Statistics

# Statistics

It is a field of study concerned with Collecting data, Analyzing Data, Interpreting data, Presenting data and making decisions based on data.

# Why Statistics?

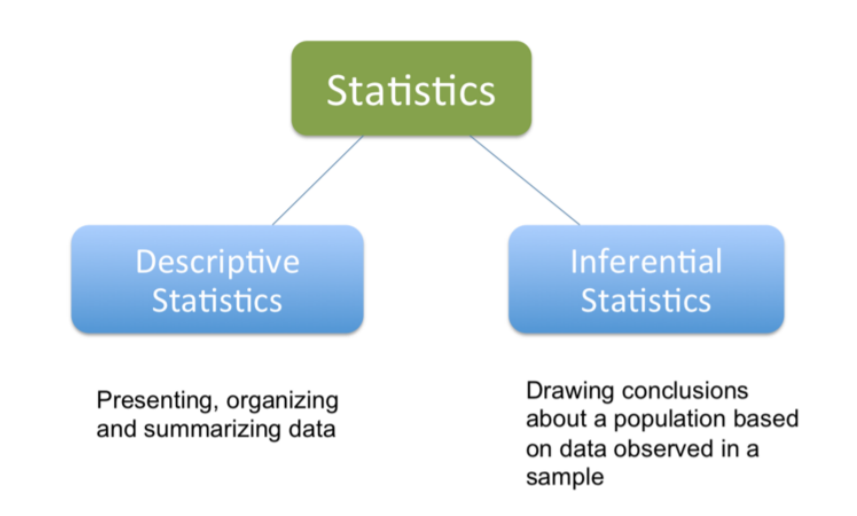
In almost every endeavor of human activity, the scientific method has proven effective for solving problems and improving performance. This approach involves the collection of information or data pertinent to the particular problem or concern. Statisticians play several important roles in these scientific studies.

First, they plan the studies to ensure that the data are collected efficiently and answer the questions relevant to the investigation.

Second, they analyze the data to discover what conclusions can be reached from the study and what issues need further investigation.

In industry, statisticians design and analyze experiments to improve the safety, reliability and performance of all types products.

# Structure of Statistics



## Descriptive Statistics (DS)

Descriptive statistics are numbers that are used to summarize and describe data. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand.

Descriptive statistics is a collection of methods for summarizing data (e.g., mean, median, mode, range, variance, graphs).

## Inferential statistics

Inferential statistics is a collection of methods for using sample data to make conclusions.

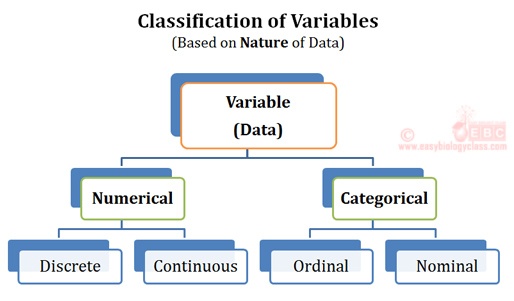
# Variables

A variable is any characteristics, number, or quantity that can be measured or counted.

## Types of Variables

* Independent or Dependent.
* Discrete or Continuous.
* (Qualitative/Numerical) or (Quantitative/Categorical).

## Classifications of Variables



# **Quantitative/Numerical**

A **Quantitative/Numerical** variable is a variable that reflects a notion of **magnitude**, that is, if the values it can take are **numbers**. A Quantitative/Numerical variable represents thus a measure and is numerical.

Quantitative/Numerical variables are divided into two types: **discrete** and **continuous**. The difference is explained in the following two sections.

## Discrete

A  **Discrete** variables are variables for which the values it can take are **countable** and have a **finite number of possibilities**. The values are often (but not always) integers. Here are some examples of discrete variables:

* Number of children per family
* Number of students in a class
* Number of citizens of a country

## Continuous

A  **Continuous** variables are variables for which the values are **not countable** and have an **infinite number of possibilities**. For example:

* Age
* Weight
* Height

# Qualitative/Categorical

**Qualitative Categorical** variables (also referred as categorical variables or [factors](https://www.statsandr.com/blog/data-types-in-r/#factor) in R) are variables that are **not numerical** and which **values fits into categories**.

