/mean)of zero and a [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of one. When the curve is standardized, we can use a Z Table to find percentages under the curve.

## **Percentages under the curve**

[](https://www.statisticshowto.datasciencecentral.com/what-is-a-z-table-used-for/)

This graph shows the standardized normal graph with the percentage of results (data) that will fall between [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) on that graph. For example, 68.27 percent of results will fall within one standard deviation of the mean. On this graph, it’s represented by two [z-scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) from the z table: the area between z = -1 and z = 1.

## **The z-table**

Obviously a graph can only give us so much information. The above graph can tell us the area under the curve for one (z = -1 to 1), two (z = -2 to 2) and three (z = -3 to 3) standard deviations from the center. But what about if we want to know the area between z = -0.78 and z = 0.78? Or z = -1.2 and z=0.44? That’s where the z-table comes in. It tells us the area under the standard normal curve for any value between the mean (zero) and any z-score.

## **Why Are There at least Two z-tables?**

Simply, it’s to make life easier. Sometimes you’ll want to know the area between the mean and some positive value. That’s when you’ll use the right-hand z-table. But other times you might want to know the area in a left tail. If that’s the case, use the z-table that shows the area to the left of z.

## **Z-Table (right)**

[Watch the video](https://youtu.be/owJi8y3UWIE) to find out how to read a z-table, or scroll down for the table:

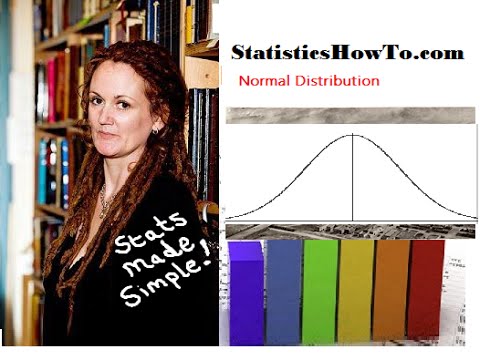
This z-table (normal distribution table) shows the area to the right hand side of the curve. Use these values to find the area between z=0 and any positive value. For an area in a **left tail**, look at [this left-tail z-table instead](https://www.statisticshowto.datasciencecentral.com/tables/z-table/#left).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/220px-Z_cumulative_from_mean.svg_.png)

## **Left Z-Table**

This table shows the area to the left of Z. In other words, the area of a left hand tail. If you want to find the value between z=0 and a positive number, use the right-hand z-table (above) instead (Hint: if you’re asked to look at the “z-table”, in most cases you’ll want to be looking at the other z-table!)

## **Normal Distributions (Bell Curve)**

[](https://www.youtube.com/watch?v=iMak-EW4HtM)

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/normal-distribution-probability.jpg)

*A normal distribution.*

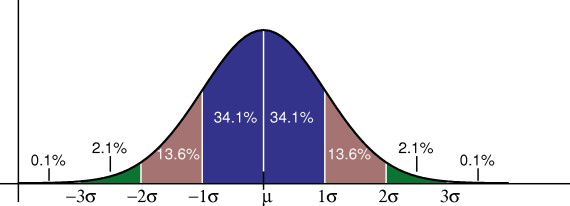
A [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), sometimes called the bell curve, is a distribution that occurs naturally in many situations. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean); half will fall to the right.

Many groups follow this type of pattern. That’s why it’s widely used in business, statistics and in government bodies like the [FDA](https://www.fda.gov/default.htm):

* Heights of people.
* Measurement errors.
* Blood pressure.
* Points on a test.
* IQ scores.
* Salaries.

The [empirical rule](https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/) tells you what percentage of your data falls within a certain number of [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the [mean](https://www.statisticshowto.datasciencecentral.com/mean):

• 68% of the data falls within one [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.datasciencecentral.com/mean).  
• 95% of the data falls within two [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.datasciencecentral.com/mean).  
• 99.7% of the data falls within three [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.datasciencecentral.com/mean).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/02/standard-normal-distribution.jpg)

The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the [mean](https://www.statisticshowto.datasciencecentral.com/mean); the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the [mean](https://www.statisticshowto.datasciencecentral.com/mean); the normal distribution will be flatter and wider.

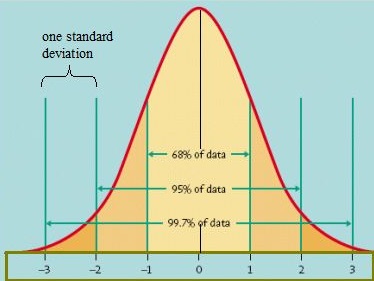
## **Properties of a normal distribution**

* The [mean, mode and median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/) are all equal.
* The curve is symmetric at the centre (i.e. around the mean, μ).
* Exactly half of the values are to the left of centre and exactly half the values are to the right.
* The total area under the curve is 1.

The Standard Normal Model - A standard normal model is a normal distribution with a mean of 1 and a standard deviation of 1.

## **Standard Normal Model: Distribution of Data**

One way of figuring out how data are distributed is to plot them in a graph. If the data is evenly distributed, you may come up with a **bell curve**. A bell curve has a small percentage of the points on both tails and the bigger percentage on the inner part of the curve. In the **standard normal model**, about 5 percent of your data would fall into the “tails” (colored darker orange in the image below) and 90 percent will be in between. For example, for test scores of students, the normal distribution would show 2.5 percent of students getting very low scores and 2.5 percent getting very high scores. The rest will be in the middle; not too high or too low. The shape of the standard normal distribution looks like this:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/standard-normal-distribution.jpg)

*Standard normal model. Image credit:*[*University of Virginia*](http://www.virginia.edu/)*.*

## **Practical Applications of the Standard Normal Model**

The standard normal distribution could help you figure out which subject you are getting good grades in and which subjects you have to exert more effort into due to low scoring percentages. Once you get a score in one subject that is higher than your score in another subject, you might think that you are better in the subject where you got the higher score. **This is not always true**.

You can only say that you are better in a particular subject if you get a score with a certain number of standard deviations above the mean. The standard deviation tells you how tightly your data is clustered around the mean; It allows you to compare different distributions that have different types of data — including different means.

For example, if you get a score of 90 in Math and 95 in English, you might think that you are better in English than in Math. However, in Math, your score is 2 standard deviations above the mean. In English, it’s only one standard deviation above the mean. It tells you that in Math, your score is far higher than most of the students (your score falls into the tail).  
Based on this data, you actually performed better in Math than in English!

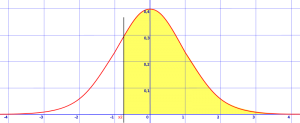
## **Probability Questions using the Standard Model**

Questions about standard normal distribution probability can look alarming but the key to solving them is understanding what the area under a standard normal curve represents. The total area under a standard normal distribution curve is 100% (that’s “1” as a decimal). For example, the left half of the curve is 50%, or .5. So the probability of a random variable appearing in the left half of the curve is .5.

Of course, not all problems are quite that simple, which is why there’s a [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/). All a z-table does is measure those probabilities (i.e. 50%) and put them in [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean). The mean is in the centre of the standard normal distribution, and a probability of 50% equals zero standard deviations.

## **Standard normal distribution: How to Find Probability (Steps)**

**Step 1:** Draw a bell curve and shade in the area that is asked for in the question. The example below shows z >-0.8. That means you are looking for the probability that z is greater than -0.8, so you need to draw a vertical line at -0.8 standard deviations from the mean and shade everything that’s greater than that number.



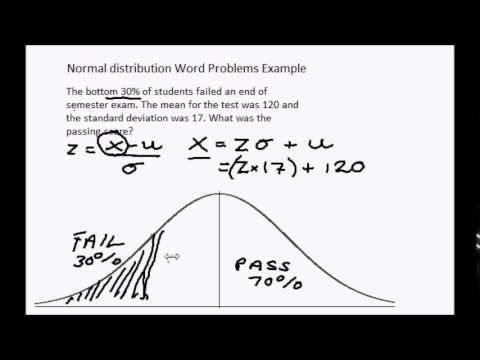
*shaded area is z > -0.8*

**Step 2:** Visit the[normal probability area index](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/find-the-area-under-a-normal-curve/)and find a picture that looks like your graph. Follow the instructions on that page to find the z-value for the graph. The z-value is the probability.

**Tip:**Step 1 is technically optional, but it’s always a good idea to sketch a graph when you’re trying to answer probability word problems. That’s because most mistakes happen not because you can’t do the math or read a z-table, but because you subtract a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) instead of adding (i.e. you imagine the probability under the curve in the wrong direction. A sketch helps you cement in your head exactly what you are looking for.

## **Normal Distribution Word Problems**

This video shows one example of a normal distribution word problem. For more examples, read on below:

[](https://www.youtube.com/watch?v=4CEqKDqKq34)

When you tackle normal distribution in a statistics class, you’re trying to find the area under the curve. The total area is 100% (as a decimal, that’s 1). **Normal distribution problems** come in **six**basic types. How do you know that a word problem involves normal distribution? Look for the key phrase “assume the [*variable*](https://www.statisticshowto.datasciencecentral.com/variable/) is normally distributed” or “assume the variable is approximately normal.” To solve a word problem you need to figure out which type you have.

1. “Between”: Contain the phrase “between” and includes an upper and lower limit (i.e. “find the number of houses priced between $50K and 200K”).
2. “More Than” or “Above”: contain the phrase “more than” or “above”.
3. “Less Than”.
4. Lower Cut Off Example (video)
5. Upper Cut Off Example (video)
6. Middle Percent Example (video)

## **1. “Between”**

This how-to covers solving problems that contain the phrase “between” and includes an upper and lower limit (i.e. “find the number of houses priced between $50K and 200K”. Note that this is different from finding the “middle percentage” of something.

## **Word problems with normal distribution: “Between”: Steps**

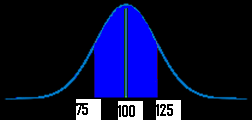
**Step 1:** Identify the parts of the word problem. The word problem will identify:

1. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)(average or μ).
2. [Standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ).
3. Number selected (i.e. “choose one at random” or “select ten at random”).
4. X: the numbers associated with “between” (i.e. “between $5,000 and $10,000” would have X as 5,000 and as $10,000).

In addition, you will be given EITHER:

* 1. [Sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) (i.e. 400 houses, 33 people, 99 factories, 378 plumbers etc.). OR
  2. You might be asked for a [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/)(in which case your sample size will most likely be everyone, i.e. “Journeyman plumbers” or “First year pilots.”

**Step 2:** Draw a graph. Put the mean you identified in Step 1 in the center. Put the number associated with “between” on the graph (take a guess at where the numbers would fall–it doesn’t have to be exact). For example, if your mean was $100, and you were asked for “hourly wages between $75 and $125”) your graph will look something like this:



**Step 3:**Figure out the [*z-scores*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/). Plug the first X value (in my graph above, it’s 75) into the z value formula and solve. The μ (the mean), is 100 from the sample graph. You can get these figures (including σ, the standard deviation) from your answers in step 1 :

z score formula

\*Note: if the formula confuses you, all this formula is asking you to do is:

1. subtract the mean from X
2. divide by the standard deviation.

**Step 4:** Repeat step 3 for the second X.

**Step 5:** Take the numbers from step 3 and 4 and use them to find the area in the [*z-table*](https://www.statisticshowto.datasciencecentral.com/tables/z-table/).

If you were asked to find a probability in your question, go to step 6a. If you were asked to find a number from a specific given sample size, go to step 6b.

**Step 6a:**

Convert the answer from step 5 into a percentage.

For example, 0.1293 is 12.93%.

That’s it–skip step 6b!

**Step 6b**

Multiply the sample size (found in step 1) by the z-value you found in step 4. For example, 0.300 \* 100 = 30.

That’s it!

## **2. “More Than” or “Above”**

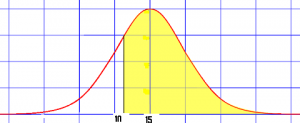
This how-to covers solving normal distribution problems that contain the phrase “**more than**” (or a phrase like “above”).

**Step 1:** Break up the word problem into parts. Find:

1. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) ([average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)or μ)
2. [Standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ)
3. A number (for example, “choose fifty at random” or “select 90 at random”)
4. X: the number associated with the “less than” statement. For example, if you were asked to find “under $9,999” then X is 9,999.

**Step 2:** Find the [sample](https://www.statisticshowto.datasciencecentral.com/sample/)from the problem. You’ll have either a specific size (like “1000 televisions”) or a general sample (“Every television”).

Draw a picture if the problem with the mean and the area you are looking for. For example, if the mean is $15, and you were asked to find what dinners cost more than $10, your graph might look like this:



**Step 3:** Calculate the [*z-score*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) (plug your values into the z value formula and solve). Use your answers from step 1 :

z score formula

Basically, all you are doing with the formula is subtracting the mean from X and then dividing that answer by the standard deviation.

**Step 4:** Find the area using the z-score from step 3. Use the [z-table](https://www.statisticshowto.datasciencecentral.com/z-table/). Not sure how to read a z-table? See the video on the [z-table page.](https://www.statisticshowto.datasciencecentral.com/tables/z-table/)

**Step 6:** Go to Step 6a to find a probability OR go to step 6b to calculate a certain number or amount.

**Step 6a**  
Turn step 5’s answer into a percentage.

For example, 0.1293 is 12.93%.

Skip step 6b: you’re done!

**Step 6b**  
Multiply the sample size from Step 1 by the z-score from step 4. For example, 0.500 \* 100 = 50.

You’re done!

## **3. Less Than**

This how-to covers solving **normal distribution word problems** that have the phrase “**less than**” (or a similar phrase such as “fewer than”).

## **Normal distribution word problems less than: Steps**

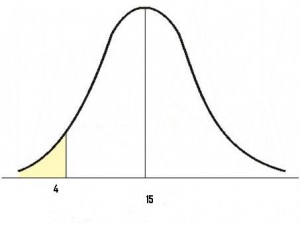
**Step 1:** Break up the word problem into parts:

1. The [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)([average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)or μ)
2. [Standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ)
3. Number selected (i.e. “choose one at random” or “select ten at random”)
4. X: the number that goes with “less than” (i.e. “under $99,000” would list X as 99,000)

Plus, you will have EITHER:

1. A specific [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/). For example, 500 boats, 250 sandwiches, 100 televisions etc.
2. Everyone in the [sample](https://www.statisticshowto.datasciencecentral.com/sample/)(you’ll be asked to find a probability). For example “first year medical students,” “Cancer patients” or “Airline pilots.”

**Step 2:** Draw a picture to help you visualize the problem. The following graph shows a mean of 15, and an area “under 4”):



**Step 3:**  Find the z value by plugging the given values into the formula. The “X” in our sample graph is 4, and the μ (or mean) is 15. You can get these figures (including σ, the standard deviation) from your answers in step 1, where you identified the parts of the problem:

z score formula

All you have to do to solve the formula is:

1. Subtract the mean from X.
2. Divide by the standard deviation.

**Step 4:** Take the number from step 3, then use the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to find the area

**Step 5:**To [*find a probability*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/), go to step 6a. To find a number from a specific given sample size, go to step 6b.

**Step 6a**

Change the number from step 5 into percentage.

For example, 0.1293 is 12.93%.

That’s it!

**Step 6b**

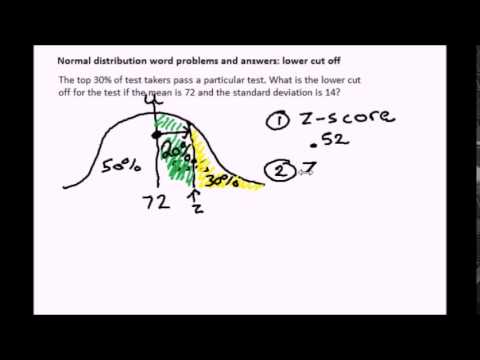
Multiply the sample size (found in step 1) by the z-value you found in step 4. For example, 0.300 \* 100 = 30.

That’s it!

## **4. Lower Cut Off**

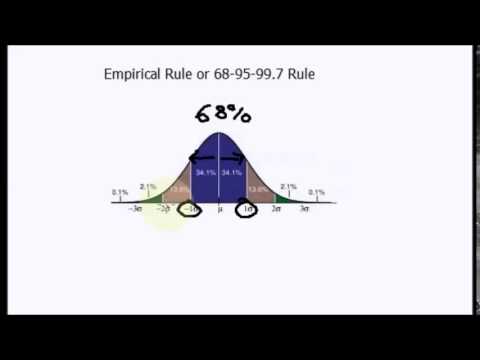
Sometimes on **a normal distribution word problem** you’ll be asked to find a “lower limit of an upper percentage” of something (i.e. “find the cut-off point to pass a certain exam where the upper 40% of test takers pass”). A lower cut off point is the point where scores will fall below that point. For example, you might want to find where the cut off point is for the bottom 10% of test takers.

X is the cut off point in z score formula

[](https://www.youtube.com/watch?v=vH0hia948ws)

## **Empirical Rule or 68-95-97.7 Rule**

Watch the video or read the article below:

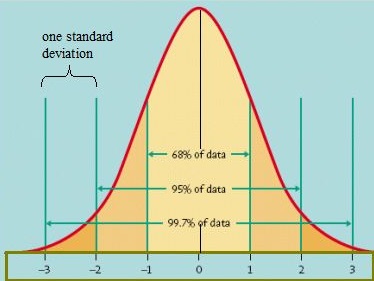
[](https://www.youtube.com/watch?v=hQTvdD8vtio)

## **Definition of the Empirical Rule**

The empirical rule states that for a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), nearly all of the data will fall within three [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.datasciencecentral.com/mean/). The empirical rule can be broken down into three parts:

* 68% of data falls within the first standard deviation from the mean.
* 95% fall within two standard deviations.
* 99.7% fall within three standard deviations.

The rule is also called the [68-95-99 7 Rule](https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/) or the **Three Sigma Rule**.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/standard-normal-distribution.jpg)

*Standard normal distribution showing standard deviations. Image credit: University of Virginia.*

## **When do we use the Empirical Rule?**

The Empirical Rule is often used in statistics for [**forecasting**](http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm), especially when obtaining the right data is difficult or impossible to get. The rule can give you a rough estimate of what your data collection might look like if you were able to survey the entire [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/).

This rule applies generally to a [random variable](https://www.statisticshowto.datasciencecentral.com/random-variable/), X, following the shape of a normal distribution, or [bell-curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/), with a mean “mu” (the Greek letter &mu) and a standard deviation “sigma” (the Greek letter σ). The rule doesn’t apply to distributions that are not normal, but you can apply it to other distributions using [Chebyshev’s Theorem](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/chebyshevs-theorem-inequality/).

