## Statistics for Data Science – Part I: Introduction

In 1970s, John Tukey produced a new definition for statistics; instead of calling it a pure mathematical science, he suggested that deriving hypotheses from data was the future. It was a reform of statistics and announcement of an as-yet unrecognized science. It has been called Data Science for a long time and it is influenced by computer science, mathematics, statistics as well as the applied sciences.

In this series of articles, I will cover the basic parts of statistics which are crucial for a data scientist and I will try to answer the following questions:

- What is statistics?
- Why should I learn statistics?
- How can statistics help me in my profession?

## Definitions and Concepts

For the sake of formality, let us start with the Wikipedia definition for statistics:

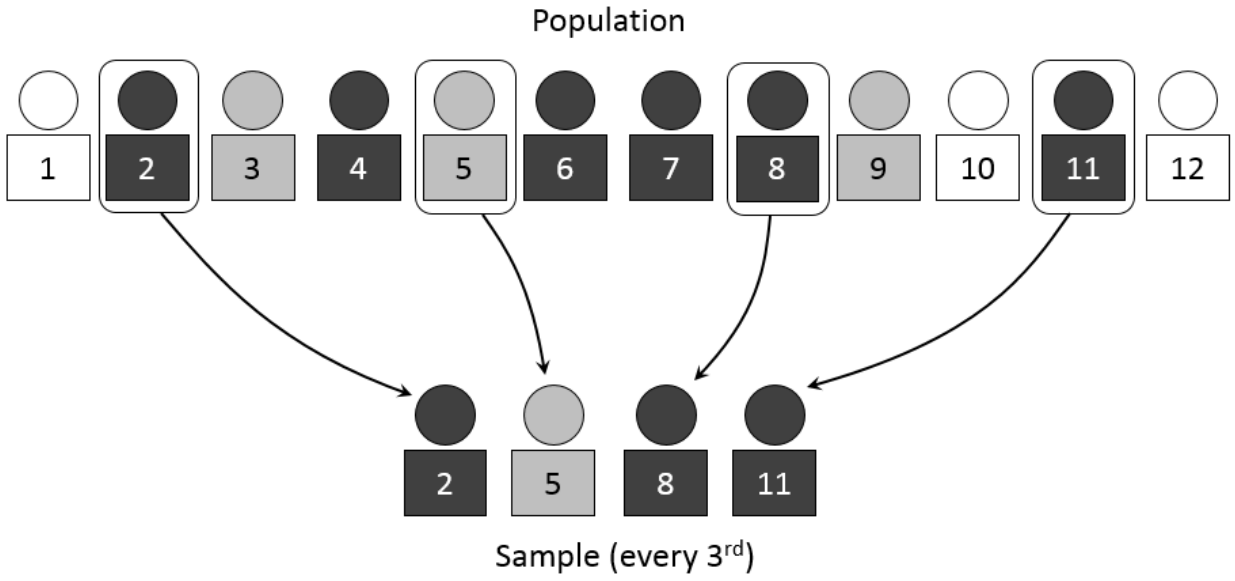
”Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.”

There are two basic types of data sets, namely, a population and a sample.

A **population** is the collection of all outcomes, responses, measurements, or counts that are of interest.

Meanwhile, a **sample** is a subset of a population. Sample data are used to form conclusions about populations.

The figure below shows an example of systematic sampling, where the sample set is generated by picking specimens by following a certain pattern (not randomly).



A **parameter** is a numerical description of a population characteristic.

A **statistic** is a numerical description of a sample characteristic.

Descriptive and Inferential Statistics are defined as the two branches of statistics.

**Descriptive statistics** is the branch that involves the organization, summarization and display of data – data analysis and visualization.

**Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about a population. Probability is widely used in inferential statistics.

Data sets contain two types of data: qualitative and quantitative data.