Qualitative variables are divided into two types: **nominal** and **ordinal**.

## Nominal

A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels. For example, the variable gender is nominal because there is no order in the levels female/male. Eye color is another example of a nominal variable because there is no order among blue, brown or green eyes.

A nominal variable can have between two levels (e.g., do you smoke? Yes/No or what is your gender? Female/Male) and a large number of levels (what is your college major? Each major is a level in that case).

## Ordinal

On the other hand, a **qualitative ordinal** variable is a qualitative variable with an **order implied in the levels**.

For instance, if the severity of road accidents has been measured on a scale such as light, moderate and fatal accidents, this variable is a qualitative ordinal variable because there is a clear order in the levels.

# Summarizing Data

When you want to measure something in the natural world you usually have to take several measurements. This is because things are variable, so you need several results to get an idea of the situation. Once you have these measurements you need to summarize them in some way because sets of raw numbers are not easily interpreted by most people.

There are four key areas to consider when summarizing a set of numbers:

* Centrality – the middle value or average.
* Dispersion – how spread out the values are from the average.
* Replication – how many values there are in the sample.
* Shape – the data distribution, which relates to how “evenly” the values are spread either side of the average.

You need to present the first three summary statistics in order to summarize a set of numbers adequately. There are different measures of centrality and dispersion – the measures you select are based on the the last item, shape (or data distribution).

## Averages

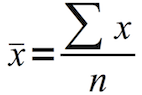
An average is a measure of the middle point of a set of values. This central tendency (centrality) is an important measure and is usually what you are comparing when looking at differences between samples for example.

There are three main kinds of average:

* Mean – the arithmetic mean, the sum of the values divided by the replication.
* Median – the middle value when all the numbers are ranked in order.
* Mode – the most frequent value(s) in a sample.
* Of these three, the mean and the median are most commonly used in statistical analysis. The most appropriate average depends on the shape of the data sample.

## Mean

The arithmetic mean is calculated by adding together the values in the sample. The sum is then divided by the number of items in the sample (the replication).

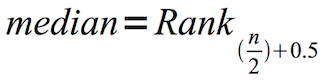


The formula is shown above. The ∑ symbol represents “sum of”. The n represents the replication. The final mean is indicated using an overbar. This shows that the mean is your estimate of the true mean. This is because you usually measure only some of the items in a “population”; this is called a sample. If you measured everything then you would be able to calculate the true mean, which would be indicated by giving it a µ symbol.

The mean should only be used when the shape of the sample is appropriate. When the data are normally distributed the mean is a good summary of the average. If the data are not normally distributed the mean is not a good summary and you should use the median instead.

## Median

The median is the middle value, taken when you arrange your numbers in order (rank). This measure of the average does not depend on the shape of the data. The “formula” for working out the median depends on the ranks of the values, you want a value whose rank is the (n/2)+0.5th like so:



If you have an odd number of values in your sample the median is simply the middle value like so:

4 6 7 8 9

The median is 7 in this case.

When you have an even number of values the middle will fall between two items:

2 3 4 7 8 9

What you do is use a value mid-way between the two items in the middle. In this case mid-way between 4 and 7, which gives 5.5.

The median is a good general choice for an average because it is not dependent on the shape of the data. When the data are normally distributed the mean and the median are coincident (or very close).

## Mode

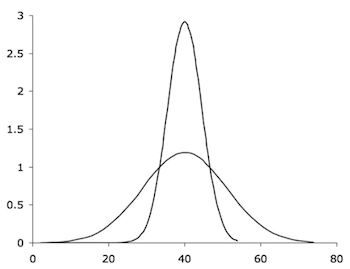
The Mode is the most frequent value in a sample. It is calculated by working out how many there are of each value in your sample. The one with the highest frequency is the mode. It is possible to get tied frequencies, in which case you report both values. The sample is then said to be bimodal. You might get more than two modal values!

The mode is not commonly used in statistical analysis. It tends to be used most often when you have a lot of values, and where you have integer values (although it can be calculated for any sample).

The mode is not dependent on the shape of your sample. Generally speaking you would expect your mode and median to be close, regardless of the sample distribution. If the sample is normally distributed the mode will usually also be close to the mean.