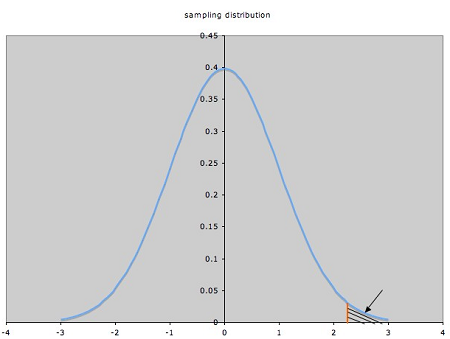
## **Empirical Rule: Notation**

When applying the Empirical Rule to a data set the following conditions are true:

* Approximately 68% of the data falls within one standard deviation of the mean (or between the mean – one times the standard deviation, and the mean + 1 times the standard deviation). In mathematical notation, this is represented as: μ ± 1σ
* Approximately 95% of the data falls within two standard deviations of the mean (or between the mean – 2 times the standard deviation, and the mean + 2 times the standard deviation). The mathematical notation for this is: μ ± 2σ
* Approximately 99.7% of the data falls within three standard deviations of the mean (or between the mean – three times the standard deviation and the mean + three times the standard deviation). The following notation is used to represent this fact: μ ± 3σ

## **Critical Value**

A critical value is a line on a graph that splits the graph into sections. One or two of the sections is the “rejection region”; if your test value falls into that region, then you reject the null hypothesis.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/11/critical-values.png)

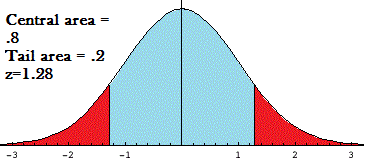
*A one tailed test with the rejection rejection in one tail. The critical value is the red line to the left of that region.*

[](https://www.youtube.com/watch?v=yk3soTy7Bko)

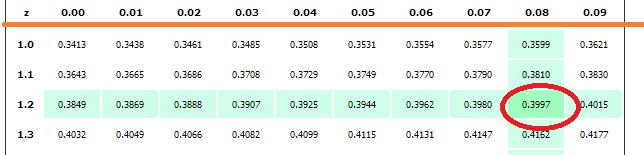
Critical values come in all shapes and sizes, but the one you’ll come across first in statistics is the critical value of Z.

## **Critical Value of Z**

The critical value of z is term linked to the area under the [standard normal model](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/). Critical values can tell you what probability any particular variable will have

.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/11/critical-value-of-z.png)  
  
The above graph of the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/)curve shows a critical value of 1.28. The graph has two parts:

* **Central region:** The [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) is equal to the number of [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean). A score of 1.28 indicates that the [variable](https://www.statisticshowto.datasciencecentral.com/variable/)is 1.28 [standard deviations](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) from the mean. If you look in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/)for a z of 1.28, you’ll find the area is .3997. This is the region to the right of the mean, so you’ll double it to get the area of the entire central region: .3997\*2 = .7994 or about 80 percent.

[](https://www.statisticshowto.datasciencecentral.com/what-is-the-critical-value-of-z/)

* **Tail region**: The area of the tails (the red areas) is 1 minus the central region. In this example, 1-.8=.20, or about 20 percent. The tail regions are sometimes calculated when you want to know how many [variables](https://www.statisticshowto.datasciencecentral.com/variable/) would be less than or more than a certain figure.

A critical value of z is sometimes written as za, where the [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/), a, is the area in the tail. For example, z.10=1.28.

**When are Critical values of z used?**

A critical value of z ([Z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)) is used when the sampling distribution is normal, or close to normal. Z-scores are used when the population standard deviation is known or when you have larger [sample sizes](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/). While the z-score can also be used to calculate probability for unknown standard deviations and small [samples](https://www.statisticshowto.datasciencecentral.com/sample/), many statisticians prefer to use the [t distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/) to calculate these probabilities.

See also: [T Critical Value.](https://www.statisticshowto.datasciencecentral.com/t-critical-value/)

**Other uses of z-scores**

Every [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/)has a probability, and every probability calculated for a sample has a[margin of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/#WhatMofE). The critical value of z can also be used to calculate the margin of error.  
Margin of error = **Critical value** \* Standard deviation of the statistic  
Margin of error = **Critical value** \* [Standard error of the sample](https://www.statisticshowto.datasciencecentral.com/what-is-the-standard-error-of-a-sample/)

## **Find a Critical Value in Any Tail**

How you look up a critical value is very straightforward as long as you know if you have a [left tailed test or right tailed test](https://www.statisticshowto.datasciencecentral.com/how-to-decide-if-a-hypothesis-test-is-a-left-tailed-test-or-a-right-tailed-test/) (or potentially, both).

## **A. Find a critical value for a confidence level**

[](https://www.youtube.com/watch?v=RAnFyF_6zHk)

**Example question:** Find a critical value for a 90% confidence level (Two-Tailed Test).

Step 1: Subtract the confidence level from 100% to find the α level: 100% – 90% = 10%.

Step 2: Convert Step 1 to a decimal: 10% = 0.10.

Step 3: Divide Step 2 by 2 (this is called “α/2”).  
0.10 = 0.05. This is the area in each tail.

Step 4: Subtract Step 3 from 1 (because we want the area in the middle, not the area in the tail):  
1 – 0.05 = .95.

Step 5: Look up the area from Step in the z-table. The area is at z=1.645. This is your critical value for a confidence level of 90%.

## **B. Common confidence levels and their critical values**

You don’t have to perform the above calculations every time. This list of critical values and their associated two-tailed test confidence levels were calculated using the above steps:

|  |  |
| --- | --- |
| **Confidence Level** | **Critical Value (**[**Z-score**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)**)** |
| 0.90 | 1.645 |
| 0.91 | 1.70 |
| 0.92 | 1.75 |
| 0.93 | 1.81 |
| 0.94 | 1.88 |
| 0.95 | 1.96 |
| 0.96 | 2.05 |
| 0.97 | 2.17 |
| 0.98 | 2.33 |
| 0.99 | 2.575 |

## **C. Find a Critical Value: Two-Tailed Test**

[Watch the video](https://www.youtube.com/watch?v=cTQCopdKnjY) or read the steps below:

**Sample question**: Find the critical value for alpha of .05.

**Step 1:** Subtract alpha from 1.

1 – .05 = **.95**

**Step 2:**Divide Step 1 by 2 (because we are looking for a two-tailed test). 95 / 2 = **.475**

**Step 3:** Look at your [*z-table*](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) and locate the answer from Step 2 in the middle section of the z-table. The fastest way to do this is to use the [find function](http://www.altairhyperworks.com/hwhelp/Altair/hw12.0/help/hw/hw.htm?using_your_browsers_find_funct.htm) of your browser (usually CTRL+F). In this example we’re going to look for .475, so go ahead and press CTRL+F, then type in .475.

**Step 4:**In this example, you should have found the number .4750. Look to the far left or the row, you’ll see the number 1.9 and look to the top of the column, you’ll see .06. Add them together to get **1.96**. That’s the critical value!

**Tip**: The critical value **appears twice in the z table** because you’re looking for both a left hand and a right hand tail, so don’t forget to add the plus or minus sign! In our example you’d get **±1.96**.

**D. Find a Critical Value: Right-Tailed Test**

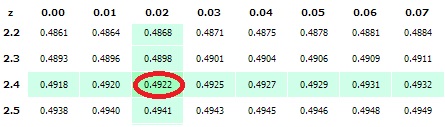
[find a critical value rt](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2011/12/find-a-critical-value-rt.png)

**Sample question**: Find a critical value in the z-table for an alpha level of 0.0079.

**Step 1:** Draw a diagram, like the one above. Shade in the area in the right tail.  
This area represents alpha, α. A diagram helps you to visualize what area you are looking for (i.e. if you want an area to the right of the mean or the left of the mean).

**Step 2:** Subtract alpha (α) from 0.5.  
0.5-0.0079 = 0.4921.

**Step 3:** Find the result from step 2 in the center part of the [*z-table*](https://www.statisticshowto.datasciencecentral.com/tables/z-table/).  
The closest area to 0.4921 is 0.4922 at z=2.42.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/critical-value-z-table.jpg)

*The closest to .4921 is the value (.4922) at z = 2.42*

That’s it!

## **E. Find a Critical Value: Left-Tailed Test**

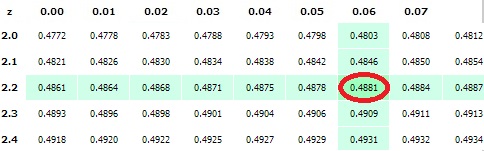
**Sample question**: find the critical value in the z-table for α=.012 (left-tailed test).

[left](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2011/12/left.png)

**Step 1:** Draw a diagram, like the one above. Shade in the area in the left tail (because you’re looking for a critical value for a ***left-tailed*** test).  
This area represents alpha, α.

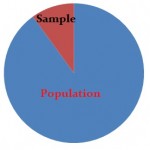
**Step 2:**Subtract alpha (α) from 0.5.  
0.5 – 0.012 = 0.488.

**Step 3:** Find the result from step 2 in the centre part of the[*z-table.*](https://www.statisticshowto.datasciencecentral.com/tables/z-table/)

The closest area to 0.488 is at z=2.26. If you can’t find the exact area, just find the closest number and read the z value for that number.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/critical-value-z-table-2.jpg)

**Step 4:**Add a negative sign to Step 3 (left-tail critical values are always negative).  
-2.26.

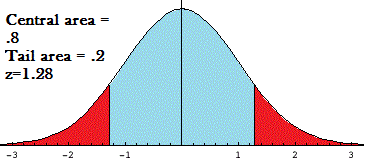
## **Critical Values and Working with Samples**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/10-percent-condition.jpg)

Critical values are used in statistics for [**hypothesis testing**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/). When you work with statistics, you’re working with a small percentage (a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)) of a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, you might have statistics for voting habits from two percent of democratic voters, or five percent of students and their test results. Because you’re working with a fraction of the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)and not the entire population, you can never be one hundred percent certain that your results reflect the actual population’s results. You might be 90 percent certain, or even 99 percent certain, but you can never be 100 percent certain. How accurate are your results? You can tell with [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/).

## **Types of Critical Values**

Various types of critical values are used to calculate [significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/), including: [t scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/) from [student’s t-tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/), [chi-square](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/), and z-tests. In each of these tests, you’ll have an area where you are able to [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/), and an area where you cannot. The line that separates these two regions is where your critical values are.

[](https://www.statisticshowto.datasciencecentral.com/what-is-the-critical-value-of-z/)

In the above image, the critical values are at 1.28 or -1.28. The blue area is where you must accept the [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state). The red areas are where you can [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). How large these areas actually are (and what test you use) is dependent on many factors, including your chosen [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/) and your [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/).

[Significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/) testing is used to figure out if your results differ from the [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state). The null hypothesis is just an accepted fact about the population.

## [**Significance**](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/)**Testing Example**

For example, your school may make a statistics course mandatory for nursing students because research has shown that patient outcomes improve when nurses have a statistics background. You might think that there’s no difference. Instead of trying to prove that there’s no difference, proper research techniques dictate that you’ll try to disprove the opposite — the null hypothesis, which in this case is that “patient outcomes improve when nurses have a statistics background.” In order to disprove, or reject, the null hypothesis, your research must pass a test of [significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/).

## **What does Significance Testing Tell Us?**

[Significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/)testing is used to calculate the probability that a relationship between two variables (like “taking a statistics class” and “improved patient outcomes”) is just due to chance. It helps to answer the question of whether you could duplicate your test results accurately in further research. By using probability and the normal curve, you can figure out what the chance is that your research is wrong.

Steps in Testing for Statistical Significance

1. State the [Alternate Hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/).
2. State the [Null Hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state).
3. Select a probability of error level ([alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/)).
4. Select and compute the test for [statistical significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/) (i.e. calculate a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).)
5. Interpret the results.

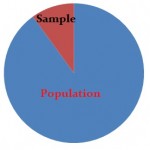
## **More Critical Values Articles**

1. [Chi square test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/).
2. [How to Find a Critical Chi-Square Value](https://www.statisticshowto.datasciencecentral.com/how-to-find-a-critical-chi-square-value/)
3. [How do I Find the Area Under a Normal Distribution Curve?](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/find-the-area-under-a-normal-curve/)
4. [Hypothesis Testing Examples](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/)
5. [One tailed Distribution: How to Find the Area.](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/#area)
6. [What is z alpha/2?](https://www.statisticshowto.datasciencecentral.com/z-alpha2-za2/)
7. [What is a Z Test?](https://www.statisticshowto.datasciencecentral.com/z-test/)
8. [One Sample Z Test](https://www.statisticshowto.datasciencecentral.com/one-sample-z-test/)
9. [Z Score to Percentile Calculator and Manual methods](https://www.statisticshowto.datasciencecentral.com/percentile-z-score/)

## **Statistical Significance**

Statistics isn’t an exact science. In fact, you can think of stats as very finely tuned guesswork. As stats is guesswork, we need to know how close our “guess” is. That’s where **statistical significance** comes in.

## **What is Statistical Significance: Overview**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/10-percent-condition.jpg)  
Stats is all about taking a piece of the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)and making a guess about what that population’s behavior might be like. If you were working with [parameters](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)([parameter vs. statistic explanation](https://www.statisticshowto.datasciencecentral.com/how-to-tell-the-difference-between-a-statistic-and-a-parameter/)), there would be no need for guesswork; You’d have all the data. In real life getting all of the data can be **costly, time-consuming**, or **impossible**.

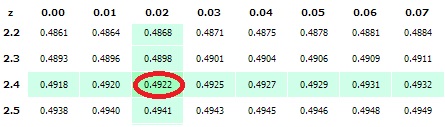
For example, [Gallup Polls](http://www.gallup.com/home.aspx) uses stats to estimate who will win the next election. Drug manufacturers use stats to estimate how many people might have side effect from their drugs. And businesses use stats to forecast sales figures for the future.

## **What is Statistical Significance a Measure of?**

Statistical significance is a measure of **whether your research findings are meaningful**. More specifically, it’s whether your stat closely matches what value you would expect to find in an entire [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). In order to test for statistical significance, perform these steps:

1. Decide on an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/). An alpha level is the error rate you are willing to work with (usually 5% or less).
2. Conduct your research. For example, conduct a poll or collect data from an experiment.
3. Calculate your [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/). A statistic is just a piece of information about your sample, like a [mean](https://www.statisticshowto.datasciencecentral.com/mean/), [mode](https://www.statisticshowto.datasciencecentral.com/mode/)or [median](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median).
4. Compare the statistic you calculated in Step 3 with a statistic from a [statistical table](https://www.statisticshowto.datasciencecentral.com/tables/).

There are many different types of tables you can use to figure out if your data is significant or not. For more on [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/), see our [hypothesis Testing index](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/critical-value-z-table.jpg)

# Alpha Level (Significance Level)

Watch the video or read on below:

[](https://www.youtube.com/watch?v=dDOKxRnCovo)

The significance level α is the probability of making the wrong decision when the [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) is true. Alpha levels (sometimes just called “significance levels”) are used in [hypothesis tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/). Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.

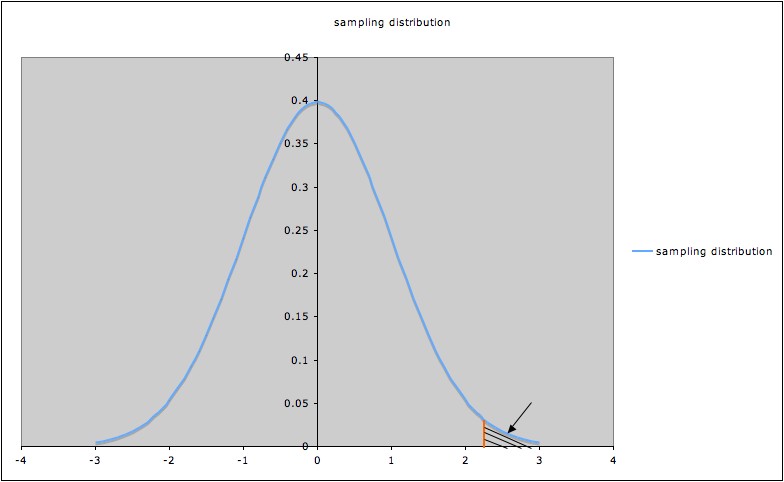
1. [Type I and II errors](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/#alphatype)
2. [How do I Calculate an Alpha Level for one- and two-tailed tests?](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/#alphacalc)
3. [Why is an Alpha Level of .05 commonly used?](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/#alphawhy)

## **1. Alpha Levels / Significance Levels: Type I and Type II errors**

In hypothesis tests, two errors are possible, [Type I and Type II errors](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/type-i-error-type-ii-error-decision/).  
**Type I error**: Supporting the [alternate hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/) when the null hypothesis is true.  
**Type II error**: Not supporting the alternate hypothesis when the alternate hypothesis is true.

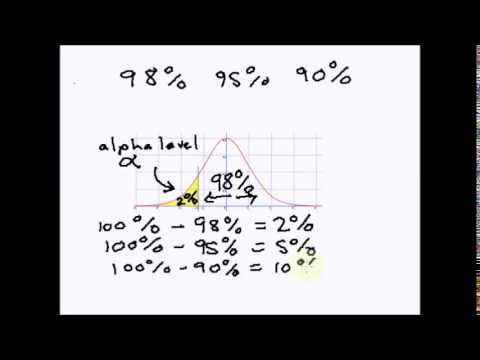
In an example of a courtroom, let’s say that the null hypothesis is that a man is innocent and the alternate hypothesis is that he is guilty. if you convict an innocent man (Type I error), you support the alternate hypothesis (that he is guilty). A type II error would be letting a guilty man go free.

An **alpha level** is the probability of a type I error, or you [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/) when it is true. A related term, [beta](https://www.statisticshowto.datasciencecentral.com/beta-level/), is the opposite; the probability of rejecting the alternate hypothesis when it is true.

[](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level)

This graph shows the [rejection region](https://www.statisticshowto.datasciencecentral.com/rejection-region/) to the far right.

## **2. How do I Calculate an Alpha Level for one- and two-tailed tests?**

[](https://www.youtube.com/watch?v=AkC4X5_nP2U)

Alpha levels can be controlled by you and are related to **confidence levels**. To get α subtract your [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/) from 1. For example, if you want to be 95 percent confident that your analysis is correct, the alpha level would be 1 – .95 = 5 percent, assuming you had a one tailed test. For two-tailed tests, divide the alpha level by 2. In this example, the two tailed alpha would be .05/2 = 2.5 percent. See: [One-tailed test or two?](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/) for the difference between a one-tailed test and a two-tailed test.

## **3. Why is an alpha level of .05 commonly used?**

Seeing as the alpha level is the probability of making a Type I error, it seems to make sense that we make this area as tiny as possible. For example, if we set the alpha level at 10% then there is large chance that we might incorrectly reject the null hypothesis, while an alpha level of 1% would make the area tiny. So why not use a tiny area instead of the standard 5%?

The smaller the alpha level, the smaller the area where you would reject the null hypothesis. So if you have a tiny area, there’s more of a chance that you will NOT reject the null, when in fact you should. This is a Type II error.

In other words, the more you try and avoid a Type I error, the more likely a Type II error could creep in. Scientists have found that an alpha level of 5% is a good balance between these two issues.

## **Null Hypothesis**

[Watch the video](https://youtu.be/tDmCFVQvv2A) or read below steps

The null hypothesis, H0 is the commonly accepted fact; it is the opposite of the [alternate hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/). Researchers work to reject, nullify or disprove the null hypothesis. Researchers come up with an alternate hypothesis, one that they think explains a phenomenon, and then work to [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/).

## **Why is it Called the “Null”?**

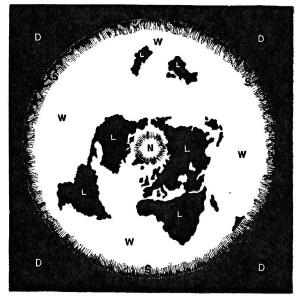
The word “null” in this context means that it’s a commonly accepted fact that researchers work to nullify. It doesn’t mean that the statement is null itself! (Perhaps the term should be called the “nullifiable hypothesis” as that might cause less confusion).