**Qualitative data** consist of attributes, labels, or nonnumerical entries.

**Quantitative data** consist of numerical measurements or counts.

## Frequency Distributions

A frequency distribution is a table that shows classes or intervals of data entries with a count of the number of entries in each class. The frequency ***f*** of a class is the number of data entries (or how often something happened) in the class.

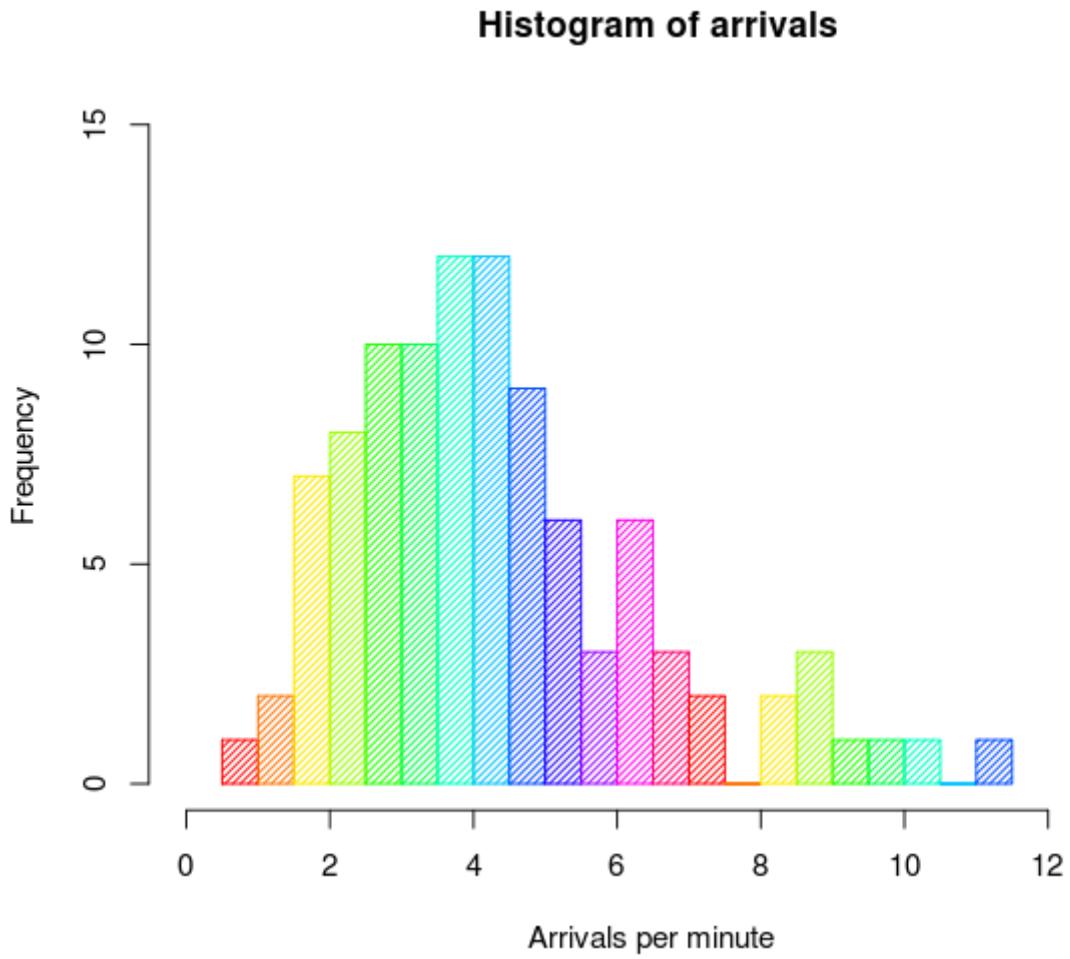
| Weights | No. of Students |
|---|---|
| 31 – 35 | 9 |
| 36 – 40 | 5 |
| 41 – 45 | 14 |
| 46 – 50 | 3 |
| 51 – 55 | 1 |
| 56 – 60 | 2 |
| 61 – 65 | 2 |
| 66 – 70 | 1 |
| 71 – 75 | 1 |
| Total | 38 |

In the frequency distribution shown, there are nine classes. Each class has a lower and an upper-class limit. The difference between the maximum and minimum data entries is called the range.

