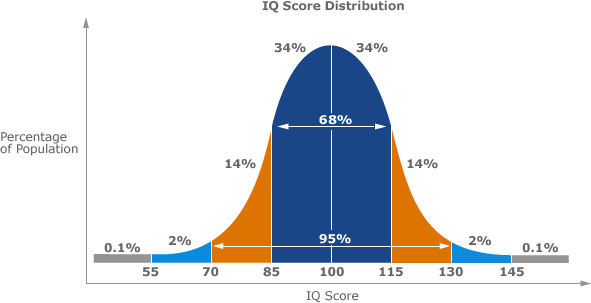
## **Statistics Overview**

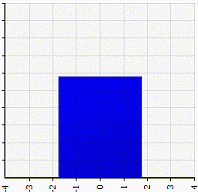
Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a [statistical population](https://en.wikipedia.org/wiki/Statistical_population) 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.*

**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).

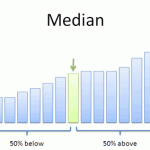
## **Central Tendency Measures**

**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)

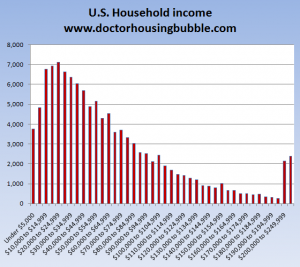
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/median.png)

**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.

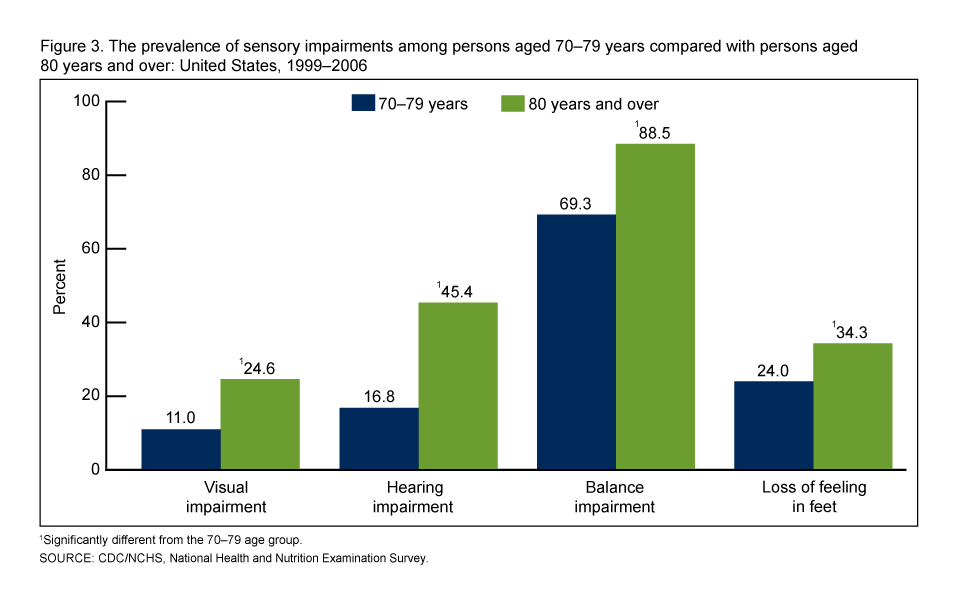
Step 5:**Check the “Labels in first row” check box**if you have titled the column in row 1, otherwise leave the box unchecked.

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.

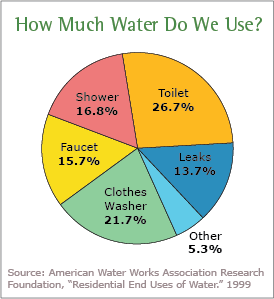
## **4. Descriptive Statistics: Charts, Graphs and Plots**

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/)



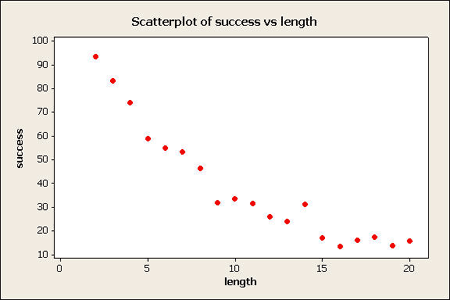
*Grouped bar graph. Image: CDC.*

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.

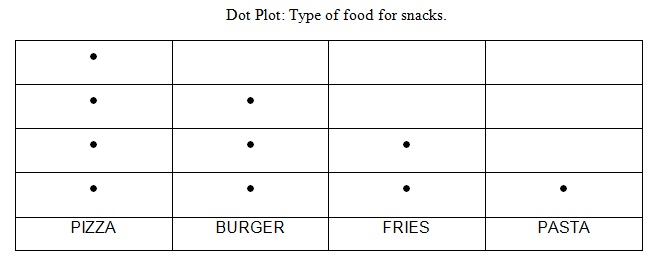


*Pie chart showing water consumption. Image courtesy of EPA.*

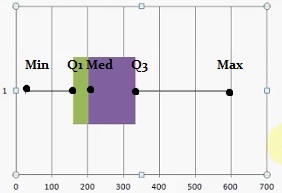
[Scatter plots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/#definition) are a good way to display data points.