## **Why Do I need to Test it? Why not just prove an alternate one?**

The short answer is, as a scientist, you are required to; It’s part of the scientific process. Science uses a battery of processes to prove or disprove theories, making sure than any new hypothesis has no flaws. Including both a null and an alternate hypothesis is one safeguard to ensure your research isn’t flawed. Not including the null hypothesis in your research is considered very bad practice by the scientific community. If you set out to prove an alternate hypothesis without considering it, you are likely setting yourself up for failure. At a minimum, your experiment will likely not be taken seriously.

## **Example**

Not so long ago, people believed that the world was flat.

[](https://www.statisticshowto.datasciencecentral.com/what-is-the-null-hypothesis/)

Null hypothesis, H0: The world is flat.  
Alternate hypothesis: The world is round.  
Several scientists, including [Copernicus](http://plato.stanford.edu/entries/copernicus/), set out to disprove the null hypothesis. This eventually led to the rejection of the null and the acceptance of the alternate. Most people accepted it — the ones that didn’t created the [Flat Earth Society](http://www.theflatearthsociety.org/)!. What would have happened if Copernicus had not disproved the it and merely proved the alternate? No one would have listened to him. In order to change people’s thinking, he first had to prove that their thinking was wrong.

## **How to State the Null Hypothesis**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=wpJ15yOAeYM)

## **How to State the Null Hypothesis from a Word Problem**

You’ll be asked to convert a word problem into a**hypothesis**statement in statistics that will include a null hypothesis and an [**alternate hypothesis**](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/). Breaking your problem into a few small steps makes these problems much easier to handle.

## **How to State the Null Hypothesis**

**Example Problem**: A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. [Average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)recovery times for knee surgery patients is 8.2 weeks.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/knee-surgery.jpg)

*Hypothesis testing is vital to test patient outcomes.*

**Step 1:** Figure out the hypothesis from the problem. The hypothesis is usually hidden in a word problem, and is sometimes a statement of what you expect to happen in the experiment. The hypothesis in the above question is “I expect the average recovery period to be greater than 8.2 weeks.”

**Step 2:** Convert the hypothesis to math. Remember that the average is sometimes written as μ.

H1: μ > 8.2

Broken down into (somewhat) English, that’s H1(The hypothesis): μ (the average) > (is greater than) 8.2

**Step 3:** State what will happen if the hypothesis doesn’t come true. If the recovery time isn’t greater than 8.2 weeks, there are only two possibilities, that the recovery time is equal to 8.2 weeks or less than 8.2 weeks.

H0: μ ≤ 8.2

Broken down again into English, that’s H0(The null hypothesis): μ (the average) ≤ (is less than or equal to) 8.2

## **How to State the Null Hypothesis: Part Two**

## **But what if the researcher doesn’t have any idea what will happen?**

**Sample Problem:** A researcher is studying the effects of radical exercise program on knee surgery patients. There is a good chance the therapy will improve recovery time, but there’s also the possibility it will make it worse. Average recovery times for knee surgery patients is 8.2 weeks.

**Step 1:** State what will happen if the experiment doesn’t make any difference. That’s the null hypothesis–that nothing will happen. In this experiment, if nothing happens, then the recovery time will stay at 8.2 weeks.

H0: μ = 8.2

Broken down into English, that’s H0(The null hypothesis): μ (the average) = (is equal to) 8.2

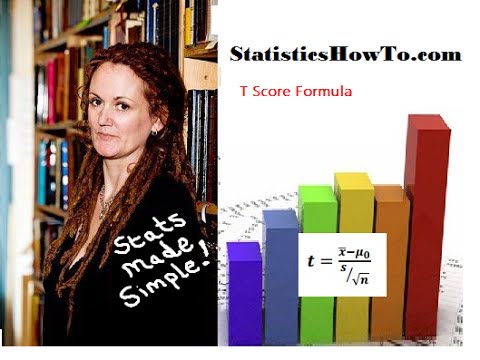
**Step 2:** Figure out the [*alternate hypothesis*](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/). The alternate hypothesis is the opposite of the null hypothesis. In other words, what happens if our experiment makes a difference?

H1: μ ≠ 8.2

In English again, that’s H1(The  alternate hypothesis): μ (the average) ≠ (is not equal to) 8.2

## **T Scores in Statistics**

Watch the video or read on below:

[](https://www.youtube.com/watch?v=GvKDJGqtqrM)

## **What is the T Score Formula?**

A t score is one form of a [standardized test statistic](https://www.statisticshowto.datasciencecentral.com/standardized-test-statistic/) (the other you’ll come across in [elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/) is the [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/)). The t score formula enables you to take an individual score and transform it into a standardized form one which helps you to compare scores.

You’ll want to use the t score formula when you don’t know the population [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) and you have a small [sample](https://www.statisticshowto.datasciencecentral.com/sample/)(under 30).

The t score formula is:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/11/t-score.jpg)  
Where  
x̄ = [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/)  
μ0 = [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/)  
s = sample [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)  
n =[sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/)

If you have only one item in your [sample](https://www.statisticshowto.datasciencecentral.com/sample/), the square root in the denominator becomes √1. This means the formula becomes:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/11/t-score-formula-2.jpg)

In simple terms, the larger the t score, the larger the difference is between the groups you are testing. It’s influenced by many factors including:

* How many items are in your [sample](https://www.statisticshowto.datasciencecentral.com/sample/).
* The [means](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)of your sample.
* The mean of the population from which your sample is drawn.

The standard deviation of your sample.

## **What is the T Score Formula used for?**

You traditionally look up a t score in a [t-table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/). The number of items in your sample, minus one, is your [degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/). For example, if you have 20 items in your sample, then df = 19. You use the degrees of freedom along with the [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/) you are willing to accept, to decide whether to [support or reject the null hypothesis.](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/)

The t score formula can also be used to solve [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/)questions. You won’t have an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/), but you can use the result from the formula, along with a calculator like the TI-83, to find probabilities.

The following example shows how to calculate a t-score formula for a single sample. Paired samples and independent samples use different formulas.

* If you have paired samples, follow the instructions in the [paired samples t-test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/#PairedTTest).
* For independent samples, see: [independent samples t-test](https://www.statisticshowto.datasciencecentral.com/independent-samples-t-test/).

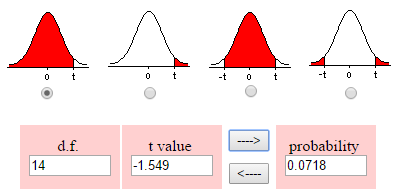
## **Example of the T Score Formula**

**Sample question:**  
A law school claims it’s graduates earn an average of $300 per hour. A sample of 15 graduates is selected and found to have a mean salary of $280 with a [sample standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/#HFSSD)of $50. Assuming the school’s claim is true, what is the probability that the [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)salary of graduates will be no more than $280?

Step 1: Plug the information into the formula and solve:  
x̄ = [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/) = 280  
μ0 = [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/) = 300  
s = sample standard deviation = 50  
n = [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) = 15

t = (280 – 300)/ (50/√15) = -20 / 12.909945 = -1.549.

Step 2: Subtract 1 from the sample size to get the [degrees of freedom:](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/)  
15 – 1 = 14. The degrees of freedom lets you know which form of the [t distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/probability-and-statistics/t-distribution/)to use (there are many, but you can solve these problems without knowing that fact!).

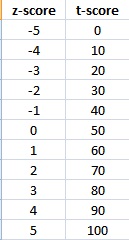
Step 3: Use a calculator to find the probability using your [degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) (8). You have several options, including the TI-83 (see [How to find a t distribution on a TI 83](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/#TI83)) and [this online calculator](https://surfstat.anu.edu.au/surfstat-home/tables/t.php). Here’s the result from that calculator. Note that I selected the radio button under the left tail, as we’re looking for a result that’s no more than $280:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/11/t-dist-formula.png)  
  
  
The probability is 0.0718, or 7.18%.

## **T Scores in Psychometrics**

A t score in psychometric (psychological) testing is a specialized term that is **not the same thing as a t score that you get from a t-test.**  
T scores in [t-tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/) can be positive or negative. **T scores in psychometric testing are always positive**, with a [mean](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)of 50.

A difference of 10 (positive or negative) from the mean is a difference of one [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). For example, a score of 70 is two standard deviations above the mean, while a score of 0 is five standard deviations below the mean.

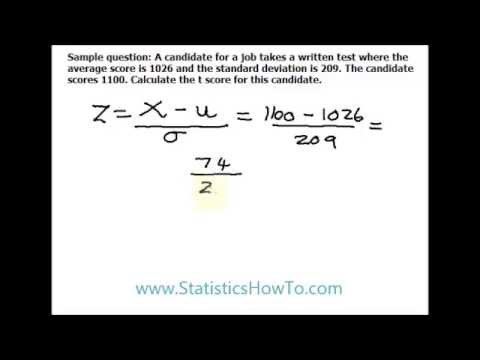
A t score is similar to a [z score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) — it represents **the number of**[**standard deviations**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)**from the mean**. While the z-score returns values from between -5 and 5 (most scores fall between -3 and 3) standard deviations from the mean, the t score has a greater value and returns results from between 0 to 100 (most scores will fall between 20 and 80). Many people prefer t scores because the lack of negative numbers means they are easier to work with and there is a larger range so decimals are almost eliminated. This table shows z-scores and their equivalent t scores.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/t-score-vs.-z-score.jpg)

***T = (Z x 10) + 50****.*

## **T Score Conversion in Psychometrics**

Watch the video or read the article below:

[](https://www.youtube.com/watch?v=7ad7ayWAVKk)

Calculating a t score is really just a **conversion** from a z score to a t score, much like [converting Celsius to Fahrenheit](http://www.metric-conversions.org/temperature/celsius-to-fahrenheit.htm). The formula to convert a z score to a t score is:

***T = (Z x 10) + 50****.*

**Sample question:** A candidate for a job takes a written test where the average score is 1026 and the standard deviation is 209. The candidate scores 1100. Calculate the t score for this candidate.

Note: If you are given the z-score for a question, skip to Step 2.

Step 1: Calculate the z score. (See: [How to calculate a z-score).](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/). The z-score for the data in this sample question is .354.

Step 2: Multiply the z score from Step 1 by 10:  
10 \* .354 = 3.54.

Step 3: Add 50 to your result from Step 2:  
3.54 + 50 = 53.54.

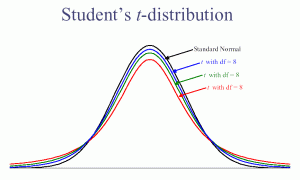
That’s it!

**Tips**:

1. z-scores and t scores both represent standard deviations from the mean, but while “0” on a z-score is 0 standard deviations from the mean, a “50” on a t score represents the same thing. That’s because t scores use a mean of 50 and z-scores use a mean of 0.
2. A t score of over 50 is above average; below 50 is below average. In general, a t score of above 60 means that the score is in the top one-sixth of the distribution; above 63, the top one-tenth. A t score  
   below 40 indicates a lowest one-sixth position; below 37, the bottom one-tenth.

## **T Test (Student’s T-Test):**

## **What is a T test?**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/students-normal.gif)

*The*[*t-distribution*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/)*, used for the t-test. Image: Carnegie Mellon.*

The t test tells you how [significant](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/)the differences between groups are; In other words it lets you know if those differences (measured in means/averages) could have happened by chance.

**A very simple example:** Let’s say you have a cold and you try a naturopathic remedy. Your cold lasts a couple of days. The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week. You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you really want to know is, are these results repeatable? A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

**Another example:** Student’s T-tests can be used in real life to compare means. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there’s always a [control group](https://www.statisticshowto.datasciencecentral.com/control-group/) (a group who are given a placebo, or “sugar pill”). The control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student’s t-test to find out if the results are repeatable for an entire population.

## **The T Score.**

The [t score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/) is a [ratio](https://www.statisticshowto.datasciencecentral.com/ratios-and-rates/)between the **difference between two groups and the difference within the groups**. The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different from each other as they are within each other. When you run a t test, the bigger the t-value, the more likely it is that the results are repeatable.

* A large t-score tells you that the groups are different.
* A small t-score tells you that the groups are similar.

## **T-Values and P-values**

How big is “big enough”? Every t-value has a [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/) to go with it. A p-value is the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/) that the results from your sample data occurred by chance. P-values are from 0% to 100%. They are usually written as a decimal. For example, a p value of 5% is 0.05. **Low p-values are good**; They indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance. In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

## **Calculating the Statistic / Test Types**

There are**three main types of t-test:**

* [**Paired t test**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/#PairedTTest) (dependent samples/ A[Paired sample t-test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/#PairedTTest)): Compares means from the same group at different times (say, one year apart). Used to compare related observations. For example, do test scores differ significantly if the test is taken at 8 a.m. or noon?
* [**Independent Samples t-test**](https://www.statisticshowto.datasciencecentral.com/independent-samples-t-test/): An [Independent Samples t-test](https://www.statisticshowto.datasciencecentral.com/independent-samples-t-test/) compares the [means](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)for two groups. For example, you could run a t test to see if the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)test scores of males and females are different;
* [**One sample t test**](https://www.statisticshowto.datasciencecentral.com/one-sample-t-test/): Tests the mean of a single group against a known mean. Used to compare a result to an [expected value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/expected-value/). For example, do males score higher than the average of 70 on a test if their exam time is switched to 8 a.m.?

You probably don’t want to calculate the test by hand (the math can get very messy, but if you insist you can find the steps for an [independent samples t test here](https://www.statisticshowto.datasciencecentral.com/independent-samples-t-test/).

## **Assumptions for the T Test**

* [**Assumption of Independence**](https://www.statisticshowto.datasciencecentral.com/assumption-of-independence/)**:** you need two independent, categorical groups that represent your [independent variable](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/). In the above example of test scores “males” or “females” would be your [independent variable](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/).
* [**Assumption of normality**](https://www.statisticshowto.datasciencecentral.com/assumption-of-normality-test/): the [dependent variable](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/) should be approximately normally distributed. The dependent variable should also be measured on a continuous scale. In the above example on average test scores, the “test score” would be the dependent variable.
* **Assumption of**[**Homogeneity of Variance:**](https://www.statisticshowto.datasciencecentral.com/homoscedasticity/) The [variances](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)of the dependent variable should be equal.

## **What is a Paired T Test (Paired Samples T Test / Dependent Samples T Test)?**

A paired t test (also called a **correlated pairs t-test**, a **paired samples t test** or **dependent samples t test**) is where you run a t test on dependent samples. Dependent samples are essentially connected — they are tests on the same person or thing. For example:

* Knee MRI costs at two different hospitals,
* Two tests on the same person before and after training,
* Two blood pressure measurements on the same person using different equipment.

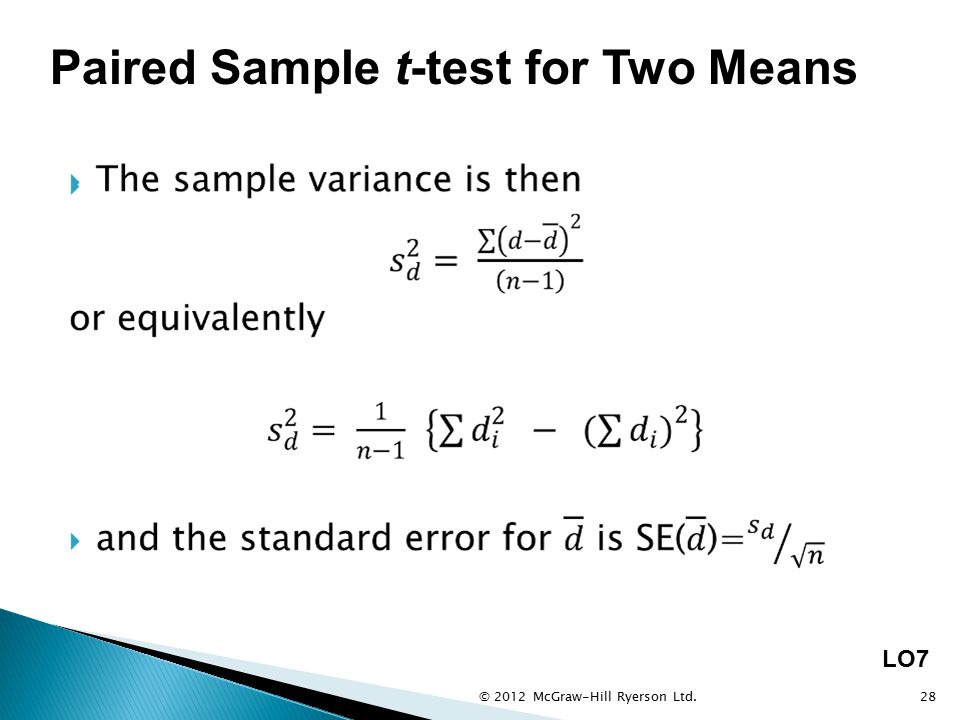
## **When to Choose a Paired T Test / Paired Samples T Test / Dependent Samples T Test**

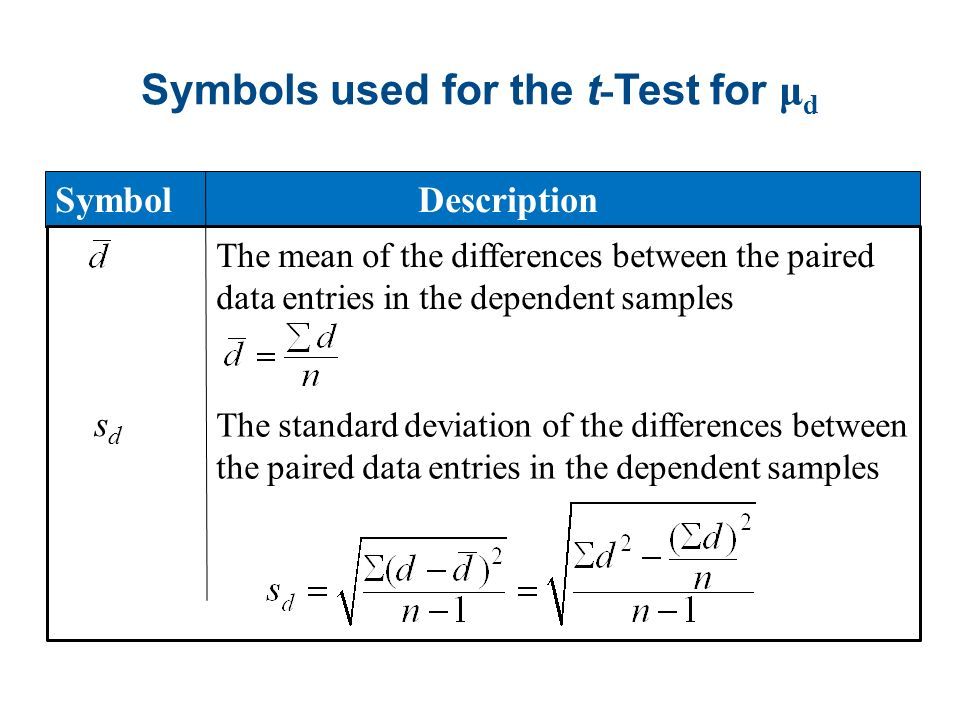
Choose the paired t-test if you have two measurements on the same item, person or thing. You should also choose this test if you have two items that are being measured with a unique condition. For example, you might be measuring car safety performance in [Vehicle Research and Testing](http://www.nhtsa.gov/VRTC) and subject the cars to a series of crash tests. Although the manufacturers are different, you might be subjecting them to the same conditions.