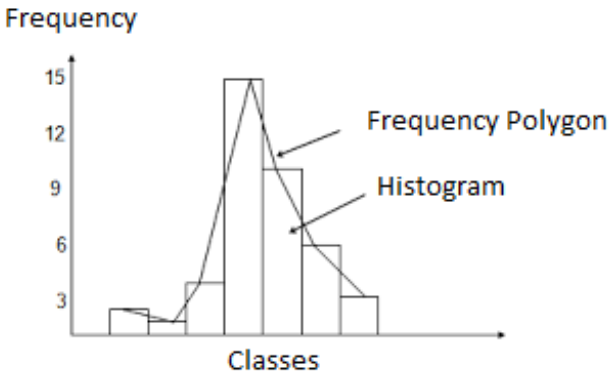
Sometimes, it is easier to identify patterns of a data set by looking at a graph of the frequency distribution. Frequency histogram is a very explanatory tool for the pattern illustration of data.

A **frequency histogram** is a bar graph that represents the frequency distribution of a data set. A histogram has the following properties:

- The horizontal axis includes the data values,
- The vertical axis includes the frequencies of the classes.




A **frequency polygon** shows the frequency distribution in an alternative way. It is a line graph that emphasizes the continuous change in frequencies. The difference is shown below, for the same data set.




## Central Tendency

A measure of **central tendency** is a value that represents a typical, or central entry of a data set.

The three most commonly used measures of central tendency are the mean, the median and the mode.

The **mean** of a data set is the sum of the data entries divided by the number of entries.

Population Mean: $\mu = \frac{\sum x}{N}$ Sample Mean: $\bar{x} = \frac{\sum x}{N}$

where, μ is the population mean and $\bar{x}$ is the sample mean.

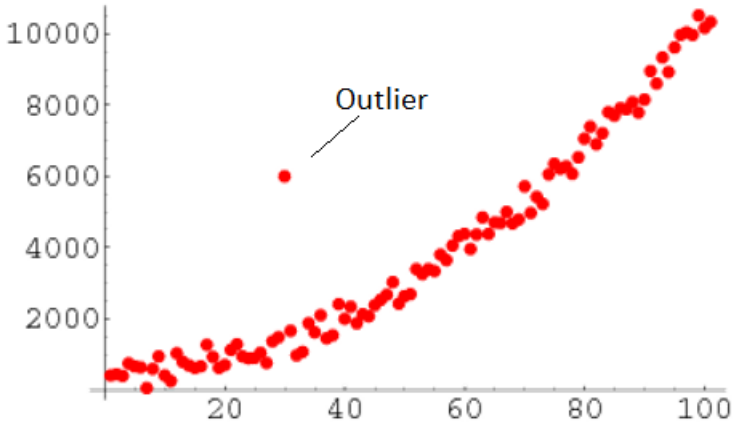
The **median** of a data set is the value that lies in the middle of the data when the data is ordered. It may be considered as the middle value of a data set.

The **mode** of a data set is the data entry that occurs with the greatest frequency. If two entries have the same frequency, each one of them is a mode and the data set is called **bimodal**.

Each of these values describe a typical property of a data set and there are advantages and disadvantages of using each of them. Most often, the mean is a reliable measure because it takes every element of the data set into account – unless there are outliers. An **outlier** is a value which is far removed from the other elements in the data set. A data set can have one or more outliers, causing gaps in an otherwise regular

distribution. The most important result is, conclusions that are drawn from a data set that contains outliers may be flawed.

Boxplots and other visualization techniques are very useful in detecting outliers and we will discuss them later in the article.