## Dispersion

The dispersion of a sample refers to how spread out the values are around the average. If the values are close to the average, then your sample has low dispersion. If the values are widely scattered about the average your sample has high dispersion.



The example figure shows samples that are normally distributed, that is, they are symmetrical around the average (mean). As far as dispersion goes, the principle is the same regardless of the shape of the data. However, different measures of dispersion will be more appropriate for different data distribution.

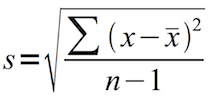
There are various measures of dispersion, such as:

* Standard deviation
* Variance
* Standard Error
* Confidence Interval
* Inter-Quartile Range
* Range

The choice of measurement depends largely on the shape of the data and what you want to focus on. In general, with normally distributed data you use the standard deviation. If the data are not normally distributed, you use the inter-quartile range.

## Standard deviation

The standard deviation is used when the data are normally distributed. You can think of it as a sort of “average deviation” from the mean. The general formula for calculating standard deviation looks like the following:



To work out standard deviation follow these steps:

* Subtract the mean from each value in the sample.
* Square the results from step 1 (this removes negative values).
* Add together the squared differences from step 2.
* Divide the summed squared differences from step 3 by n-1, which is the number of items in the sample (replication) minus one.
* Take the square root of the result from step 4.

The final result is called s, the standard deviation. In most cases you will have taken a sample of values from a larger “population”, so your value of s is your estimate of standard deviation (the sample standard deviation). This is also why you used n-1 as the divisor in the formula. If you measured the entire population you can use n as the divisor. You would then have σ, which is the “true” standard deviation (called the population standard deviation).

In effect the -1 is a compensation factor. As n gets larger and therefore closer to the entire population, subtracting 1 has a smaller and smaller effect on the result. In most statistical analyses you will use sample standard deviation (and so n-1).

## Inter-Quartile range

The inter-quartile range (IQR) is a useful measure of the dispersion of data that are not normally distributed (see shape). You start by working out the median; this effectively splits the data into two chunks, with an equal number of values in each part. For each half you can now work out the value that is half-way between the median and the “end” (the maximum or minimum). This gives you values for the two inter-quartiles. The difference between them is the IQR, which you usually express as a single value.

The IQR essentially “knocks off” the most extreme portions of the data sample, leaving you with a core 50% of your original data. A small IQR denotes a small dispersion and a large IQR a large dispersion.

As a by-product of working out the IQR you’ll usually end up with five values:

* Minimum – the 0th quartile (or 0% quantile).
* Lower quartile – the 1st quartile (or 25% quantile).
* Median – the 2nd quartile (or 50% quantile).
* Upper quartile – the 3rd quartile (or 75% quantile).
* Maximum – the 4th quartile (or 100% quantile).

These 5 values split the data sample into four parts, which is why they are called quartiles. You can calculate the quartiles from the ranks of the data values like so:

* Rank the values in ascending order. Use the mean rank for tied values.
* The median corresponds to the item that has rank 0.5n + 0.5 (where n = replication).
* The lower quartile corresponds to the item that has rank 0.25n + 0.75.
* The upper quartile corresponds to the item that has rank 0.75n + 0.25.
* If you are using Excel you can compute the quartiles using the QUARTILE function.

## Range

The range is simply the difference between the maximum and the minimum values. It is quite a crude measure and not very useful. The inter-quartile range is much more useful, and makes use of the maximum and minimum values in the calculation.

## Replication

This is the simplest of the summary statistics but it is still important. The replication is simply how many items there are in your sample (that is, the number of observations).

The value n, the replication, is used in calculating other summary statistics, such as standard deviation and IQR, but it is also helpful in its own right. You should look at the dispersion and replication together. A certain value for dispersion might be considered “high” if n is small but quite “low” if n is very large.

## Shape

The shape of the data affects the type of summary statistics that best summarize them. The “shape” refers to how the data values are distributed across the range of values in the sample. Generally you expect there to be a “cluster” of values around the average. It is important to know if the values are more or less symmetrically arranged around the average, or if there are more values to one side than the other.

There are two main ways to explore the shape (distribution) of a sample of data values:

* Graphically – using frequency histograms or tally plots draws a picture of the sample shape.
* Shape statistics – such as skewness and kurtosis. These give values to how central the average is and how clustered around the average the data are.