With a “regular” [two sample t test](https://www.statisticshowto.datasciencecentral.com/two-sample-t-test-difference-means/), you’re comparing the [means](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)for two different [samples](https://www.statisticshowto.datasciencecentral.com/sample/). For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills. If you take a [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) each group separately and they have different conditions, your samples are independent and you should run an [independent samples t test](https://www.statisticshowto.datasciencecentral.com/independent-samples-t-test/) (also called between-samples and unpaired-samples).

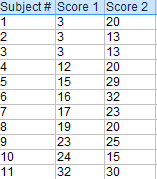
The [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state) for the for the independent samples t-test is μ1 = μ2. In other words, it assumes the means are equal. With the paired t test, the null hypothesis is that the [*pairwise difference*](https://www.statisticshowto.datasciencecentral.com/pairwise-independent-mutually/#PWD) between the two tests is equal (H0: µd = 0). The difference between the two tests is very subtle; which one you choose is based on your [data collection method](https://www.statisticshowto.datasciencecentral.com/data-collection-methods/).

## **Paired Samples T Test By hand**

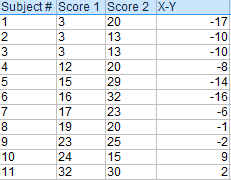




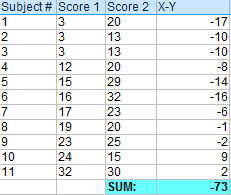
**Sample question:**Calculate a paired t test by hand for the following data:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-2.png)  
  
Note : Alternative hypothesis is there is a difference in group means. So it is two-tailed test(not equal to zero)

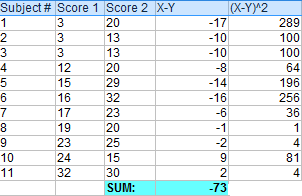
Step 1: Subtract each Y score from each X score.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-3.png)

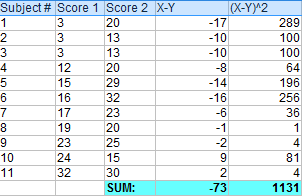
Step 2: Add up all of the values from Step 1.  
Set this number aside for a moment.

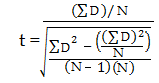
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-4.png)

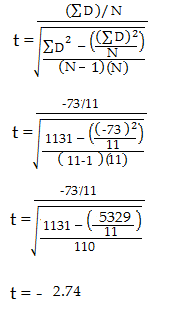
Step 3: Square the differences from Step 1.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-5.png)

Step 4: Add up all of the squared differences from Step 3.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-6.png)

Step 5: Use the following formula to calculate the t-score:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example.png)  
  
  
ΣD: Sum of the differences (Sum of X-Y from Step 2)  
ΣD2: Sum of the squared differences (from Step 4)  
(ΣD)2: Sum of the differences (from Step 2), squared.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-7v3.png)

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/) in the [t-table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/), using the [degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) in Step 6. If you don’t have a specified [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/), use 0.05 (5%). For this sample problem, with df=10, the t-value is 2.228(two-tailed)

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

In order to fully reject the null hypothesis, use both values (p and t) in combination. In other words, if you think you might reject the null based on the t-value, but your p-value is large, then don’t reject the null.

**Note**: You can ignore the minus sign when comparing the two t-values, as ± indicates the direction; the p-value remains the same for both directions.

## **What is an Independent Samples T Test?**

The independent samples t test (also called the unpaired samples t test) is the most common form of [the T test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/). It helps you to compare the [means](https://www.statisticshowto.datasciencecentral.com/mean/)of two sets of data. For example, you could run a t test to see if the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)test scores of males and females are different; the test answers the question, “Could these differences have occurred by random chance?”

This test is extremely useful because for the [z test](https://www.statisticshowto.datasciencecentral.com/z-test/) you need to know facts about the population, like the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)[standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/). With the independent samples t test, you don’t need to know this information. You should use this test when:

* You do not know the [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/) or standard deviation.
* You have two independent, separate [samples](https://www.statisticshowto.datasciencecentral.com/sample/).

## **Assumptions for the Independent Samples T Test**

* [**Assumption of Independence**](https://www.statisticshowto.datasciencecentral.com/assumption-of-independence/)**:** you need two independent, categorical groups that represent your [independent variable](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/). In the above example of test scores “males” or “females” would be your [independent variable](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/).
* [**Assumption of normality**](https://www.statisticshowto.datasciencecentral.com/assumption-of-normality-test/): the [dependent variable](https://www.statisticshowto.datasciencecentral.com/dependent-variable-definition/) should be approximately normally distributed. The dependent variable should also be measured on a continuous scale. In the above example on average test scores, the “test score” would be the dependent variable.
* **Assumption of**[**Homogeneity of Variance:**](https://www.statisticshowto.datasciencecentral.com/homoscedasticity/) The [variances](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)of the dependent variable should be equal.

## **Null and alternative hypotheses for the independent t-test**

The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal:

H0: u1 = u2

In most cases, we are looking to see if we can show that we can reject the null hypothesis and accept the alternative hypothesis, which is that the population means are not equal:

HA: u1 ≠ u2

To do this, we need to set a significance level (also called alpha) that allows us to either reject or accept the alternative hypothesis. Most commonly, this value is set at 0.05.

## **How to Run an Independent Samples T Test**

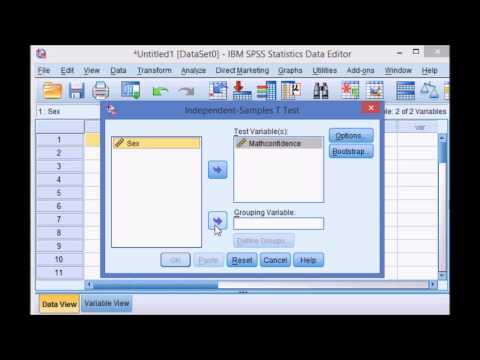
Technology is usually used to run this test. For Excel instructions, see: [How to run a T Test in Excel](https://www.statisticshowto.datasciencecentral.com/how-to-do-a-t-test-in-excel/). For instructions by hand, scroll down.

## **How to Run an Independent Samples T Test in SPSS**

Before you perform a t test in SPSS for independent samples, you should:

1. Write a hypothesis statement. For the above research question, the [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) would be that there is no significant difference
2. [Determine if your test is one-tailed or two-tailed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/)
3. Specify an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/).

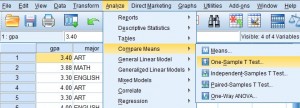
Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=DciJm7hA2fk)

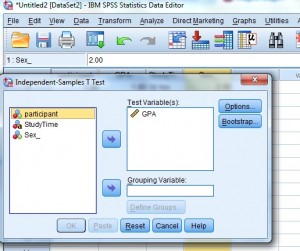
Step 1: **Open the worksheet**with the data you want to perform the t test in SPSS.

Step 2: [**Define the SPSS variables**](https://www.statisticshowto.datasciencecentral.com/how-to-enter-data-into-spss/)you want to perform a t-test on. For example, you might want to compare GPAs between male and female high school students. Therefore, you’ll want to define the variables “sex” (i.e. other male or female). [If you aren’t sure how to define variables in SPSS, click here to find out how.](https://www.statisticshowto.datasciencecentral.com/how-to-enter-data-into-spss/)

Step 3: **Click “Analyse,” then click “Compare Means,” then click “Independent Sample T Test.”**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/t-test-in-spss-12.jpg)

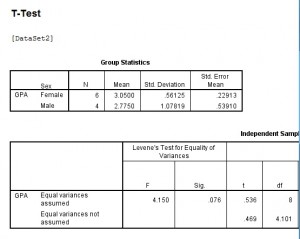
Step 4: **Select the dependent variable from the left window pane**and then click the top arrow button to move the [variables](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/types-of-variables/)over to the Test Variable(s): window. For this example, we are comparing GPAs, so the test variable we want to select is GPA.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/SPSS-t-test-2.jpg)

Step 5: **Select the independent variable in the left window**and then click the arrow to the left of the “Grouping Variable” box. The [grouping variable](https://www.statisticshowto.datasciencecentral.com/grouping-variable/) is the variable you divided into groups when you defined variables. For this example, the groups are “male” and “female” so the grouping variable you want to select is “Sex.”

Step 6: **Click “Define Groups.”**For this example, type “1” into the Group 1 box (for female) and then type “2” into the Group 2 box (for male).

Step 7: **Click “Continue” and then click “OK.”**The test is calculated and the results will appear in a new window.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/SPSS-t-test-3.jpg)

## **T Test in SPSS: Output**

Your output will include:

* The Levine’s test for equal variance (the first section of the Independent Samples Test box). If the [significance level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) is larger than .05, you should use the first line in the output table, Equal [*variances*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)assumed. If the value is .05 or lower, use the second row of results.
* [Sig (2 Tailed)](https://www.statisticshowto.datasciencecentral.com/sig2-tailed-interpreting-results/): use the value indicated in Levine’s test. If this p-value is above .05, then there is **not** a significant difference in test scores.

**Tip:** Click “Options” on the t-test window to change the [confidence interval](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/).

Check out out [YouTube Channel](https://www.youtube.com/channel/UCs3IhN8VOA_5WxpAgbSmFkg)for more SPSS videos!

## **Calculating an Independent Samples T Test By hand**

**Sample question:** Calculate an independent samples t test for the following data sets:  
Data set A: 1,2,2,3,3,4,4,5,5,6  
Data set B: 1,2,4,5,5,5,6,6,7,9

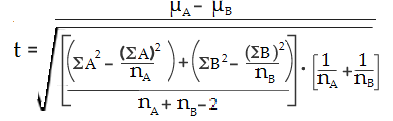
Step 1: Sum the two groups:  
A: 1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6 = 35  
B: 1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9 = 50

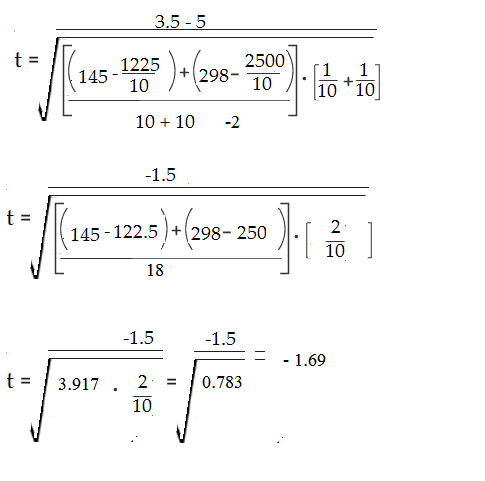
Step 2: Square the sums from Step 1:  
352 = 1225  
492 = 2500  
Set these numbers aside for a moment.

Step 3: Calculate the [means](https://www.statisticshowto.datasciencecentral.com/mean/)for the two groups:  
A: (1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5 + 6)/10 = 35/10 = 3.5  
B: (1 + 2 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 9) = 50/10 = 5  
Set these numbers aside for a moment.

Step 4: Square the individual scores and then add them up:  
A: 11 + 22 + 22 + 33 + 33 + 44 + 44 + 55 + 55 + 66 = 145  
B: 12 + 22 + 44 + 55 + 55 + 55 + 66 + 66 + 77 + 99 = 298  
Set these numbers aside for a moment.

Step 5: Insert your numbers into the following formula and solve:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/t-test-formula-a.png)  
  
  
(ΣA)2: Sum of data set A, squared (Step 2).  
(ΣB)2: Sum of data set B, squared (Step 2).  
μA: Mean of data set A (Step 3)  
μB: Mean of data set B (Step 3)  
ΣA2: Sum of the squares of data set A (Step 4)  
ΣB2: Sum of the squares of data set B (Step 4)  
nA: Number of items in data set A  
nB: Number of items in data set B

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/07/t-test-formula-2-1.png)  
  
Step 6: Find the [Degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) (nA-1 + nB-1) = 18

Step 7: Look up your degrees of freedom (Step 6) in the [t-table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/). If you don’t know what your [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) is, use 5% (0.05).  
18 degrees of freedom at an alpha level of 0.05 = 2.10.

Step 8: Compare your calculated value (Step 5) to your table value (Step 7). The calculated value of -1.79 is less than the cutoff of 2.10 from the table. Therefore p > .05. As the[p-value](https://www.statisticshowto.datasciencecentral.com/p-value/) is greater than the alpha level, we cannot conclude that there is a difference between means.

## **One Sample T Test: How to Run It, Step by Step**

The one sample t test compares the [mean](https://www.statisticshowto.datasciencecentral.com/mean/)of your sample data to a known value. For example, you might want to know how your [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/) compares to the [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/). You should run a one sample t test when you don’t know the population [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) or you have a small [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/). For a full rundown on which test to use, see:[T-score vs. Z-Score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/t-score-vs-z-score/).

Assumptions of the test (your data should meet these requirements for the test to be valid):

* Data is independent.
* Data is collected randomly.
* The data is approximately [normally distributed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/).

## **One Sample T Test Example**

**Sample question**: your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/).

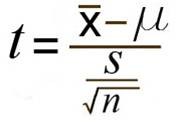
Step 1: Write your null hypothesis statement ([How to state a null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state)). The accepted hypothesis is that there is no difference in sales, so:  
H0: μ = $100.

Step 2: Write your [alternate hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/). This is the one you’re testing. You think that there is a difference (that the mean sales increased), so:  
H1: μ > $100.

Step 3: Identify the following pieces of information you’ll need to calculate the test statistic. The question should give you these items:

1. **The sample mean**(x̄). This is given in the question as $130.
2. **The population mean**(μ). Given as $100 (from past data).
3. **The sample standard deviation**(s) = $15.
4. **Number of observations**(n) = 25.

Step 4: Insert the items from above into the [t score formula](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/09/t-score.jpg)  
  
  
t = (130 – 100) / ((15 / √(25))  
t = (30 / 3) = 10  
This is your **calculated t-value**.

Step 5: Find the [t-table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/) value. You need two values to find this:

1. The alpha level: given as 5% in the question.
2. The [degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/), which is the number of items in the sample (n) minus 1: 25 – 1 = 24.

Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1.711.This is your one-tailed critical t-value.

What this critical value means is that we would expect most values to fall under 1.711. If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.

Step 6: Compare Step 4 to Step 5. The value from Step 4 **does not** fall into the range calculated in Step 5, so we can [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). The value of 10 falls into the rejection region (the left tail).

In other words, it’s highly likely that the mean sale is greater. The sales training was probably a success.

## [**T Test in Excel: Easy Steps with Video**](https://www.statisticshowto.datasciencecentral.com/how-to-do-a-t-test-in-excel/)

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/excel-pivot-table.jpg)

The [t test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/)is a way to tell if the difference between before and after results is [significant](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/)or if those results could have happened by chance. For example, a drug manufacturer might test a new drug and compare the before and after results to see if the drug was effective. It’s mostly used to test if [means](https://www.statisticshowto.datasciencecentral.com/mean/)are different. The larger the t-value, the larger the difference in the two [samples](https://www.statisticshowto.datasciencecentral.com/sample/).

You have three options in the Data Analysis Toolpak ([How to load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/)) for a t test in Excel. Read below if you aren’t sure which test to choose or skip to the section you need:

1. [Paired Two Sample for means in Excel.](https://www.statisticshowto.datasciencecentral.com/how-to-do-a-t-test-in-excel/#ttestexcel1)
2. [Two Sample T test in Excel assuming Equal Variances.](https://www.statisticshowto.datasciencecentral.com/how-to-do-a-t-test-in-excel/#ttestexcelequal)
3. [Two-sample T test in Excel assuming Unequal variances.](https://www.statisticshowto.datasciencecentral.com/how-to-do-a-t-test-in-excel/#ttestexcelunequal)

T-testing is used in [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/), when you are deciding if you should [support or reject a null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). Which t test in Excel you use depends mostly on what type of data you have. If your data has **two sets of observations from the same group**(for example, medical testing before and after a drug is administered to the same group of people), you would use the paired two sample for means. Otherwise, use a two sample test for variances.

**Paired Two Sample For Means**is used when your sample observations are naturally paired. The usual reason for performing this test is when you are testing the same group twice. For example, if you are testing a new drug, you’ll want to compare the sample before and after they take the drug to see if the results are different. This particular t test in Excel used a paired two-sample test to determine if the before and after observations are likely to have been derived from distributions with equal population means.

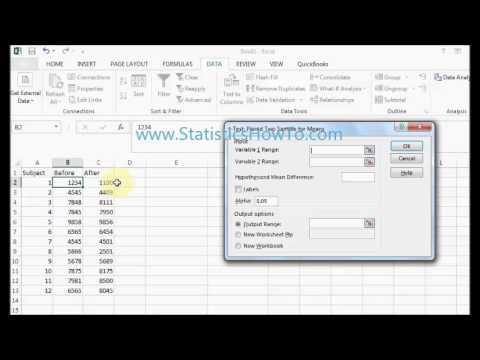
The other two tests are used when you have different groups (i.e. you aren’t testing one group twice over time). The **Two-Sample assuming Equal Variances** test is used when you know (either through the question or you have analyzed the [variance](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)in the data) that the variances are the same. The **Two-Sample assuming UNequal Variances** test is used when either:

1. You know the [variances](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/variance/)are not the same.
2. You do not know if the variances are the same or not.

In most cases, you don’t know if the variances are equal or not, so you would use the **Two-Sample assuming UNequal Variances** test.

## **Paired Two Sample for means in Excel 2013.**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=RHBIQ2reACM)

## **Two sample t test in Excel for means: Overview**

A two sample t test for means is normally used when you are testing **twice on the same subject**. For example, in a medical trial you might want to know if a particular medicine is effective so you test patients before the medication is administered and after. The t-test can tell you if the results from the trial have [statistical significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/) (i.e. it worked) or if the results probably occurred by chance.

## **Two sample t test for means in Excel: Steps**

Step 1: Type your data into Excel. As the two sample t test paired two sample for means is usually used for “before” and “after” data, you’ll probably have three columns: the first column for the subject identifier (i.e. a name or a number), the second column for the Before results and the third column for the After Results.

Step 2: State your null hypothesis ([How to state the null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state)). For example, your null hypothesis might be that the means are the same.

Step 3: Click the “Data” tab and then click “Data analysis.” If you don’t see the Data Analysis option, [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/).

Step 4: Click “t test paired two sample for means” from the options window then click “OK.”

Step 5: Click the “Variable 1 Range” box and then select your first variable list (usually the Before list).

Step 6: Click the “Variable 2 Range” box and then select your second variable list (usually the After list).

Step 7: Type a number into the [Hypothesized Mean Difference](https://www.statisticshowto.datasciencecentral.com/mean-difference/#HypothesizedMD)box. For example, if your [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) stated that there was no difference between the means, enter “0.” Otherwise, if you are hypothesizing there is a difference, type that difference into the box.

Step 8: Check the “Labels” box if you have included labels.

Step 9: Type an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) into the alpha level box. An alpha level of 0.05, or 5%, is standard in hypothesis testing so if you aren’t sure what alpha level you need, leave this at 0.05.