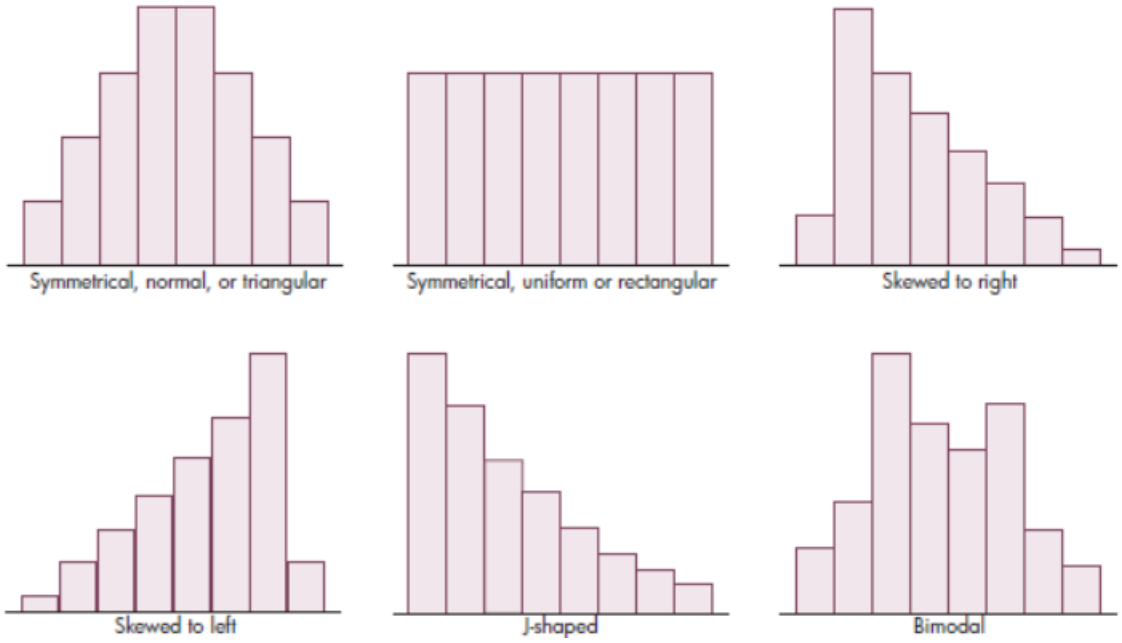

## Shape of Distributions

Graphics may reveal several characteristics of a frequency distribution – shape of the distribution is one of them.

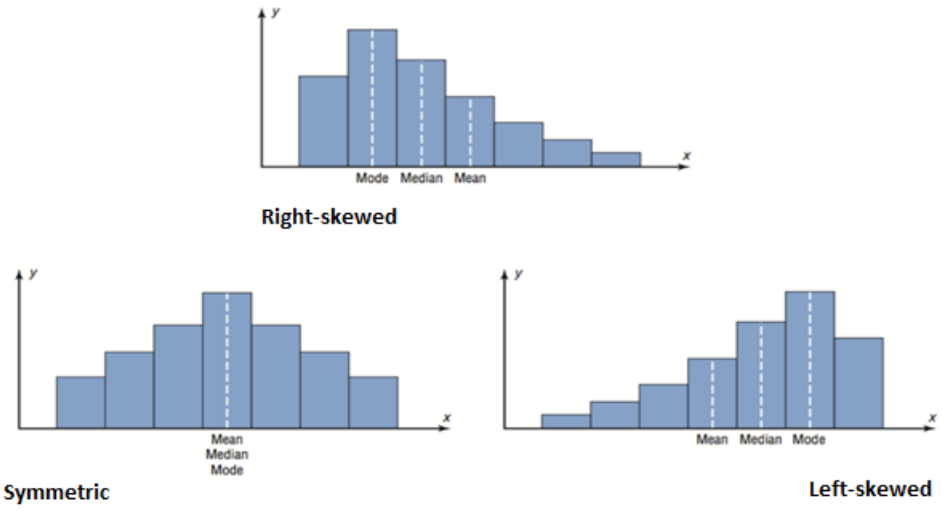
A frequency distribution is **symmetric** if a virtual vertical line through the middle of the distribution slices the shape into identical mirror images.