The ultimate goal is to determine what kind of distribution your data forms. If you have normal distribution you have a wide range of options when it comes to data summary and subsequent analysis.

## Types of data distribution

There are many “shapes” of data, commonly encountered ones are:

* Normal (also called Gaussian)
* Poisson
* Binomial

In general, your aim is to work out if you have normal distribution or not. If you do have normal distribution you can use mean and standard deviation for summary. If you do not have normal distribution you need to use median and IQR instead.

The normal distribution (also called Gaussian) has well-explored characteristics and such data are usually described as parametric. If data are not parametric they can be described as skewed or non-parametric.

## Drawing the distribution

There are two main ways to visualize the shape of your data:

* Tally plots
* Histograms

In both cases the idea is to make a frequency plot. The data values are split into frequency classes, usually called bins. You then determine how many data items are in each bin. There is little difference between a tally plot and a histogram, they show the same information but are constructed is slightly different ways.

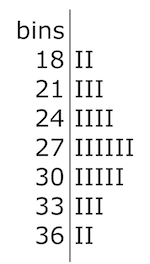
## Tally plots

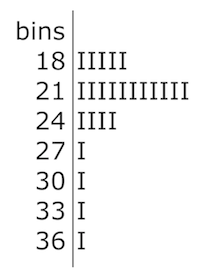
A tally plot is a kind of frequency graph that you can sketch in a notebook. This makes it a very useful tool for times when you haven’t got a computer to hand.

To draw a tally plot follow these steps:

* Determine the size classes (bins), you want around 7 bins.
* Draw a vertical line (axis) and write the values for the bins to the left.
* For each datum, determine which size class it fits into and add a tally mark to the right of the axis, opposite the appropriate bin.

You will now be able to assess the shape of the data sample you’ve got.





The tally plot in the preceding figure shows a normal (parametric) distribution. You can see that the shape is more or less symmetrical around the middle. So here the mean and standard deviation would be good summary values to represent the data. The original dataset was:

17 26 28 27 29 28 25 26 34 32 23 29 24 21 26 31 31 22 26 19 36 23 21 16 30

The first bin, labelled 18, contains values up to 18. There are two in the dataset (17, and 16). The next bin is 21 and therefore contains items that are >18 but not greater than 21 (there are three: 21, 19 and 21).

The following dataset is not normally distributed:

21 36 18 17 16 22 20 19 20 22 25 19 17 21 19 21 31 22 19 19 16 23 21 16 30

These data produce a tally plot like so:

Note that the same bins were used for the second dataset. The range for both samples was 16-36. The data in the second sample are clearly not normally distributed. The tallest size class is not in the middle and there is a long “tail” towards the higher values. For these data the median and inter-quartile range would be appropriate summary statistics.

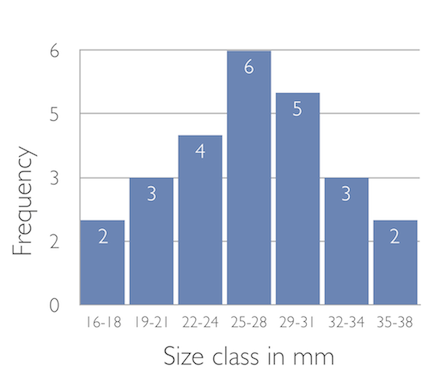
## Histograms

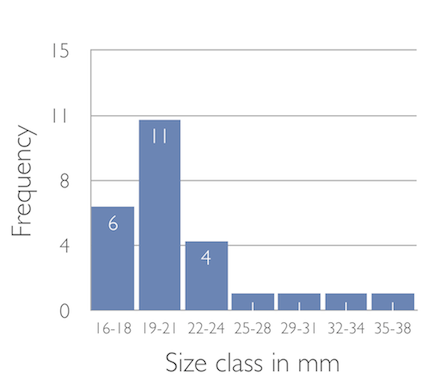
A histogram is like a bar chart. The bars represent the frequency of values in the data sample that correspond to various size classes (bins). Generally the bars are drawn without gaps between them to highlight the fact that the x-axis represents a continuous variable. There is little difference between a tally plot and a histogram but the latter can be produced easily using a computer (you can sketch one in a notebook too).