Step 10: Click the Output Range box and select an area to the right of your data.

Step 11: Click “OK.”

## **Reading The Results from the two sample t Test for means in Excel 2013**

Your results will include a lot of data, some that’s obvious (like the number of data items). But when you run a t-test you’re really only looking for two things: [t-scores](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/) and alpha levels.

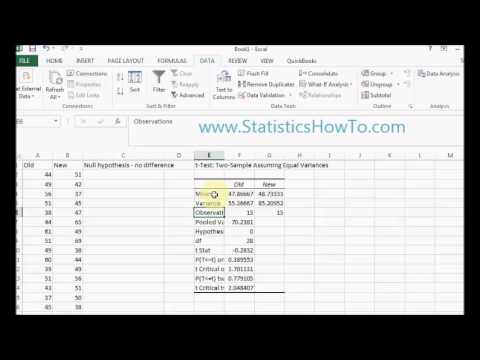
Step 1: Compare the alpha level you chose (i.e. 0.05) to the [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/)in the output. If the p-value in the output is smaller than the alpha level you chose, [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/).

Step 2: Compare the [t-critical value](https://www.statisticshowto.datasciencecentral.com/t-critical-value/)in the output with the t-value. If the t-value is larger than the t-critical value, [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). There are two t-critical values, one-tail and two-tail. If you aren’t sure if you have a[one-tailed test or a two-tailed test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/), always compare the t-value to the two-tail t critical value.

In order to fully reject the null hypothesis, use both values (p and t) in combination. In other words, if you think you might reject the null based on the t-value, but your p-value is large, then don’t reject the null.

## **Two Sample t test assuming Equal Variances.**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=OHHhzLHakKA)

## **Two sample t test Assuming Equal variances in Excel 2013: Overview**

A two sample t test assuming equal variances is used to test data to see if there is [statistical significance](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/) or if the results may have occurred randomly. This is one of three [t tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/) available in Excel and of the three, it’s the one least likely to be used. Why? In the vast majority of cases in hypothesis testing, you don’t know the [population variances](https://www.statisticshowto.datasciencecentral.com/population-variance/). This test should only be used if you have been explicitly informed that the population variances are equal. If you don’t have this information, you should be running the other t test (Two sample t test Assuming **Unequal** variances).

## **How to do a Two sample t test in Excel Assuming Equal variances: Steps**

Step 1: Type your data into a worksheet. Generally, you’ll have a list in one column and another list in a second column. The t-test will allow you to compare the means from these two columns.

Step 2: Write the null hypothesis ([How to state the null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state)). For example, your null hypothesis might be that the means are different by a certain amount.

Step 3: Click the “Data” tab and then click “Data analysis.” If you don’t see the Data Analysis button on the toolbar (to the far right of the Data tab), [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/).

Step 4: Click “t test two sample Assuming Equal variances ” from the options window then click “OK.”

Step 5: Click the “Variable 1 Range” box and then select your first data list.

Step 6: Click the “Variable 2 Range” box and then select your second data list.

Step 7: Type a number into the Hypothesized Mean Difference box. For example, if your null hypothesis stated that there was no difference between the means, type “0.”

Step 8: Check the “Labels” box (you’ll usually want to include labels so you can easily compare the two sets of data).

Step 9: Type an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) into the alpha level box. If you don’t know what alpha level you should be using, leave it at 0.05.

Step 10: Click the Output Range box and select an area for your output.

Step 11: Click the “OK” button.

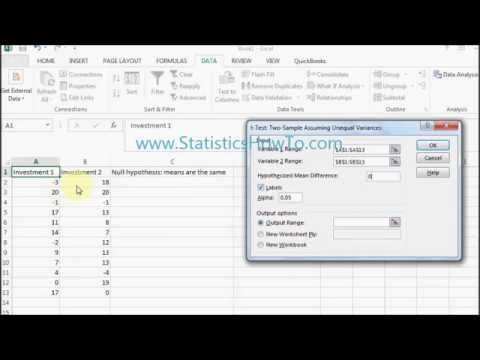
## **Reading The Results from two sample t test Assuming Equal variances in Excel 2013**

Step 1: Compare the alpha level you types into the two sample t test Assuming Equal variances window (i.e. 0.05) to the alpha level listed in the output on the worksheet. If the alpha level in the output is larger than the alpha level you chose, you will be unable to [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/).

Step 2: Compare the t-critical value in the output on the worksheet with the t-value listed. If the t-value is larger than the t-critical value, you can reject the null hypothesis. There are two t-critical values, one-tail and two-tail. If you aren’t sure if you have a[one-tailed test or a two-tailed test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/), always compare the t-value to the two-tail t critical value.

## **Two-sample t test assuming Unequal variances.**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=L-jfenou5hI)

## **Two Sample T Test in Excel Unequal Variances: Overview**

A two sample t test assuming unequal variances is the most common type of t test in Excel 2013. You have three options in Excel for [t tests](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/): assuming equal variances, assuming unequal variances and a paired two sample. The paired two sample for means in Excel is generally used if you have a sample you’re testing twice (i.e. a “Before” and an “After”) while the two sample test assuming equal variances is only used on the very rare occasion you know the population variance.

## **Two sample t test in Excel 2013 for unequal variance: Steps**

Step 1: Type your data into a worksheet in two columns.

Step 2: State your null hypothesis ([How to state the null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state)) (i.e. the means for both sets of data are the same).

Step 3: Click “Data” and then click “Data analysis.” If you don’t see Data Analysis, [load the Data Analysis Toolpak](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/).

Step 4: Click “t test two sample Assuming unequal variances ” and then click the “OK” button. This will open the t test two sample Assuming unequal variances dialog box.

Step 5: Type the location for your first set of data into the “Variable 1” box.

Step 6: Type the location for your first set of data into the “Variable 2” box.

Step 7: Type a number into the Hypothesized Mean Difference box. The hypothesized mean difference should have been stated when you wrote your null hypothesis. For example, if you think the means are the same then the hypothesized mean difference is 0.

Step 8: Check the “Labels” box (assuming you included labels for your data, which is always a good idea).

Step 9: Click the “Alpha level” box and then type an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/). the default is 0.05, which is a standard alpha level for these tests.

Step 10: Click the “Output Range” box and select an area for your output.

Step 11: Click the “OK” button.

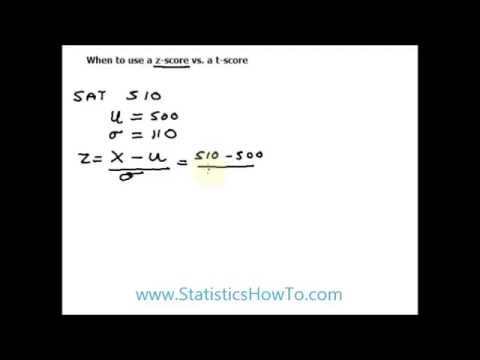
## **Reading The Results from two sample t test unequal variance Excel 2013**

1. [Reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/) if the alpha level in the output is smaller than your stated alpha level. For example, if the alpha level in the output is 0.03 and your alpha level from Step 9 was 0.05, you can reject the null hypothesis.
2. Compare the t-value with the t-critical value. If the t-value is larger than the t-critical value, reject the null hypothesis. There are two t-critical values—one for a one-tailed test and one for a two-tailed test. If you don’t know if you have a one or two tailed test, use the two tailed test figure ([How to tell if you have a one-tailed test or a two-tailed test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/)).

[For more excel statistics refer this link](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/excel-statistics/)

## **T-score vs. Z-score:**

Watch the video or read the article below:

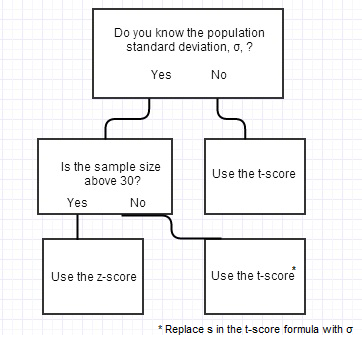
[](https://www.youtube.com/watch?v=I_hXQ1vatV0)

## **When to use a t score**

The general rule of thumb for when to use a [t score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/)is when your sample:

* Has a [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) below 30,
* Has an unknown population [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/).

You **must** know the standard deviation of the **population** and your sample size **should**be above 30 in order for you to be able to use the z-score. Otherwise, use the t-score.



*When to use a t-score vs. z-score.*

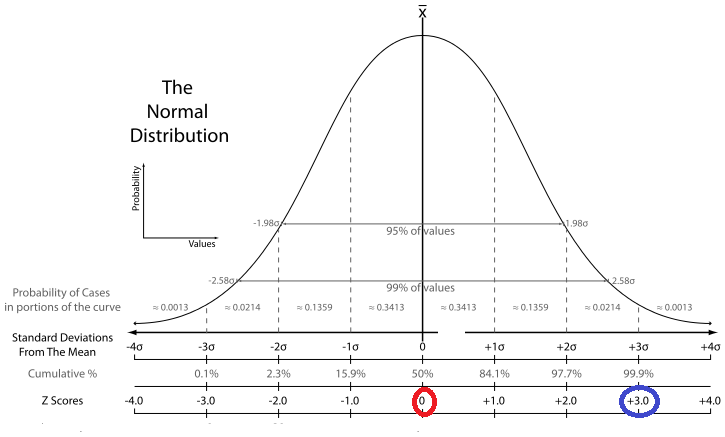
In real life though, it’s more common just to use the [t-distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/) as we usually don’t know sigma (SoSci, 1999).

*“When a sample has more than 30 observations, the normal distribution*can *be used in place of the t distribution.” (Meier et.al, p. 191).*

Note the use of the word can in the above quote; The use of the t-distribution is theoretically sound for all sample sizes, but you \*can\* choose to use the normal for sample above 30.

### **T-Score vs. Z-Score: Z-score**

Technically, z-scores are a conversion of individual scores into a standard form. The conversion allows you to more easily compare different data; it is based on your knowledge about the **population’s standard deviation and mean.**A z-score tells you how many standard deviations from the mean your result is. You can use your knowledge of normal distributions (like the [68 95 and 99.7 rule](https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/)) or the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to determine what percentage of the population will fall below or above your result.



The z-score is calculated using the formula:

**z = (X-μ)/σ**

**Where**:

* σ is the population standard deviation and
* μ is the population mean.

The z-score formula doesn’t say anything about sample size; The rule of thumb applies that your sample size should be above 30 to use it.

### **T-Score vs. Z-Score: T-score**

Like z-scores, t-scores are also a conversion of individual scores into a standard form. However, t-scores are used **when you don’t know the population standard deviation**; You make an estimate by using your sample.

**T = (X – μ) / [ s/√(n) ].**

Where:

* s is the standard deviation of the sample

.

If you have a larger sample (over 30), the t-distribution and z-distribution look pretty much the same. Therefore, you can use either. That said, if you know σ, it doesn’t make much sense to use a sample estimate instead of the “real thing”, so just substitute σ into the equation in place of s:

**T = (X – μ) / [ σ/√(n) ].**

This makes the equation identical to the one for the z-score; the only difference is you’re looking up the result in the T table, not the Z-table. For sample sizes over 30, you’ll get the same result.

# Confidence Interval

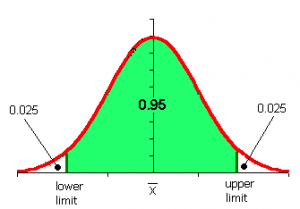
If you’re just beginning statistics, you’ll probably be finding confidence intervals using the normal distribution (see #3 below). But in reality, most confidence intervals are found using the t-distribution (especially if you are working with small samples). If you aren’t sure which technique you should be looking at, start with #1 below (how to find a confidence interval for a sample).

### **What is the Definition of a Confidence Interval?**

A confidence interval is how much [uncertainty](https://www.statisticshowto.datasciencecentral.com/uncertainty-in-statistics/)there is with any particular [statistic](https://www.statisticshowto.datasciencecentral.com/statistic/). Confidence intervals are often used with a [margin of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/#WhatMofE). It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to **survey the entire**[**population**](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). Confidence intervals are intrinsically connected to [confidence levels](https://www.statisticshowto.datasciencecentral.com/confidence-level/).

### **Confidence Intervals vs. Confidence Levels**

Confidence levels are expressed as a percentage (for example, a 95% confidence level). It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population (in other words, your statistics would be sound!). Confidence intervals are your results…usually numbers. For example, you survey a group of pet owners to see how many cans of dog food they purchase a year. You test your statistics at the 99 percent confidence level and get a confidence interval of (200,300). That means you think they buy between 200 and 300 cans a year. You’re super confident (99% is a very high level!) that your results are sound, statistically.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/ci95.png)

*Image: WUSTL.EDU*

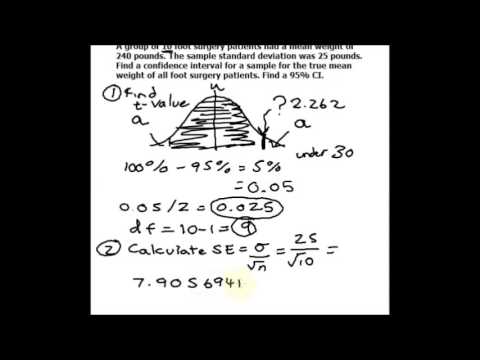
### **Real Life Examples of Confidence Intervals**

A [2008 Gallup survey found that TV ownership may be good for wellbeing](http://www.gallup.com/poll/105850/ownership-may-good-wellbeing.aspx). The results from the poll stated that the confidence level was 95% +/-3, which means that if Gallup repeated the poll over and over, using the same techniques, 95% of the time the results would fall within the published results. The 95% is the [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/) and the +/-3 is called a margin of error. At the beginning of the article you’ll see statistics (and [bar graphs](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/)). At the bottom of the article you’ll see the **confidence intervals**. For example, “For the European data, one can say with 95% confidence that the true population for wellbeing among those without TVs is between 4.88 and 5.26.” The confidence interval here is “**between 4.88 and 5.26**“.

The[U.S. Census Bureau](https://www.census.gov/programs-surveys/saipe.html)routinely uses [confidence levels](https://www.statisticshowto.datasciencecentral.com/confidence-level/) of 90% in their surveys. One survey of the number of people in poverty in 1995 stated a confidence level of 90% for the statistics “The number of people in poverty in the United States is 35,534,124 to 37,315,094.” That means if the Census Bureau repeated the survey using the same techniques, 90 percent of the time the results would fall between 35,534,124 and 37,315,094 people in poverty. **The stated figure (35,534,124 to 37,315,094) is the confidence interval.**

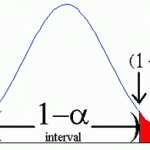
### **Confidence Interval For a Sample: Overview**

Watch the video or read on below:

[](https://www.youtube.com/watch?v=YbXEALO2OdM)

When you don’t know anything about a population’s behaviour (i.e. you’re just looking at data for a sample), you need to use the [**t-distribution**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/) to find the **confidence interval**. That’s the vast majority of cases: you usually don’t know population [parameters](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/), otherwise you wouldn’t be looking at statistics!

**The confidence interval tells you how confident you are in your results**. With any survey or experiment, you’re never 100% sure that your results could be repeated. If you’re 95% sure, or 98% sure, that’s usually considered “good enough” in statistics. That percentage of sureness is the confidence interval.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/construct-a-confidence-interval.gif)

### **Confidence Interval For a Sample: Steps**

**Question**:

*A group of 10 foot surgery patients had a*[*mean*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean)*weight of 240 pounds. The sample*[*standard deviation*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/)*was 25 pounds. Find a confidence interval for a sample for the true mean weight of all foot surgery patients. Find a 95% CI.*

**Step 1:**Subtract 1 from your [*sample size*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/). 10 – 1 = 9. This gives you [degrees of freedom,](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) which you’ll need in step 3.

**Step 2:**Subtract the confidence level from 1, then divide by two.  
(1 – .95) / 2 = .025

**Step 3:**Look up your answers to step 1 and 2 in the [*t-distribution table*](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/). For 9 degrees of freedom (**df**) and **α =**0.025, my result is 2.262.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **df** | **α = 0.1** | **0.05** | **0.025** | **0.01** | **0.005** | **0.001** | **0.0005** |
| **∞** | tα=1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.091 | 3.291 |
| **1** | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 | 636.578 |
| **2** | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.328 | 31.600 |
| **3** | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 | 12.924 |
| **4** | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| **5** | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 6.869 |
| **6** | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| **7** | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| **8** | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| **9** | 1.383 | 1.833 | 2.262 |  |  |  |  |

**Step 4:**Divide your sample [*standard deviation*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) by the square root of your sample size.  
25 / √(10) = 7.90569415

**Step 5:**Multiply step 3 by step 4.  
2.262 × 7.90569415 = 17.8826802

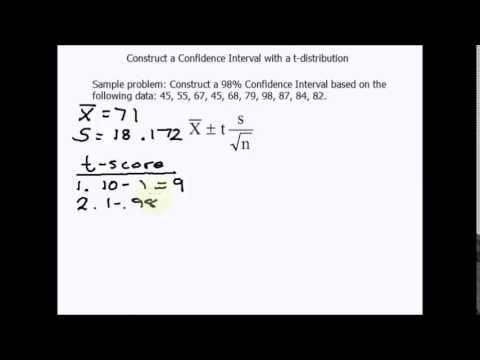
**Step 6:**For the lower end of the range, subtract step 5 from the [*sample mean*](https://www.statisticshowto.datasciencecentral.com/sample-mean/).  
240 – 17.8826802 = 222.117

**Step 7:**For the upper end of the range, add step 5 to the sample mean.  
240 + 17.8826802 = 257.883

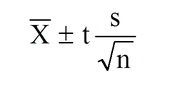
That’s how to find the confidence interval for a sample!

### **How to Find a Confidence Interval Example 2 (small sample)**

Watch the video or read the steps below:

[](https://www.youtube.com/watch?v=gDcMG4QK-m4)

If you have one small set of data (under 30 items), you’ll want to use the [t-distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/) instead of the [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) to construct your confidence interval.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/t-dist.jpg)

*The formula for constructing a CI with the t-distribution.*

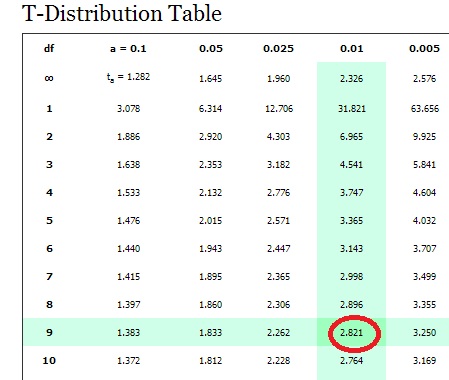
**Sample problem**: Construct a 98% Confidence Interval based on the following data: 45, 55, 67, 45, 68, 79, 98, 87, 84, 82.

**Step 1:**Find the mean, μ and [standard deviation, σ](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) for the data.  
σ: 18.172.  
μ: 71  
Set these numbers aside for a moment.

**Step 2:**Subtract 1 from your [*sample size*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) to find the [*degrees of freedom*](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) (df). We have 10 numbers listed, so our [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) is 10, so our **df = 9**. Set this number aside for a moment.