A frequency distribution is **uniform** when all entries have approximately equal frequencies. A uniform distribution is also symmetric.

A frequency distribution is **skewed** if the tail of the graph elongates more to one side than to the other.




When a distribution is symmetric and unimodal, the mean, median and mode are equal. If a distribution is skewed left, the mean is less than the median and the median is usually less than the mode. If a distribution is skewed right, the mean is greater than the median and the median is usually greater than the mode.



There may be more different shapes of distributions. In some cases, the shape cannot be classified as symmetric, uniform or skewed. A distribution can have several gaps caused by outliers, or clusters of data. Clusters usually occur when several types of data are included in one data set.

The table shows which measure to consider according to the shape of distribution.

| | Mean | Median | Mode |
|---|---|---|---|
| **Categorical Data** | | | **X** |
| **Outliers are Present** | | **X** | **X** |
| **Symmetrical** | **X** | **X** | **X** |
| **Skewed** | | **X** | |

## Measures of Variation

Range is the simplest measure of variation of a data set. The **range** is the difference between the maximum and minimum data entries in the set (for quantitative data).

Using the range to measure the variation has a major drawback: it uses only two entries from the data set and does not give much indication of the spread of observations about the mean. Using all the values in the data set will give a more reliable value.

The **deviation** of an entry *x* in a data set is the difference between the entry and the mean.

$$\text{Deviation of } x = x - \mu$$

The **variance** of a data set of N entries is:

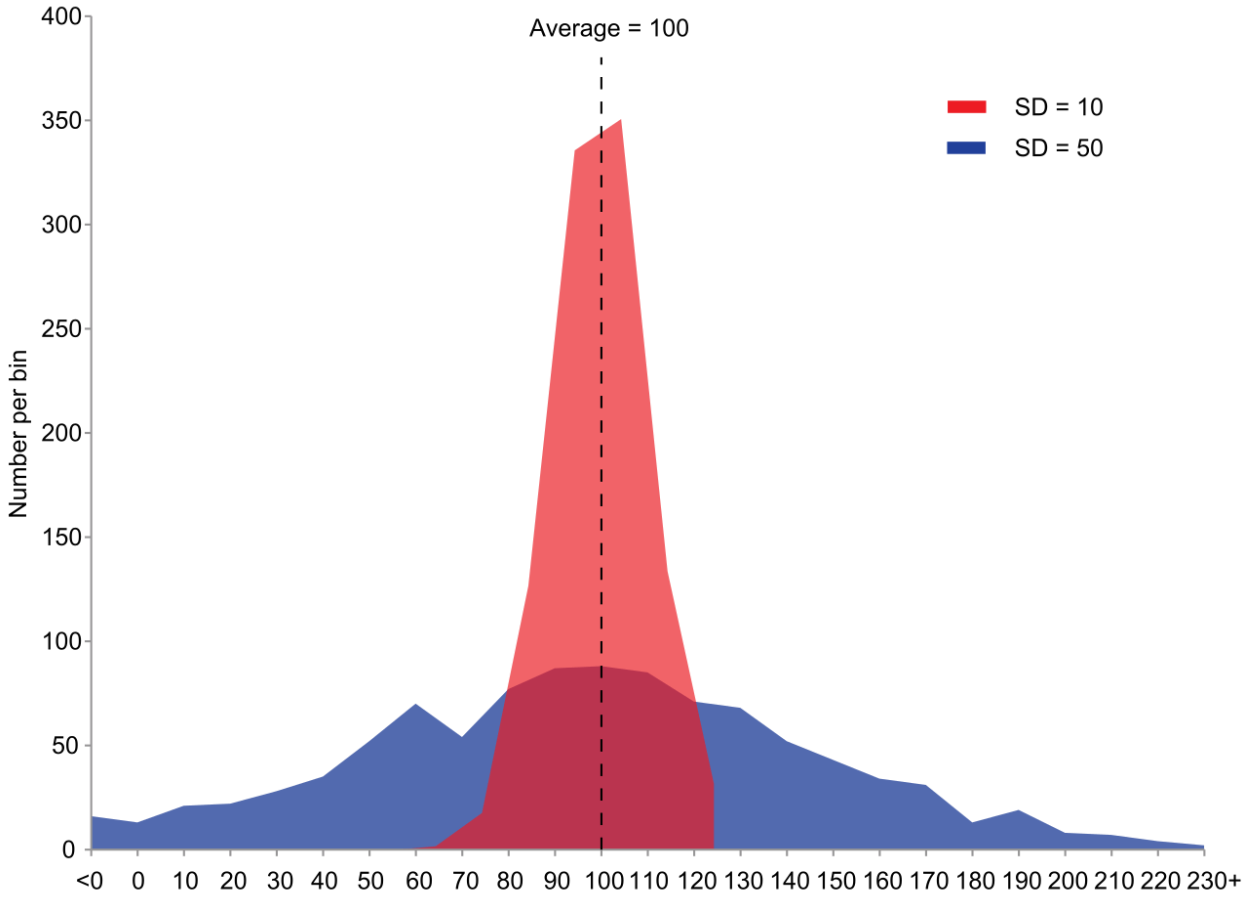
$$Variance = \sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

