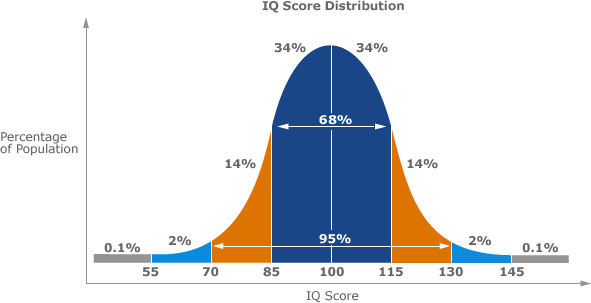
## **Statistics Overview**

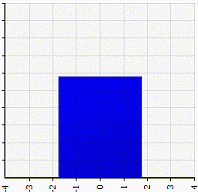
Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a [statistical population](https://en.wikipedia.org/wiki/Statistical_population)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

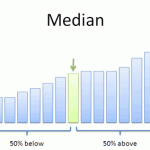
## **Central Tendency Measures**

**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)

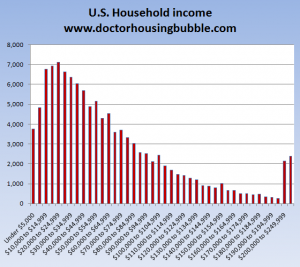
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/median.png)

**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.

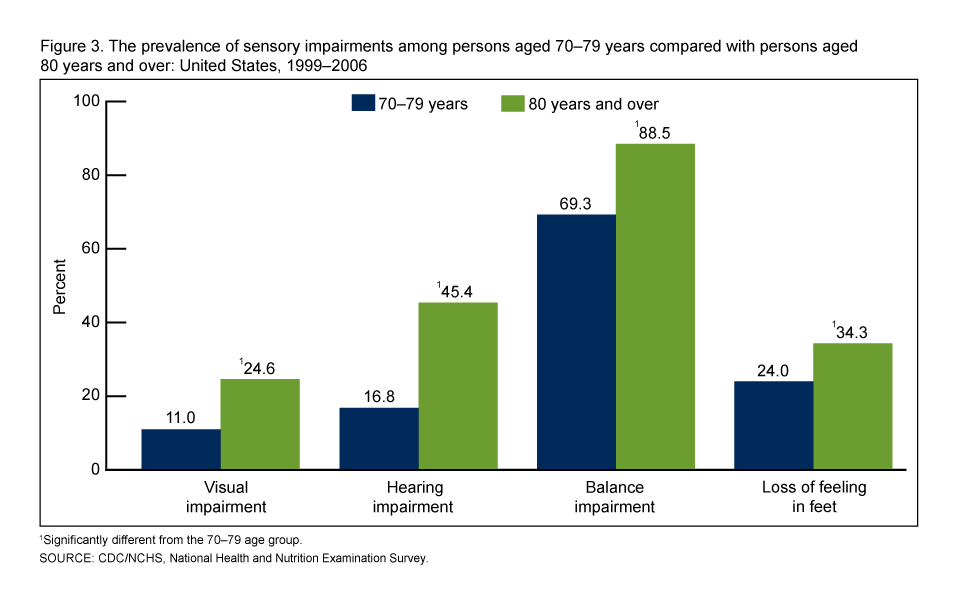
Step 5:**Check the “Labels in first row” check box**if you have titled the column in row 1, otherwise leave the box unchecked.

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.

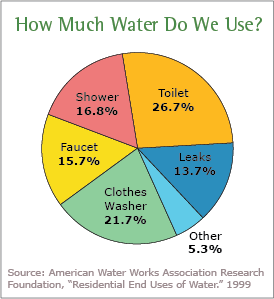
## **4. Descriptive Statistics: Charts, Graphs and Plots**

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/)



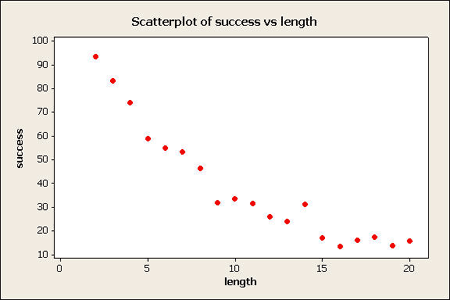
*Grouped bar graph. Image: CDC.*

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.

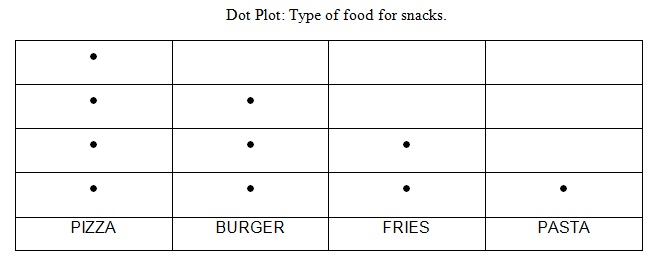


*Pie chart showing water consumption. Image courtesy of EPA.*

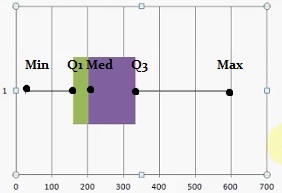
[Scatter plots](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/scatter-plot-chart/#definition) are a good way to display data points.

 *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):

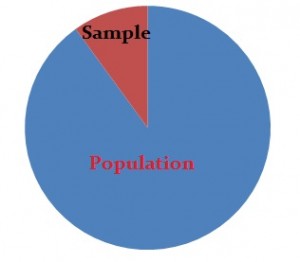
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/10/dot-plot-2.jpg)

*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**



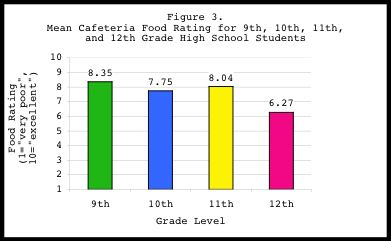
[Descriptive statistics](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/) describes data (for example, a chart or graph) and **inferential statistics** allows you to make predictions (“inferences”) from that data. With inferential statistics, you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

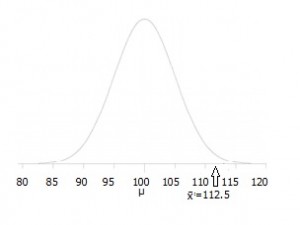
Let’s say you have some sample data about a potential new cancer drug. You could use descriptive statistics to describe your sample, including:

* 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)

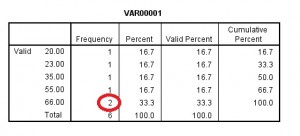
Step 2: **Type your data into the worksheet.** You can type the data into one column or multiple columns if you have multiple data sets. For this example, type 20, 23, 35, 66, 55, 66 into column 1. Do not leave spaces between the data (i.e. don’t leave any empty rows).

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.

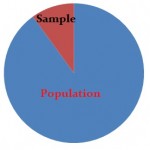
If you scroll down, the frequency table will also show you the mode. The mode is defined in statistics as the number with the highest frequency (for this sample data set, the number appearing the most is 66, with two results in the frequency column).  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/08/spss-mean-6.jpg)

**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.

**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/).

**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)

**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.

References:

1. <https://www.statisticshowto.datasciencecentral.com/statistics-basics/>
2. <https://stattrek.com/tutorials/ap-statistics-tutorial.aspx>