**Step 3:**Subtract the [*confidence level*](https://www.statisticshowto.datasciencecentral.com/confidence-level/) from 1, then divide by two. This is your [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/).  
(1 – .98) / 2 = **.01**

**Step 4:**Look up **df** (Step 2) and **α** (Step 3) in the [t-distribution table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/). For **df** = 9 and **α** = .01, the table gives us **2.821**.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/t-distribution.jpg)

*Degrees of freedom in the left column of the t distribution table.*

**Step 5:**Divide your std dev (step 1) by the square root of your sample size.  
18.172 / √(10) = **5.75**

**Step 6:**: Multiply step 4 by step 5.  
2.821 × 5.75 = **16.22075**

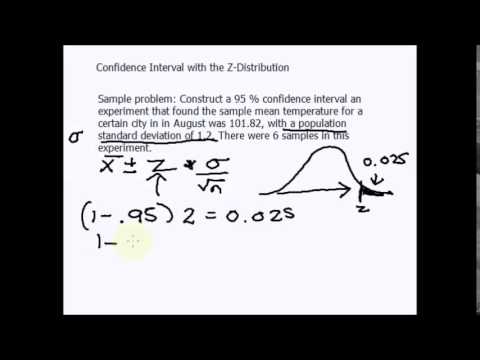
**Step 7:**For the lower end of the [*range*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/range-statistics/), subtract step 6 from the mean (Step 1).  
71 – 16.22075 = **54.77925**

**Step 8:**For the upper end of the range, add step 6 to the mean (Step 1).  
71 + 16.22075 = **87.22075**

That’s how to find a confidence interval using the t-distribution!

### **Confidence Interval with the Normal Distribution / Z-Distribution**

Watch the video or read the article below:

[](https://www.youtube.com/watch?v=vRtC5BEbhRM)

If you don’t know your [population mean](https://www.statisticshowto.datasciencecentral.com/population-mean/) (μ) but you do know the [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) (σ), you can find a confidence interval for the population mean, with the formula:

x̄ ± z\* σ / (√n),

**Sample problem**: Construct a 95 % confidence interval an experiment that found the sample mean temperature for a certain city in August was 101.82, with a population standard deviation of 1.2. There were 6 samples in this experiment.

**Step 1:**Subtract the confidence level (Given as 95 percent in the question) from 1 and then divide the result by two. This is your [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/), which represents the area in one tail.  
(1 – .95) / 2 = **.025**

**Step 2:**Subtract your result from Step 1 from 1 and then look that area up in the middle of the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to get the [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/):

1. 1 – 0.025 = 0.975
2. z score = 1.96.

**Step 3:**Plug the numbers into the second part of the formula and solve:  
z\* σ / (√n)  
= 1.96 \* 1.2/√(6)  
= 1.96 \* 0.49  
= 0.96

**Step 4:**For the lower end of the range, subtract step 3 from the mean.  
101.82 – 0.96 = 100.86

**Step 5:**For the upper end of the range, add step 3 to the  mean.  
101.82 + 0.96 = 102.78.

The CI is (100.86,102.78)

### **How to Find a Confidence Interval for a Proportion: Overview**

When we talk about a confidence interval (CI), we’re dealing with data. For example, let’s say the manager for that job you applied for told you he would get back with you in a “couple of days.” A couple of days could mean two. Or three. Or there might be a paperwork backlog and it could be a week. It definitely doesn’t mean in an hour. So your CI would probably be between**2 and 4 days.**

Perhaps the trickiest part of CIs is recognizing the various parts needed for the formula, like z a/2. This section breaks everything down into simple steps and shows you how to find a confidence interval for population proportions.

### **How to Find a Confidence Interval for a Proportion: Steps**

[Watch the video](https://youtu.be/gClMYzniK3M)

Question*: 510 people applied to the Bachelor’s in Elementary Education program at Florida State College. Of those applicants, 57 were men. Find the 90% CI of the true proportion of men who applied to the program.*

**Step 1:** Read the question carefully and figure out the following [*variables*](https://www.statisticshowto.datasciencecentral.com/variable/):

* **α** : subtract the given CI from 1.  
  1-.9=.10
* **z α/2**: divide α by 2, then look up that area in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/). Not sure how to read a z table? See this article, [**area to the right of a z-score**](https://www.statisticshowto.datasciencecentral.com/area-to-the-right-of-a-z-score-how-to-find-it/)for an explanation on how to read the z-table.  
  .1/2=.0500. The closest z-value to an area of .0500 is 0.13
* phat: Divide the proportion given (i.e. the smaller number)by the [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/).  
  57/510=0.112
* qhat : Subtract phatfrom 1.  
  1-0.112 = 0.888

**Step 2:** Multiply *phat*by *qhat*.  
0.112 x 0.888 = 0.099456

**Step 3:** Divide step 2 by the sample size.  
0.099456 / 510 = 0.000195011765

**Step 4:**Take the square root of step 3:  
sqrt(0.000195011765) = 0.0139646613

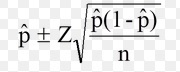
**Step 5:** Multiply step 4 by z a/2:  
0.0139646613 x 0.13 = 0.00182

**Step 6:**: For the lower percentage, subtract step 5 from phat.    
0.112-0.00182 =  .11018 = 11.018%

**Step 7:**For the upper percentage, add step 5 to phat.  
0.112 + 0.00182 =  0.11382 = 11.382%

This next method involves plugging in numbers into the actual formula. You’ll get the same results if you use the “formula free” method above or if you use the steps below.

Confidence intervals for a proportion are calculated using the following formula:

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/confidence-interval-for-a-proportion-1.jpg)

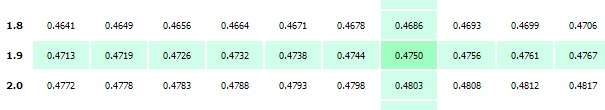
The formula might look daunting, but all you really need are two pieces of information: the [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) and the P-hat. You should be familiar with looking up z-scores from previous sections on the normal distribution (if you need a refresher, be sure to watch the above video) and P-hat is just dividing the number of events by the number of trials. Once you’ve figured those two items out, the rest is basic math.

### **Confidence Interval for a Proportion Example 2: Steps**

**Sample question:** Calculate a 95% confidence interval for the true [population proportion](https://www.statisticshowto.datasciencecentral.com/population-proportion/) using the following data:  
Number of trials(n) = 160  
Number of events (x) = 24

Step 1: **Divide your confidence level by 2**: .95/2 = 0.475.

Step 2: Look up the value you calculated in Step 1 in the [**z-table**](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) and find the corresponding z-value. The z-value that has an area of .475 is 1.96.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/confidence-interval-for-a-proportion-2.jpg)

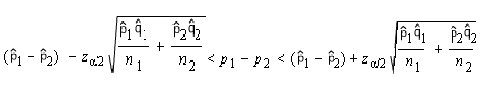
Step 3: **Divide the number of events by the number of trials** to get the “P-hat” value: 24/160 = 0.15.

Step 4: **Plug your numbers into the formula and solve**:

1. 0.15±(1.96) √ ((0.15(1-0.15)/160))=
2. 0.15±(1.96) √ ((0.15(**0.85**)/160))=
3. 0.15±(1.96) √ ((**0.1275**)/160))=
4. 0.15±(1.96) √ (**0.000796875**)=
5. 0.15±(1.96) **0.0282289744765905**=
6. 0.15±**0.0553** =
7. 0.15-0.0553=0.0947 <--this is your lower confidence interval for a proportion
8. 0.15+0.0553=0.2053 <--this is your upper confidence interval for a proportion

Your answer can be expressed as: **(0.0947,0.2.053).**

### **How to Find a Confidence Interval for Two Populations (Proportions)**

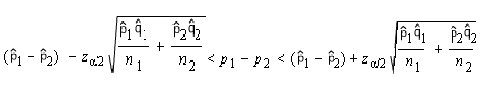
Finding confidence intervals for two populations can look daunting, especially when you take a look at the ugly equation below.  
  
It looks a lot worse than it is, because the right side of the equation is actually a repeat of the left! Finding confidence intervals for two populations can be broken down to an easy three steps.

**Example question:** A study revealed that 65% of men surveyed supported the war in Afghanistan and 33% of women supported the war. If 100 men and 75 women were surveyed, find the 90% **confidence interval** for the data’s true difference in proportions.

**Step 1:** Find the following [*variables*](https://www.statisticshowto.datasciencecentral.com/variable/)from the information given in the question:  
n1 (population 1)=100

Phat1 (population 1, positive response): 65% or 0.65  
Qhat1 (population 1, negative response): 35% or 0.35  
n2(population 2)=75  
Phat2 (population 2, positive response): 33% or 0.33  
Qhat2 (population 2, negative response): 67% or 0.67

**Step 2:** Find zα/2  
(If you’ve forgotten how to find α/2, see the directions in: How to Find a Confidence Interval for a Proportion above)  
zα/2=0.13

**Step 3:** Enter your data into the following formula and solve:  


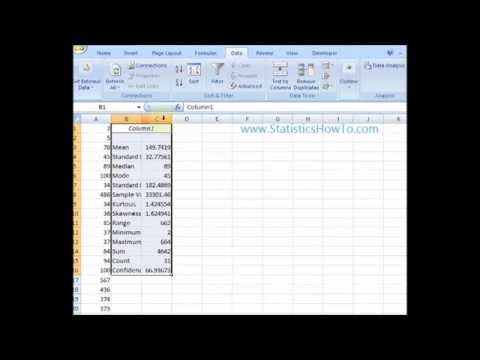
If formulas scare you, here’s the step-by-step to solve the equation (refer back to step 1 for the variables):

1. multiply phat1 and qhat1 together (.65 x .35 = .2275)
2. divide your answer to (1) by n1. Set this number aside. (.2275 x 100=.00275)
3. multiply phat2 and qhat2 together (.33 x .67=.2211).
4. divide your answer to (3) by n2 (.2211/75=.002948).
5. Add (3) and (4) together (.00275 + .002948=.005698)
6. Take the square root of (5): (sqrt.005698=.075485)
7. Multiply (6) by zα/2 found in Step 2. (.075485 x 0.13=.0098). Set this number aside.
8. Subtract phat2 from phat1 (.65-.33=.32).
9. Subtract (8) from (7) to get the left limit (.32-0.0098 = 31.9902)
10. Add (7) to (8) to get the right limit (.32+.0.0098=32.0098)

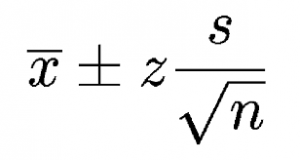
That’s it!

### **Confidence Interval for the Mean in Excel**

Watch the video or read the article below:

[](https://www.youtube.com/watch?v=yzvjz9hXvVY)

### **How to Find a Confidence Interval for the Mean in Excel: Overview**

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/ci-for-the-mean-formula.png)

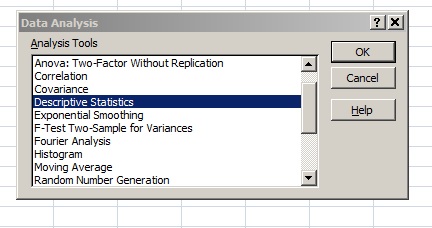
*CI for the mean formula.*

**A confidence interval for the mean is a way of estimating the true population mean**. Instead of a single number for the mean, a confidence interval gives you a lower estimate and an upper estimate. For example, instead of “6” as the mean you might get {5,7}, where 5 is the lower estimate and 7 is the upper. The narrower the estimate, the more precise your estimate is. The equations involved in statistics often involve a lot of minor calculations (such as summation), plus you would also need to calculate the margin of error and the mean of the sample. It’s very easy for errors to slip in if you calculate the confidence interval by hand. However, [Excel](https://office.microsoft.com/en-us/excel/)can calculate the mean of the sample, the margin or error and confidence interval for the mean for you. All you have to do is provide the data —which for this technique must be a sample greater than about 30 to give an accurate confidence interval for the mean.

### **How to Find a Confidence Interval for the Mean in Excel: Steps**

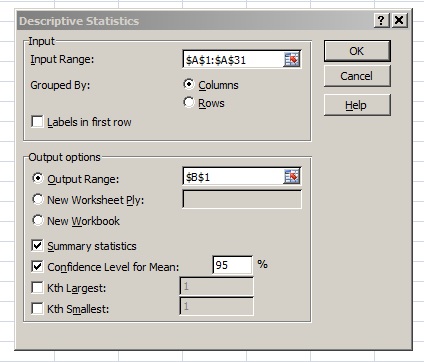
**Sample problem:**Calculate the 95 percent confidence interval for the mean in Excel using the following sample: 2,5,78,45,69,100,34,486,34,36,85,37,37,84,94,100,567,436,374,373,664,45,68,35,56,67,87,101,356,56,31.

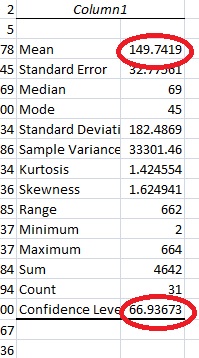
Step 1: **Type your data into a single column in Excel.** For this example, type the data into cells A1:A31.

Step 2: **Click the “Data” tab,**then click “Data Analysis,” then click “Descriptive Statistics” and “OK.” If you don’t see Data Analysis, [*load the Excel data analysis toolpak*](https://www.statisticshowto.datasciencecentral.com/excel-data-analysis-toolpak/).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/09/confidence-interval-for-the-mean-excel.jpg)

Step 3: **Enter your input range into the Input Range box**. For this example, your input range is “A1:A31”.

Step 4: **Type an output range into the Output Range box.** This is where you want your answer to appear. For example, type “B1.”

Step 5: **Click the “Summary Statistics” check box** and then place your chosen confidence level into the ‘Confidence Level for Mean’ check box. For this example, type “95”.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/09/ci-mean-excel.jpg)

Step 5: **Click “OK.”**Microsoft Excel will return the confidence interval for the mean and the margin of error for your data. For this sample, the mean (Xbar) is 149.742 and the margin of error is 66.9367. So the mean has a lower limit of 149.742-66.936 and an upper limit of 149.742+66.936.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2012/09/confidence-interval-for-the-mean-excel-results.jpg)  
That’s it!

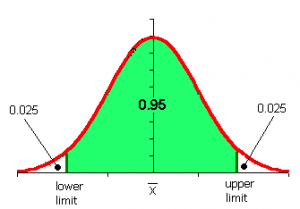
**Warning:** A 99 percent confidence interval doesn’t mean that there’s a 99 percent probability that the calculated interval has the actual mean. Your sample is either going to contain the actual mean, or it isn’t. Over the long-term, if you ran tests on many, many samples, there is a 99 percent probability that the calculated intervals would contain the true mean.

### **The 95% Confidence Interval Explained**

The terms confidence **level**and confidence **interval**are often confused.

A **95% confidence level** means is that if the survey or experiment were repeated, 95 percent of the time the data would match the results from the entire [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). Sometimes you just can’t survey everyone because of time or cost (think about how much it would cost to do a telephone survey of over 300 million Americans!).Therefore, you take a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)of the population. Having a 95% confidence level means that you’re almost certain your results are the same as if you had surveyed everyone.

A **95% confidence interval** gives you a very specific set of numbers for your confidence level. For example, let’s suppose you were surveying a local school to see what the student’s stats test scores are. You set a 95% confidence level and find that the 95% confidence interval is (780,900). That means if you repeated this over and over, 95 percent of the time the scores would fall somewhere between 780 and 900.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/ci95.png)

*Image: WUSTL.EDU*

The above image shows a 95% confidence interval on a [normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) graph. The red “tails” are the remaining 5 percent of the interval. Each tail has 2.5 percent (that’s .025 as a decimal). You don’t have to draw a graph when you’re working with confidence intervals, but it can help you visualize exactly what you are doing — especially in [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/). If your results fall into the red region, then that’s outside of the 95% confidence level that you, as a researcher, set.

If you have a small sample or if you don’t know the population [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/) which in most real-life cases is true), then you’ll find the 95% Confidence Interval with a [t-distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/).

### **Asymmetric Confidence Interval**

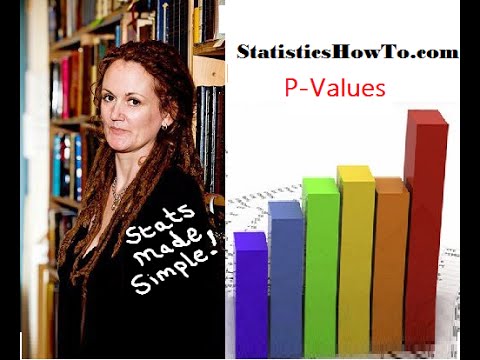
An asymmetric confidence interval just means that the [point estimate](https://www.statisticshowto.datasciencecentral.com/point-estimate/) doesn’t lie in the exact center of the CI. You can end up with asymmetric CIs for many reasons, including:

* You transform your data (for example, using log transformations).
* You incorporate [random error](https://www.statisticshowto.datasciencecentral.com/systematic-error-random-error/).
* You incorporate [systematic bias](https://www.statisticshowto.datasciencecentral.com/systematic-error-random-error/)into the interval:
  + A positive systematic bias will increase the right side of the interval.
  + A negative systematic bias will increase the left side of the interval.

# P-Value in Statistical Hypothesis Tests: What is it?

## **P Value Definition**

Watch the video or read the article below:

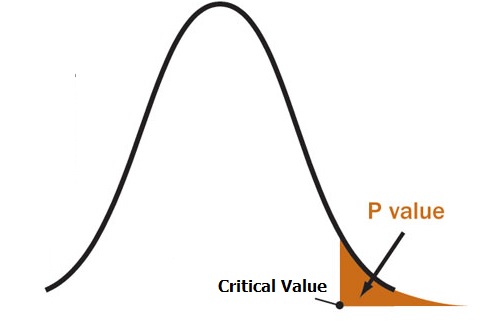
[](https://www.youtube.com/watch?v=oAfpRO5WZ7o)

A p value is used in [hypothesis testing](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/) to help you [support or reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). The p value is the evidence **against** a [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/). The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you [convert them to a percentage](http://wagner.nyu.edu/files/students/Math_Review_-_Review_topics_-_Percents.pdf). For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That’s pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important (“significant”) your results.

When you run a [hypothesis test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/), you compare the p value from your test to the [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/)you selected when you ran the test. Alpha levels can also be written as percentages.

Graphically, the p value is the area in the **tail** of a [probability distribution](https://www.statisticshowto.datasciencecentral.com/probability-distribution/). It’s calculated when you run hypothesis test and is the area to the right of the test statistic (if you’re running a two-tailed test, it’s the area to the left and to the right).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/01/p-value1.jpg)

## **P Value vs Alpha level**

Alpha levels are controlled by the researcher and are related to [confidence levels](https://www.statisticshowto.datasciencecentral.com/confidence-level/). You get an alpha level by subtracting your confidence level from 100%. For example, if you want to be 98 percent confident in your research, the alpha level would be 2% (100% – 98%). When you run the hypothesis test, the test will give you a value for p. Compare that value to your chosen alpha level. For example, let’s say you chose an alpha level of 5% (0.05). If the results from the test give you:

* **A small p**(≤ 0.05), [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). This is strong evidence that the null hypothesis is invalid.
* **A large p**(> 0.05) means the [alternate hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/) is weak, so you do not reject the null.