The **standard deviation** of a data set of N entries is the square foot of the variance.

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Standard deviation is usually interpreted as how measurements for a group are spread out from the average value. It is also defined as measure of the typical amount an entry deviates from the mean. The figure shows an example of two sample populations with the same mean and different standard

deviations. Both populations have a mean of 100, but the red population's standard deviation is 10; whereas the blue population has a standard deviation of 50. The result is a very distributed data for the blue data set.

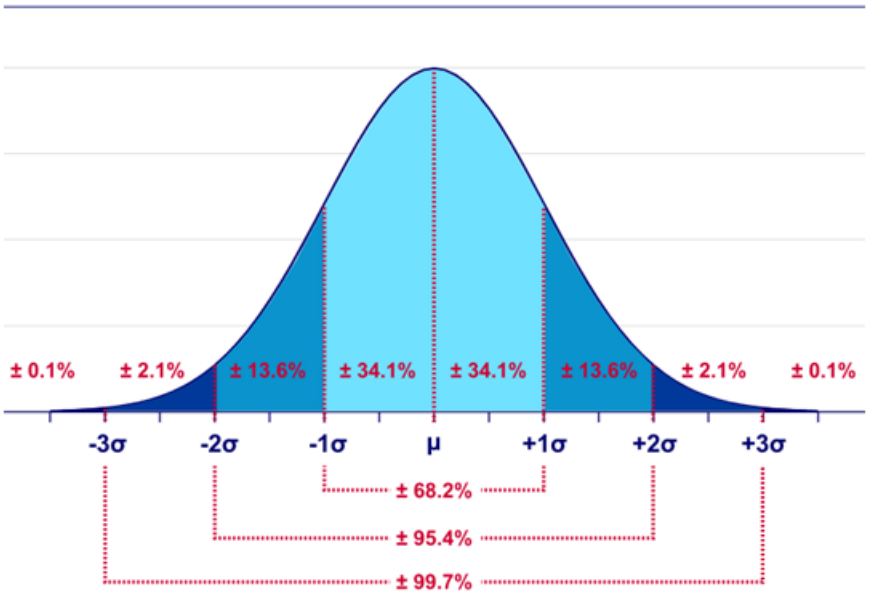



By JRBrown - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10777712.