 *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):

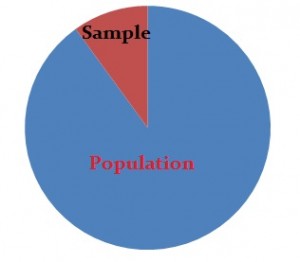
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/10/dot-plot-2.jpg)

*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**



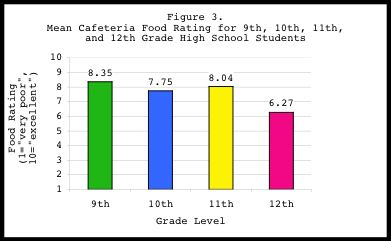
[Descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/) describes data (for example, a chart or graph) and **inferential statistics** allows you to make predictions (“inferences”) from that data. With inferential statistics, you 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.

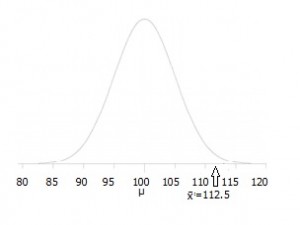
Let’s say you have some sample data about a potential new cancer drug. You could use descriptive statistics to describe your sample, including:

* 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.

## **[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)

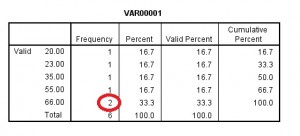
Step 2: **Type your data into the worksheet.** You can type the data into one column or multiple columns if you have multiple data sets. For this example, type 20, 23, 35, 66, 55, 66 into column 1. Do not leave spaces between the data (i.e. don’t leave any empty rows).

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.

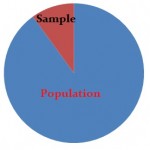
If you scroll down, the frequency table will also show you the mode. The mode is defined in statistics as the number with the highest frequency (for this sample data set, the number appearing the most is 66, with two results in the frequency column).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/spss-mean-6.jpg)

**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**

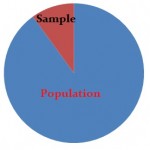
The most common techniques you’ll likely meet in [elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/) or AP statistics include [taking a sample with and without replacement](https://www.statisticshowto.datasciencecentral.com/sampling-with-replacement-without/).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

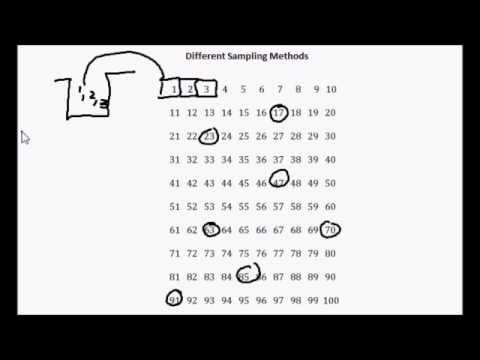
* [**Bernoulli samples**](https://www.statisticshowto.datasciencecentral.com/bernoulli-sampling/) have independent [Bernoulli trials](https://www.statisticshowto.datasciencecentral.com/bernoulli-distribution/#trial) on population elements. The trials decide whether the element becomes part of the [sample](https://www.statisticshowto.datasciencecentral.com/sample/). 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.

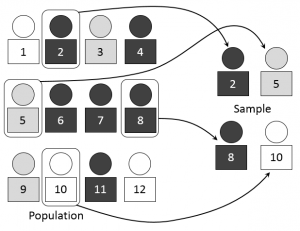
**Another Type of Error**

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](https://www.statisticshowto.datasciencecentral.com/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.

**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).

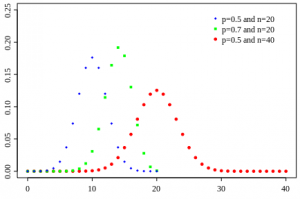
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/10/dice-probability.jpg)

*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)

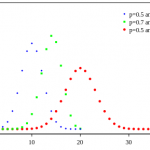
*A Bernoulli Distribution represents a Bernoulli trial where there is either (S)uccess or (F)ailure.*

**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

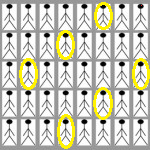
* **Single-stage cluster sampling**: all the elements in each selected cluster are used.
* **Two-stage cluster sampling**: where a random sampling technique is applied to the selected clusters. For example, once you’ve decided on your clusters, you could use [simple random sampling](https://www.statisticshowto.datasciencecentral.com/simple-random-sample/) 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:

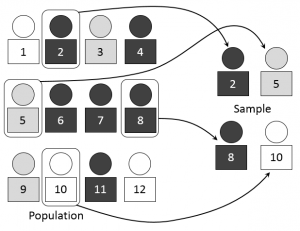
*made, done, happening, or chosen without method or conscious decision.  
“a random sample of 100 households”*

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.

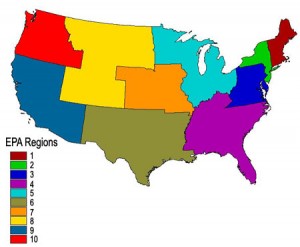
**Warning**: If you compromise (say, by not including ALL trauma centers in your sampling frame), it could open your results to bias.

# 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:

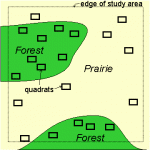
* Category 1: Low socioeconomic status — 39 percent
* Category 2: Middle class — 38 percent
* Category 3: Upper income — 23 percent

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>