## **P Values and**[**Critical Values**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/)

The p value is just one piece of information you can use when deciding if your [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) is true or not. You can use other values given by your test to help you decide. For example, if you run an[f test two sample for variances in Excel](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/f-test/#excel), you’ll get a p value, an f-critical value and a [f-value](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/f-statistic-value-test/#ANOVA).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/01/p-value.jpg)  
  
In the above image, the results from the [f-test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/f-test/) show a large p value (.244531, or 24.4531%), so you would not reject the null. However, there’s also another way you can decide: compare your f-value with your f-critical value. If the f-critical value is smaller than the f-value, you should reject the null hypothesis. In this particular test, the p value and the f-critical values are both very large so you do not have enough evidence to reject the null.

## **What if I Don’t Have an Alpha Level?**

In an ideal world, you’ll have an alpha level. But if you do not, you can still use the following rough guidelines in deciding whether to support or reject the null hypothesis:

* If p > .10 → “not significant”
* If p ≤ .10 → “marginally significant”
* If p ≤ .05 → “significant”
* If p ≤ .01 → “highly significant.”

## **When to Reject the Null hypothesis**

Basically, you reject the null hypothesis when your test value falls into the [rejection region](https://www.statisticshowto.datasciencecentral.com/rejection-region/). There are four main ways you’ll compute test values and either support or reject your null hypothesis. Which method you choose depends mainly on if you have a proportion or a [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/).

[support or reject null hypothesis](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/10/xXbar-mu1.bmp)

## **Support or Reject the Null Hypothesis: Steps**

[](https://www.youtube.com/watch?v=UgqwBYSqc-A)

If you have a [P-value](https://www.statisticshowto.datasciencecentral.com/p-value/), or are asked to find a p-value, follow these instructions to support or reject the null hypothesis. This method works if you are given an [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) and if you are not given an alpha level. If you are given a [confidence level](https://www.statisticshowto.datasciencecentral.com/confidence-level/), just subtract from 1 to get the alpha level. See: [How to calculate an alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/#alphacalc).

**Step 1:** [*State the null hypothesis*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state) and the[*alternate hypothesis*](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/) (“the claim”).  
If you aren’t sure how to do this, follow [this link for How To State the Null and Alternate Hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state).

**Step 2:** Find the [*critical value*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-critical-values/). We’re dealing with a [normally distributed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) population, so the critical value is a [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).  
Use the following formula to find the [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/).

null hypothesis z formula 

[Click here if you want easy, step-by-step instructions for solving this formula.](https://www.statisticshowto.datasciencecentral.com/practically-cheating-statistics-handbook/help-with-statistics-equations/)

**Step 4:** Find the [P-Value](https://www.statisticshowto.datasciencecentral.com/p-value/) by looking up your answer from step 2 in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/). To get the p-value, subtract the area from 1. For example, if your area is .990 then your p-value is 1-.9950 = 0.005. Note: for a two-tailed test, you’ll need to halve this amount to get the p-value in one tail.

**Step 5:** Compare your answer from step 4 with the α value given in the question. Should you support or reject the null hypothesis?  
If step 4 is less than or equal to α, reject the null hypothesis, otherwise do not reject it.

## **[P-Value Guidelines](https://www.statisticshowto.datasciencecentral.com/p-value/)**

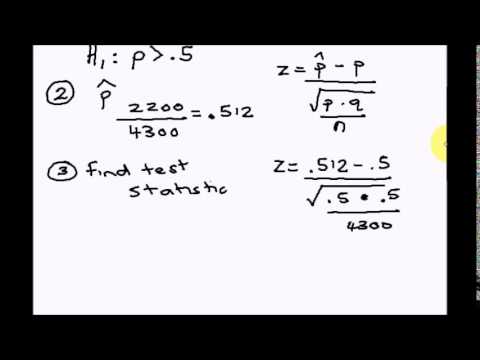
Use these general guidelines to decide if you should reject or keep the null:

If p value > .10 → “not [significant](https://www.statisticshowto.datasciencecentral.com/what-is-statistical-significance/)”  
If p value ≤ .10 → “marginally significant”  
If p value ≤ .05 → “significant”  
If p value ≤ .01 → “highly significant.”

That’s it!

## **Support or Reject Null Hypothesis for a Proportion**

Sometimes, you’ll be given a proportion of the population or a percentage and asked to support or reject null hypothesis. In this case you can’t compute a test value by calculating a [**z-score**](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) (you need actual numbers for that), so we use a slightly different technique.

[](https://www.youtube.com/watch?v=H-Nl-PpPOcU)

**Sample question:** A researcher claims that Democrats will win the next election. 4300 voters were polled; 2200 said they would vote Democrat. Decide if you should support or reject null hypothesis. Is there enough evidence at α=0.05 to support this claim?

**Step 1:** State the [*null hypothesis*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) and the [*alternate hypothesis*](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/) (“the claim”).  
Ho:p ≤ 0.5  
H1:p > .5

**Step 2:** Compute *phat* by dividing the number of positive respondents from the number in the [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/):  
2200/4300 = 0.512.

**Step 3:** Use the following formula to calculate your test value.

test value with a proportion

Where:  
Phat is calculated in Step 2  
P the [null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/) [p value](https://www.statisticshowto.datasciencecentral.com/p-value/) (.05)  
Q is 1 – p

The z-score is:  
.512 – .5 / √(.5(.5) / 4300)) = 1.57

Step 4: Look up Step 3 in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to get .9418.

Step 5: Calculate your [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/)by subtracting Step 4 from 1.  
1-.9418 = .0582

**Step 6:** Compare your answer from step 5 with the α value given in the question. Support or reject the null hypothesis? If step 5 is less than α, reject the null hypothesis, otherwise do not reject it. In this case, .582 (5.82%) is not less than our α, so we do not reject the null hypothesis.

## **Support or Reject Null Hypothesis for a Proportion: Second example**

**Sample question:** A researcher claims that more than 23% of community members go to church regularly. In a recent survey, 126 out of 420 people stated they went to church regularly. Is there enough evidence at α = 0.05 to support this claim? Use the P-Value method to support or reject null hypothesis.

**Step 1:** [*State the null hypothesis*](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state) and the[*alternate hypothesis*](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/) (“the claim”). Ho:p ≤ 0.23; H1:p > 0.23 (claim)

**Step 2:** Compute *phat* by dividing the number of positive respondents from the number in the random sample:  
126 / 420 = 0.3.

**Step 3:**Find ‘p’ by converting the stated claim to a decimal:  
23% = 0.23.  
Also, find ‘q’ by subtracting ‘p’ from 1: 1 – 0.23 = 0.77.

**Step 4:**Use the following formula to calculate your test value.

HYPOTHESIS test value with a proportion

[Click here if you want easy, step-by-step instructions for solving this formula.](https://www.statisticshowto.datasciencecentral.com/practically-cheating-statistics-handbook/help-with-statistics-equations/)

If formulas confuse you, this is asking you to:

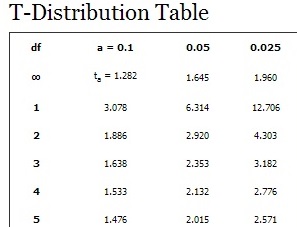
1. Subtract p fromphat(0.3 – 0.23 = 0.07). Set this number aside.
2. Multiply p and q together, then divide by the number in the random sample. (0.23 x 0.77) / 420 = 0.00042
3. Take the square root of your answer to 2. √(0.00042) = 0.0205
4. Divide your answer to 1. by your answer in 3. 0.07 / 0.0205 = 3.41

**Step 5:** Find the [P-Value](https://www.statisticshowto.datasciencecentral.com/p-value/) by looking up your answer from step 5 in the [z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/).The z-score for 3.41 is .4997. Subtract from 0.500: 0.500-.4997 = 0.003.

**Step 6:** Compare your [*P-value*](https://www.statisticshowto.datasciencecentral.com/p-value/) to α. Support or reject null hypothesis? If the [P-value](https://www.statisticshowto.datasciencecentral.com/p-value/) is less, reject the null hypothesis. If the P-value is more, keep the null hypothesis.  
0.003 < 0.05, so we have enough evidence to reject the null hypothesis and accept the claim.

**Note:** In Step 5, I’m using the [z-table on this site](https://www.statisticshowto.datasciencecentral.com/tables/z-table/) to solve this problem. Most textbooks have the [right of z-table](https://www.statisticshowto.datasciencecentral.com/tables/z-table/). If you’re seeing .9997 as an answer in your textbook table, then your textbook has a “whole z” table, in which case don’t subtract from .5, subtract from 1. 1-.9997 = 0.003.

## **Degrees of Freedom**



*Degrees of freedom in the left column of the t distribution table.*

Degrees of freedom of an estimate is **the number of independent pieces of information that went into calculating the estimate**. It’s not quite the same as the number of items in the sample. In order to get the df for the estimate, you have to subtract 1 from the number of items. Let’s say you were finding the mean weight loss for a low-carb diet. You could use 4 people, giving 3 degrees of freedom (4 – 1 = 3), or you could use one hundred people with df = 99.

In math terms (where “n” is the number of items in your set):

*Degrees of Freedom = n – 1*

**Why do we subtract 1 from the number of items?** Another way to look at degrees of freedom is that they are **the number of values that are free to vary**in a data set. What does “free to vary” mean? Here’s an example using the mean (average):

**Q**. Pick a set of numbers that have a mean (average) of 10.  
**A**. Some sets of numbers you might pick: 9, 10, 11 or 8, 10, 12 or 5, 10, 15.  
Once you have chosen the first two numbers in the set, the third is fixed. In other words, **you can’t choose the third item in the set**. The only numbers that are free to vary are the first two. You can pick 9 + 10 or 5 + 15, but once you’ve made that decision you **must** choose a particular number that will give you the mean you are looking for. So degrees of freedom for a set of three numbers is TWO.

For example: if you wanted to [find a confidence interval for a sample](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/), degrees of freedom is n – 1. “N’ can also be the number of classes or categories. See:[Critical chi-square value](https://www.statisticshowto.datasciencecentral.com/how-to-find-a-critical-chi-square-value/) for an example.

## **Degrees of Freedom: Two Samples**

If you have two [samples](https://www.statisticshowto.datasciencecentral.com/sample/)and want to find a [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/), like the [mean](https://www.statisticshowto.datasciencecentral.com/mean/), you have two “n”s to consider (sample 1 and sample 2). Degrees of freedom in that case is:

*Degrees of Freedom (Two Samples): (N1 + N2) – 2.*

## **Degrees of Freedom in ANOVA**

Degrees of freedom becomes a little more complicated in [ANOVA](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/anova/)tests. Instead of a simple parameter (like finding a mean), ANOVA tests involve comparing known means in sets of data. For example, in a one-way ANOVA you are comparing two means in two cells. The [grand mean](https://www.statisticshowto.datasciencecentral.com/grand-mean/) (the [average](https://www.statisticshowto.datasciencecentral.com/arithmetic-mean/)of the averages) would be:  
Mean 1 + mean 2 = grand mean.

What if you chose mean 1 and you knew the grand mean? You wouldn’t have a choice about Mean2, so your degrees of freedom for a two-group ANOVA is 1.

*Two Group ANOVA df1 = n – 1*

For a three-group ANOVA, you can vary two means so degrees of freedom is 2.

It’s actually a little more complicated because there are **two** degrees of freedom in ANOVA: df1 and df2. The explanation above is for df1. Df2 in ANOVA is the total number of observations in all cells – degrees of freedoms lost because the cell means are set.

*Two Group ANOVA df2 = n – k*

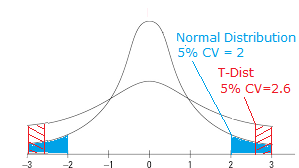
The “k” in that formula is the number of cell means or groups/conditions.  
For example, let’s say you had 200 observations and four cell means. Degrees of freedom in this case would be: Df2 = 200 – 4 = 196.

## **Why Do Critical Values Decrease While DF Increase?**

Let’s take a look at the t-score formula in a hypothesis test:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/11/t-score.jpg)  
When n increases, the t-score goes up. This is because of the square root in the denominator: as it gets larger, the fraction s/√n gets smaller and the t-score (the result of another fraction) gets bigger. As the degrees of freedom are defined above as n-1, you would think that the [t-critical value](https://www.statisticshowto.datasciencecentral.com/t-critical-value/) should get bigger too, but they don’t: they get smaller. This seems counter-intuitive.

However, **think about what a t-test is actually for**. You’re using the t-test because you don’t know the standard deviation of your population and therefore you don’t know the shape of your graph. **It could have short,**[**fat tails**](https://www.statisticshowto.datasciencecentral.com/fat-tail-distribution/)**. It could have long skinny tails**. You just have no idea. The degrees of freedom affect the shape of the graph in the t-distribution; **as the df get larger, the area in the tails of the distribution get smaller. As df approaches infinity, the t-distribution will look like a normal distribution. When this happens, you can be certain of your standard deviation (which is 1 on a normal distribution)**.

Let’s say you took repeated sample weights from four people, drawn from a population with an unknown standard deviation. You measure their weights, calculate the mean difference between the sample pairs and repeat the process over and over. The tiny sample size of 4 will result a t-distribution with fat tails. The fat tails tell you that you’re more likely to have extreme values in your sample. You test your hypothesis at an alpha level of 5%, which **cuts off the last 5% of your distribution**. The graph below shows the t-distribution with a 5% cut off. This gives a critical value of 2.6. (**Note**: I’m using a hypothetical t-distribution here as an example–the CV is not exact).

  
Now look at the normal distribution. We have less chance of extreme values with the normal distribution. Our 5% alpha level cuts off at a CV of 2.

Back to the original question “Why Do Critical Values Decrease While DF Increases?” Here’s the short answer:

*Degrees of freedom are related to sample size (n-1). If the df increases, it also stands that the sample size is increasing; the graph of the t-distribution will have skinnier tails, pushing the critical value towards the mean.*

## **Raw Score**

A raw score is simply unaltered data from a test or observation. It is recorded in its original form by a researcher before being subjected to any statistical analysis. For instance, if a participant is given a set of ten questions and answers seven right, their raw score might be 7.

Typically raw scores are converted to standard scores or percentiles before being used in comparisons.

## **Probability Distribution**

A probability distribution tells you what the probability of an event happening is. Probability distributions can show **simple events**, like tossing a coin or picking a card. They can also show much more **complex events**, like the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/) of a certain drug successfully treating cancer.

There are many different types of probability distributions in statistics including:

* Basic probability distributions which can be shown on a [probability distribution table.](https://www.statisticshowto.datasciencecentral.com/what-is-a-probability-distribution-table/)
* [Binomial distributions,](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/) which have “Successes” and “Failures.”
* [Normal distributions,](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/) sometimes called a Bell Curve.

The sum of all the probabilities in a probability distribution is always 100% (or 1 as a decimal).

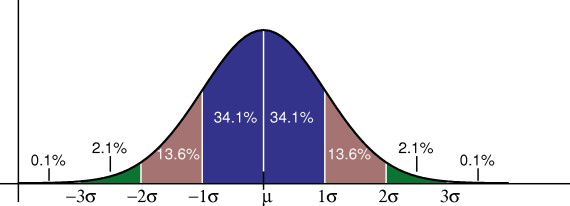
### **Ways of Displaying Probability Distributions**

Probability distributions can be shown in **tables**and **graphs** or they can also be described by a **formula**. For example, the [binomial formula](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/) is used to calculate binomial probabilities.

The following table shows the probability distribution of a tomato packing plant receiving rotten tomatoes. Note that if you add all of the probabilities in the second row, they add up to 1 (.95 + .02 +.02 +0.01 = 1).

[probability distribution toms](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2009/08/probabilitydistributiontoms2.bmp)

The following graph shows a [standard normal distribution](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/#SNM), which is probably the most **widely used probability distribution**. The standard normal distribution is also known as the “[bell curve](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/).” Lots of natural phenomenon fit the bell curve, including heights, weights and IQ scores. The normal curve is a [continuous probability distribution](https://www.statisticshowto.datasciencecentral.com/continuous-probability-distribution/), so instead of adding up individual probabilities under the curve we say that the total area under the curve is 1.

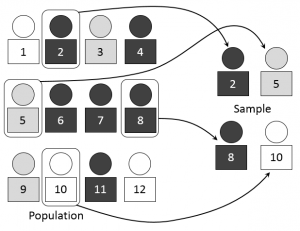
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/02/standard-normal-distribution.jpg)

*In a normal distribution, the percentages of scores you can expect to find for any standard deviations from the*[*mean*](https://www.statisticshowto.datasciencecentral.com/mean/)*are the same.*

**Randomization**

Randomization in an experiment is where you **choose your experimental participants randomly**. For example, you might use [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/), where participants names are drawn randomly from a pool where everyone has an even probability of being chosen. You can also assign treatments randomly to participants, by assigning random numbers from a random number table.

If you use randomization in your experiments, you guard against [**bias**](https://www.statisticshowto.datasciencecentral.com/what-is-bias/). For example,  [selection bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/#SelectionB) (where some groups are underrepresented) is eliminated and [**accidental bias**](https://www.statisticshowto.datasciencecentral.com/accidental-bias/) (where chance imbalances happen) is minimized. You can also run a variety of statistical tests on your data (to test your hypotheses) if your sample is random.



## *Simple random sampling of a sample “n” of 3 from a population “N” of 12.*

**Bernoulli Sampling**

Bernoulli sampling is an **equal probability, without replacement sampling design**. In this method, independent [Bernoulli trials](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/#trial)on [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)members determines which members become part of a [sample](https://www.statisticshowto.datasciencecentral.com/sample/). All members have an equal chance of being part of the [sample](https://www.statisticshowto.datasciencecentral.com/sample/). The sample sizes in Bernoulli sampling are not fixed, because each member is considered separately for the sample

**Example of Bernoulli Sampling**: A researcher has a list of 1,000 candidates for a clinical trials. He wants to get an overview of the candidates and so decides to take a Bernoulli sample to narrow the field. For each candidate, he tosses a die: if it’s a 1, the candidate goes into a pile for further analysis. If it’s any other number, it goes into another pile that isn’t looked at. The EV for the sample size is 1/6 \* 1,000 = 167.

An **advantage**to Bernoulli sampling is that it is one of the simplest types of [sampling methods](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/sampling-in-statistics/#diff). One **disadvantage**is that it’s not known how large the sample is at the outset.



**The Bernoulli Distribution**

A [**Bernoulli Distribution**](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/) is the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/) an experiment produces a particular outcome. It is a binomial distribution with a single event (n = 1).

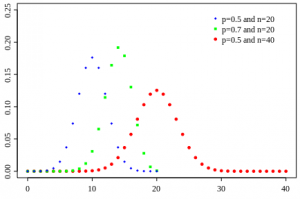
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/10/dice-probability.jpg)

*A die roll can have a Bernoulli distribution.*

There are **two**[**variables**](http://www3.nd.edu/~agervais/documents/Math%20Review.pdf)in a Bernoulli Distribution: n and p.

* “n” represents how many times an experiment is repeated. In a Bernoulli, n = 1.
* “p” is the probability of a specific outcome happening. For example, rolling a die to get a six gives a probability of 1/6. The Bernoulli Distribution for a die landing on an odd number would be p= 1/2.

The Bernoulli and binomial distribution are often confused with each other. However, the difference between the two is slim enough for both to be used interchangeably. Technically, the Bernoulli distribution is the Binomial distribution with n=1.