**Empirical Rule (68 – 95 – 99.7 Rule)**

For data with a bell-shaped distribution, the standard deviation has the following characteristics:

- About 68 % of the data lie within one standard deviation of the mean.
- About 95 % of the data lie within two standard deviations of the mean.
- About 99.7 % of the data lie within three standard deviations of the mean.



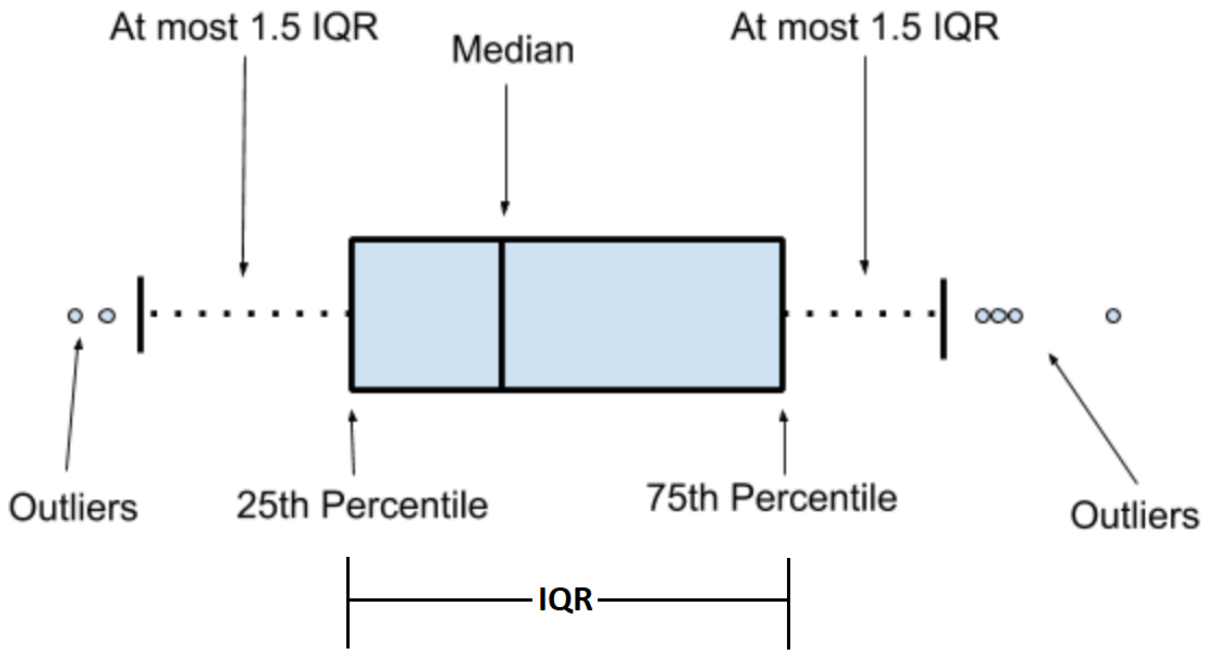

## Measures of Position

The three **quartiles**, *Q1*, *Q2* and *Q3*, approximately divide an ordered data set into four equal parts. About one quarter of the data fall on or below the **first quartile** *Q1*. About one half of the data fall on or below the **second quartile** *Q2* (the second quartile is the same as the median of the data set). About three quarters of the data fall on or below the **third quartile** *Q3*.

The **interquartile range (IQR)** of a data set is the difference between the third and first quartiles.

IQR is a measure of variation that gives an idea of how much the middle 50 % of the data varies. It can also be used to identify the outliers. Any data value that lies more than 1.5 IQRs to the left of *Q1* or to the right of *Q3* is an outlier.

A **box-and-whisker plot** (shown below) is an exploratory data analysis tool that highlights the important features of a data set.




You can access to this article and similar ones here.


Photo by [Luke Chesser](#) on [Unsplash](#)