To make a histogram you follow the same general procedure as for a tally plot but with subtle differences:

* Determine the size classes.
* Work out the frequency for each size class.
* Draw a bar chart using the size classes as the x-axis and the frequencies on the y-axis.

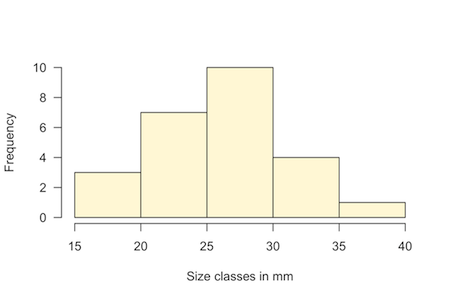
You can draw a histogram by hand or use your spreadsheet. The following histograms were drawn using the same data as for the tally plots in the preceding section. The first histogram shows normally distributed data.





The next histogram shows a non-parametric distribution.

In both these examples the bars are shown with a small gap, more properly the bars should be touching. The x-axis shows the size classes as a range under each bar. You can also show the maximum value for each size class. Ideally your histogram should have the labels at the divisions between size classes like so:



Note that this histogram uses slightly different size classes to the earlier ones.

## Shape statistics

Visualizing the shape of your data samples is usually your main goal. However, it is possible to characterize the shape of a data distribution using shape statistics. There are two, which are used in conjunction with each other:

* Skewness – a measure of how central the average is in the distribution.
* Kurtosis – a measure of how pointy the distribution is ( think of it as how clustered the values are around the middle).

If you are producing a numerical data summary these two values are useful statistics.

## Skewness

The skewness of a sample is a measure of how central the average is in relation to the overall spread of values. The formula to calculate skewness uses the number of items in the sample (the replication, n) and the standard deviation, s.

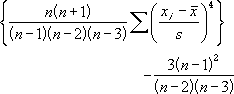
https://www.dataanalytics.org.uk/wp-content/uploads/2019/05/skewness.png

In practice you’ll use a computer to calculate skewness; Excel has a SKEW function that will compute it for you.

A positive value indicates that the average is skewed to the left, that is, there is a long “tail” of more positive values. A negative value indicates the opposite. The larger the value the more skewed the sample is.

## Kurtosis

The kurtosis of a sample is a measure of how pointed the distribution is (see drawing the distribution). It is also a way to think about how clustered the values are around the middle. The formula to calculate kurtosis uses the number of items in the sample (the replication, n) and the standard deviation, s.



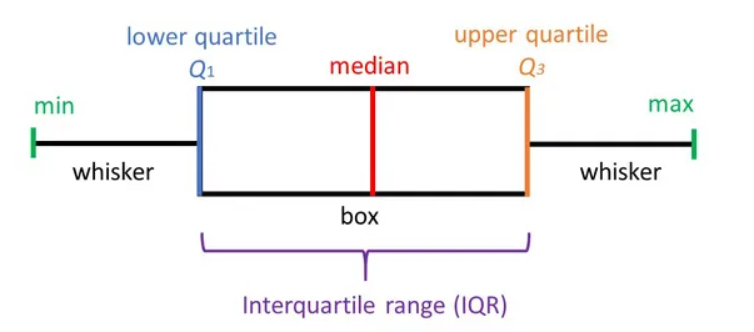
In practice you’ll use a computer to calculate kurtosis; Excel has a KURT function that will compute it for you.

A positive result indicates a pointed distribution, which will probably also have a low dispersion. A negative result indicates a flat distribution, which will probably have high dispersion. The higher the value the more extreme the pointedness or flatness of the distribution.

## Box plot

A box plot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.



Definitions

**Minimum Score**

The lowest score, excluding outliers (shown at the end of the left whisker).

**Lower Quartile**

Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).

**Median**

The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value and half are less.

**Upper Quartile**

Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.

**Maximum Score**

The highest score, excluding outliers (shown at the end of the right whisker).

**Whiskers**

The upper and lower whiskers represent scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores).

**The Interquartile Range (or IQR)**

This is the box plot showing the middle 50% of scores (i.e., the range between the 25th and 75th percentile).