[](https://www.statisticshowto.datasciencecentral.com/what-is-a-bernoulli-distribution)

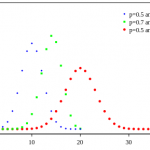
*A Bernoulli Distribution represents a Bernoulli trial where there is either (S)uccess or (F)ailure.*

**Bernoulli Trial**

A Bernoulli distribution is a **Bernoulli trial**. Each Bernoulli trial has a single outcome, chosen from S, which stands for success, or F, which stands for failure. For example, you might try to find a parking space. You are either going to be successful, or you are going to fail. Many real-life situations can be simplified to either success, or failure, which can be represented by Bernoulli Distributions.

**Binomial Distribution**

 A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has **two possible outcomes** (the prefix “[bi](http://membean.com/wrotds/bi-twice)” means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/07/Binomial_distribution_pmf.svg_.png)

*A Binomial Distribution shows either (S)uccess or (F)ailure.*

The first variable in the binomial formula, n, stands for the number of times the experiment runs. The second variable, p, represents the probability of one specific outcome. For example, let’s suppose you wanted to know the probability of getting a 1 on a die roll. if you were to roll a die 20 times, the probability of rolling a one on any throw is 1/6. Roll twenty times and you have a binomial distribution of (n=20, p=1/6). SUCCESS would be “roll a one” and FAILURE would be “roll anything else.” If the outcome in question was the probability of the die landing on an even number, the binomial distribution would then become (n=20, p=1/2). That’s because your probability of throwing an even number is one half.

## **Criteria**

Binomial distributions must also meet the following three criteria:

1. **The number of observations or trials is fixed.** In other words, you can only figure out the [probability](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/probability-main-index/) of something happening if you do it a certain number of times. This is common sense—if you toss a coin once, your probability of getting a tails is 50%. If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.
2. **Each observation or trial is** [independent](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/dependent-events-independent/#or). In other words, none of your trials have an effect on the probability of the next trial.
3. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.

Once you know that your distribution is binomial, you can apply the **binomial distribution formula**to calculate the probability.

## **Bernoulli Distribution**

The binomial distribution is closely related to the [Bernoulli distribution](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/). According to Washington State University, “If each Bernoulli trial is independent, then the number of successes in Bernoulli trails has a binomial Distribution. On the other hand, the Bernoulli distribution is the Binomial distribution with n=1.”

A [Bernouilli distribution](https://www.statisticshowto.datasciencecentral.com/bernoulli-sampling/" \t "_blank) is a set of Bernouilli trials. Each Bernouilli trial has one possible outcome, chosen from S, success, or F, failure. In each trial, the probability of success, P(S)=p, is the same. The probability of failure is just 1 minus the probability of success: P(F) = 1-p. (Remember that “1” is the total probability of an event occurring…probability is always between zero and 1). Finally, all Bernouilli trials are independent from each other and the probability of success doesn’t change from trial to trial, even if you have information about the other trials’ outcomes.

## **Real Life Examples**

Many instances of binomial distributions can be found in real life. For example, if a new drug is introduced to cure a disease, it either cures the disease (it’s successful) or it doesn’t cure the disease (it’s a failure). If you purchase a lottery ticket, you’re either going to win money, or you aren’t. Basically, anything you can think of that can only be a success or a failure can be represented by a binomial distribution.

# Cluster Sampling in Statistics

Cluster sampling is used in statistics when **natural groups** are present in a population. The whole population is subdivided into clusters, or groups, and random samples are then collected from each group.

# Use

Cluster sampling is typically used in market research. It’s used when a researcher **can’t get information about the population as a whole**, but they can get information about the clusters. For example, a researcher may be interested in data about [city taxes in Florida](http://www.sale-tax.com/Florida). The researcher would compile data from selected cities and compile them to get a picture about the state. The individual cities would be the clusters in this case. Cluster sampling is often **more economical** or **more practical** than stratified sampling or [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/).

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/money.jpg)

*“Cost efficiency” is the #1 reason that researchers like this form of sampling.*

# Requirements

1. Cluster elements should be as heterogenous as possible. In other words, the population should contain distinct subpopulations of different types.
2. Each cluster should be a small representation of the entire population.
3. Each cluster should be [mutually exclusive](https://www.statisticshowto.datasciencecentral.com/mutually-exclusive-event/). In other words, it should be impossible for each cluster to occur together. In the city tax example, it would be impossible for Miami city taxes and Jacksonville city taxes to occur together, so it fits the requirements for mutual exclusivity.

# Types

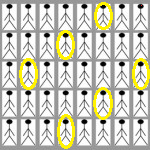
* **Single-stage cluster sampling**: all the elements in each selected cluster are used.
* **Two-stage cluster sampling**: where a random sampling technique is applied to the selected clusters. For example, once you’ve decided on your clusters, you could use [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) to select your sample.

# Difference Between Cluster Sampling and Stratified Sampling

For a [stratified random sample](https://www.statisticshowto.datasciencecentral.com/stratified-random-sample/), a population is divided into stratum, or sub-populations, before sampling. At first glance, the two techniques seem very similar. However, in cluster sampling the actual **cluster is the**[**sampling unit**](https://www.statisticshowto.datasciencecentral.com/sampling-unit/); in stratified sampling, analysis is done on **elements within each strata**. In cluster sampling, a researcher will only study selected clusters; with stratified sampling, a random sample is drawn from each strata.

# Systematic Sampling

When you’re sampling from a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/), you want to make sure you’re getting a fair representation of that population. Otherwise, your statistics will be [biased](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)or [skewed](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/) and perhaps meaningless. One way to get a fair and [random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) is to assign a number to every population member and then choose the nth member from that population. For example, you could choose every 10th member, or every 100th member.**This method of choosing the nth member is called systematic sampling**.



*A systematic sample where every 6th person is chosen (highlighted in yellow).*

Systematic sampling is quick and convenient when you have a complete list of the members of your population (for example, [this one of the members of Congress](https://www.congress.gov/members)). However, if there’s some kind of pattern to the original list, then [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)may creep in to your statistics. For example, if a list of people is ordered as MFMFMFMF, then choosing every 10th number will give you a sample consisting entirely of females. How to perform systematic sampling without this type of [sampling bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)? You could randomly shuffle the list before choosing the nth item or you could use [repeated systematic sampling](https://www.statisticshowto.datasciencecentral.com/systematic-sampling/#repeated), where you take several small samples from the same population. It’s used if you aren’t sure you have a completely random list and you want to avoid sample bias.

# How to Perform Systematic Sampling: Steps

Step 1: **Assign a number to every element in your population**. For this simple example, let’s say you have a population of 100 people, so you’ll assign the numbers 1 to 100 to the group.

Step 2: **Decide how large your sample size should be. See:**[Sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) (how to find one). For this example, let’s say you need a sample of 10 people.

Step 3: **Divide the population by your sample size**. For this example, your population is 100 and your sample size is 10, so:  
100 / 10 = 10  
This is your “nth” sampling digit (i.e. you’ll choose every 10th item)

1 2 3 4 5 6 7 8 9 10  
11 12 13 14 15 16 17 18 19 20  
21 22 23 24 25 26 27 28 29 30  
31 32 33 34 35 36 37 38 39 40  
41 42 43 44 45 46 47 48 49 50  
51 52 53 54 55 56 57 58 59 60  
61 62 63 64 65 66 67 68 69 70  
71 72 73 74 75 76 77 78 79 80  
81 82 83 84 85 86 87 88 89 90  
91 92 93 94 95 96 97 98 99 100

That’s how to perform systematic sampling!

# Repeated Systematic Sampling

Step 1: **Assign a number to every element in your population**.

Step 2: **Decide how large your sample size should be. See:**[Sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) (How to find one).

Step 3: **Divide the population by your sample size**. For example, if your population is 100 and your sample size is 10, then:

100 / 10 = 10  
This is your “nth” sampling digit (i.e. you’ll choose every 10th item)

# Sample from Group 1

Step 4: **Use the sampling digit from Step 3 up to a certain point**. This is usually a judgment call; Exactly where you stop is usually quite arbitrary. The goal is to divide your population into parts. For this example, we’ll sample up to 50; By stopping at 50 we are splitting the entire group into two sections.

First, you’ll sample from the first half of the group (in Step 5, you’ll sample from the remainder).

1 2 3 4 5 6 7 8 9 10  
11 12 13 14 15 16 17 18 19 20  
21 22 23 24 25 26 27 28 29 30  
31 32 33 34 35 36 37 38 39 40  
41 42 43 44 45 46 47 48 49 50  
51 52 53 54 55 56 57 58 59 60  
61 62 63 64 65 66 67 68 69 70  
71 72 73 74 75 76 77 78 79 80  
81 82 83 84 85 86 87 88 89 90  
91 92 93 94 95 96 97 98 99 100

# Sample from Group 2

Step 5: **Switch to a different starting point and then continue sampling with the nth digit**. Again, this is usually a judgment call. For this example, we’ll switch from 50 to 51.

1 2 3 4 5 6 7 8 9 10  
11 12 13 14 15 16 17 18 19 20  
21 22 23 24 25 26 27 28 29 30  
31 32 33 34 35 36 37 38 39 40  
41 42 43 44 45 46 47 48 49 50  
51 52 53 54 55 56 57 58 59 60  
61 62 63 64 65 66 67 68 69 70  
71 72 73 74 75 76 77 78 79 80  
81 82 83 84 85 86 87 88 89 90  
91 92 93 94 95 96 97 98 99 100

Note that we only have 9 in our sample (we wanted 10), so **return to the beginning of the list**and continue:  
1 2 3 4 5 6 7 8 9 10

That’s How to Perform Repeated Systematic Sampling!

# Simple Random Sample

# Random Sample

A random sample is a [sample](https://www.statisticshowto.datasciencecentral.com/sample/)that is chosen randomly. It could be more accurately called a randomly **chosen**sample. Random samples are used to avoid [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)and other unwanted effects. Of course, it isn’t quite as simple as it seems: choosing a random sample isn’t as simple as just picking 100 people from 10,000 people. You have to be sure that your random sample is truly random!

Note that the word “random” in random sample doesn’t exactly fit the dictionary definition of the word. If you Google “define:random” then you’ll read that it means:

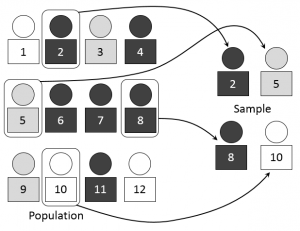
*made, done, happening, or chosen without method or conscious decision.  
“a random sample of 100 households”*

It isn’t true that a random sample is chosen “without method of conscious decision.” Simple random sampling is one way to choose a random sample.

# Simple Random Sample

A simple random sample is often mentioned in[elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/) classes, but it’s actually one of the least used techniques. In theory, it’s easy to understand. However, in practice it’s tough to perform.

Technically, a simple random sample is a set of n objects in a [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/)of N objects where all possible samples are equally likely to happen. Here’s a basic example of how to get a simple random sample: put 100 numbered bingo balls into a bowl (this is the [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/) N). Select 10 balls from the bowl without looking (this is your [sample](https://www.statisticshowto.datasciencecentral.com/sample/)n). Note that it’s important not to look as you could (unknowingly) [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/)the sample. While the “lottery bowl” method can work fine for smaller populations, in reality you’ll be dealing with much larger populations.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/Simple_random_sampling.png)

*Simple random sampling of a sample “n” of 3 from a population “N” of 12. Image: Dan Kernler |Wikimedia Commons*

Imagine the people illustrated in the image above are game pieces. Place the 12 game pieces in a bowl and (again, without looking) choose 3. This is simple random sampling.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/randomSampling.jpg)

*Image: CSUS.edu*

A simple random sample is chosen in such a way that every set of individuals has an equal chance to be in the selected sample. It sounds easy, but SRS is often difficult to employ in surveys or experiments. In addition, it’s very easy for [bias](https://www.statisticshowto.datasciencecentral.com/what-is-bias/) to creep into samples obtained with simple random sampling. Sometimes it’s impossible (either financially or time-wise) to get a realistic [sampling frame](http://srmo.sagepub.com/view/the-sage-encyclopedia-of-social-science-research-methods/n884.xml) (the population from which the sample is to be chosen). For example, if you wanted to study **all** the adults in the U.S. who had high cholesterol, the list would be practically impossible to get unless you surveyed every person in the country. Therefore other sampling methods would probably be better suited to that particular experiment.  
The simplest example of SRS would be working with things like dice or cards — rolling the die or dealing cards from a deck can give you a simple random sample. But in real life you’re usually dealing with people, not cards, and that can be a challenge.

# How to Perform Simple Random Sampling: Example

A larger population might be “All people who have had [strokes](http://www.cdc.gov/stroke/)in the United States.” That list of participants would be extremely hard to obtain. Where would you get such a list in the first place? You could contact individual hospitals (of which there are thousands and thousands…) and ask for a list of patients (would they even supply you with that information? If you could somehow obtain this list then you will end up with a list of 800,000 people which you then have to put into a “bowl” of some sort and choose random people for your sample. This type of situation is the type of real-life situation you’ll come across and is what makes getting a simple random sample so hard to undertake.

**Sample question:** Outline the steps for obtaining a simple random sample for outcomes of strokes in U.S. trauma hospitals.

Step 1: **Make a list**of all the trauma hospitals in the U.S. (there are several hundred: the CDC keeps a list).

Step 2: **Assign a sequential number**to each trauma center (1,2,3…n). This is your [sampling frame](https://www.statisticshowto.datasciencecentral.com/sampling-frame/) (the list from which you draw your simple random sample).

Step 3: **Figure out what your sample size is going to be**. See: ([Sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/)) (how to find one).

Step 4: **Use a**[**random number generator**](https://www.statisticshowto.datasciencecentral.com/random-seed-definition/#excel)to select the [sample](https://www.statisticshowto.datasciencecentral.com/sample/), using your sampling frame (population size) from Step 2 and your sample size from Step 3. For example, if your sample size is 50 and your population is 500, generate 50 random numbers between 1 and 500.

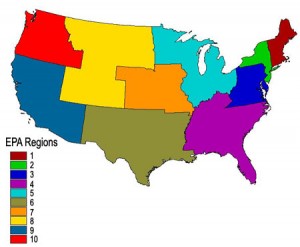
**Warning**: If you compromise (say, by not including ALL trauma centers in your sampling frame), it could open your results to bias.

# Simple Random Sample vs. Random Sample

A simple random sample is similar to a random sample. The difference between the two is that with a simple random sample, each object in the population has **an equal chance of being chosen**. With random sampling, each object does not necessarily have an equal chance of being chosen. Unequal probability sampling isn’t usually addressed in basic statistics courses, but if you’re interested in an example of when it might be used, read [this article.](https://onlinecourses.science.psu.edu/stat506/node/14)

# Stratified Random Sampling

Stratified random sampling is used when your population is divided into strata (characteristics like male and female or education level), and you want to include the stratum when taking your [sample](https://www.statisticshowto.datasciencecentral.com/sample/). The [stratum](https://www.statisticshowto.datasciencecentral.com/stratum/)may be already defined (like [census](https://www.statisticshowto.datasciencecentral.com/what-is-a-census/)data) or you might make the stratum yourself to fit the purposes of your research. Stratified random sampling is **very similar to**[**random sampling**](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/). However, these samples are more difficult to create as **you must have detailed information about what categories** your population falls into.

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/02/stratified-random-sample.jpg)

*The stratum in this map are defined by EPA region. Image:*[*USGS*](http://landcover.usgs.gov/accuracy/)

# How to Perform Stratified Random Sampling

To perform stratified random sampling, take a[random sample](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/)from within each category or stratum. Let’s say you have a population divided into the following strata:

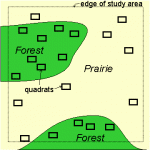
* Category 1: Low socioeconomic status — 39 percent
* Category 2: Middle class — 38 percent
* Category 3: Upper income — 23 percent

To get the stratified random sample, you would randomly sample the categories so that your eventual sample size has 39 percent of participants taken from category 1, 38 percent from category 2 and 23 percent from category 3. What you end up with is a mini representation of your population. According to [University of California at Davis](http://psychology.ucdavis.edu/), the following steps should be taken to obtain the stratified sample:

1. Name the target [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/).
2. Name the categories (stratum) in the population.
3. Figure out what [sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) you need.
4. List all of the cases within each stratum.
5. Make a [decision rule](https://www.statisticshowto.datasciencecentral.com/decision-rule/)to select cases (for example, you might select the items using the largest set of random numbers).
6. Assign a random number to each case.
7. Sort each case by random number.
8. Follow your decision rule (#5 above) to choose your participants.

Stratified random sampling for larger data sets is usually performed using statistical software. For example, [click here](http://www.ats.ucla.edu/stat/sas/faq/statified_samples_v9.htm) for the procedures in SAS.

# How to Get a Stratified Random Sample: Example

[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/12/stratified-random-sampling.gif)

*Stratified random sampling is useful when you can subdivide areas. Image: Oregon State*

“Stratified” means “in layers,” so in order to get a stratified random sample you first need to make the layers. What layers you have depends on characteristics of your [population](https://www.statisticshowto.datasciencecentral.com/what-is-a-population/). For example, if you are surveying U.S. residents about their plans for retirement, you might want your layers to represent different age groups. The sample size for each strata (layer) is proportional to the size of the layer:

*Sample size of the strata = size of entire sample / population size \* layer size.*

# How to Get a Stratified Random Sample: Steps

**Sample question:** You work for a small company of 1,000 people and want to find out how they are saving for retirement. Use stratified random sampling to obtain your [sample](https://www.statisticshowto.datasciencecentral.com/sample/).

Step 1: **Decide how you want to**[**stratify**](http://www.merriam-webster.com/dictionary/stratify)**(divide up) your population.** For example, people in their twenties might have different saving strategies than people in their fifties.

Step 2: **Make a table representing your**[**strata**](https://www.statisticshowto.datasciencecentral.com/stratum/). The following table shows age groups and how many people in the population are in that strata:

|  |  |
| --- | --- |
| Age | Total Number of People in Strata |
| 20-29 | 160 |
| 30-39 | 220 |
| 40-49 | 240 |
| 50-59 | 200 |
| 60+ | 180 |

Step 3: **Decide on your sample size.** If you don’t know how to find a sample size, see: [Sample size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/) (how to find one). For this example, we’ll assume your sample size is 50.

Step 4: **Use the stratified sample formula**(Sample size of the strata = size of entire sample / population size \* layer size) to calculate the proportion of people from each group:

|  |  |  |
| --- | --- | --- |
| Age | Number of People in Strata | Number of People in Sample |
| 20-29 | 160 | 50/1000 \* 160 = 8 |
| 30-39 | 220 | 50/1000 \* 220 = 11 |
| 40-49 | 240 | 50/1000 \* 240 = 12 |
| 50-59 | 200 | 50/1000 \* 200 = 10 |
| 60+ | 180 | 50/1000 \* 180 = 9 |

Note that all of the individual results from the stratum add up to your sample size of 50: 8 + 11 + 12 + 10 + 9 = 50

Step 5: **Perform random sampling** (i.e. [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/)) in each stratum to select your survey participants.

That’s how to get a stratified random sample!

**Tip:** Each element in your population should only fit into one stratum. In other words, one person cannot be in more than one group.

References:

1. <https://www.statisticshowto.datasciencecentral.com/statistics-basics/>
2. <https://stattrek.com/tutorials/ap-statistics-tutorial.aspx>