**Descriptive Statistics**

‘Descriptive’, the term originated from the word ‘describe’ means explain. It is used to explain the characteristics of a data set using measures such as mean, mode, median, variance etc. These measures summarize data and help you draw meaningful patterns.

#types of variable

Descriptive Statistics can further be divided into two parts:

* Measures of Central Tendency(location)
* Measures of Spread
* Measure of Shape

**Measure of Central Tendency**

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position. It sometimes called measures of central location.

The mean, median and mode are measures of central tendency, under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

**Mean (Average):**  It is calculated as the sum of all values in a data set / number(count) of values. It can be used with both discrete and continuous variables, but, it is often used with continuous variables.

Mean = Sum of all Observations / No. of Observations

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

**When not to use the mean**

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Staff | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Salary | 15k | 18k | 16k | 14k | 15k | 15k | 12k | 17k | 90k | 95k |

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

So mean is, it is highly affected by outliers. Also, mean cannot be used while dealing with skewed distributions.  In such cases, we make use of mode/ median.

Let’s find, the average value of Fare (variable in data set) charged to board Titanic Ship.

#set working directory

setwd("C:/Users/manish/Desktop/RData")

#load the data set

train <- read.csv("C:/Users/manish/Desktop/Titanic/train1.csv", stringsAsFactors = FALSE, header = T)

#find the average(mean) Fare

mean(train$Fare)

[1] 32.20421

Interpretation: On an average, passengers have paid $32 to board the titanic.

**Median:**

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 | 92 |

We first need to rearrange that data into order of magnitude (smallest first):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 35 | 45 | 55 | 55 | **56** | 56 | 65 | 87 | 89 | 92 |

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 |

We again rearrange that data into order of magnitude (smallest first):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 35 | 45 | 55 | **55** | **56** | 56 | 65 | 87 | 89 |

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Let’s compute the median of Fare in titanic data.

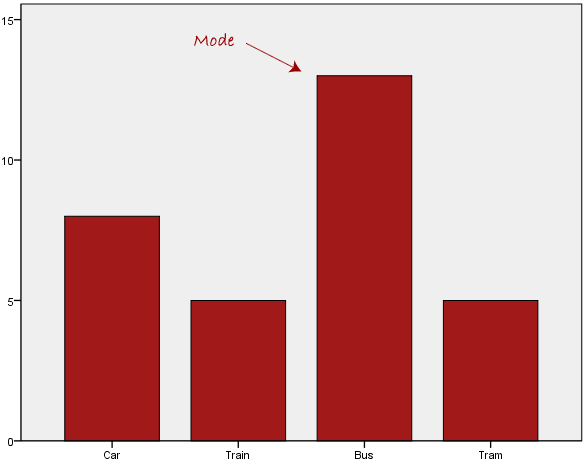
#find out the median

median(train$Fare)

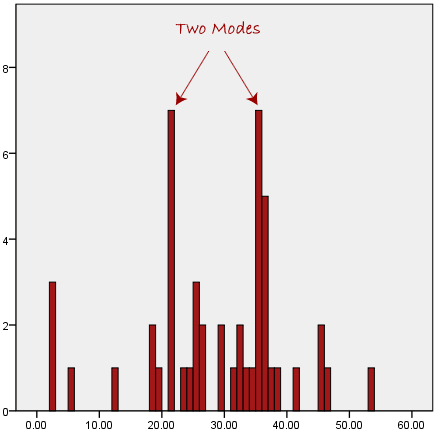
[1] 14.4542

Interpretation: The mid value of Fare variable is $14.45. This means $14.45 divides the data into two halves. If you compute its range, along with this statistic, you’ll find that this variable is highly skewed (don’t worry, if you don’t understand – we will discuss these terms shortly).

**Mode:** The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:

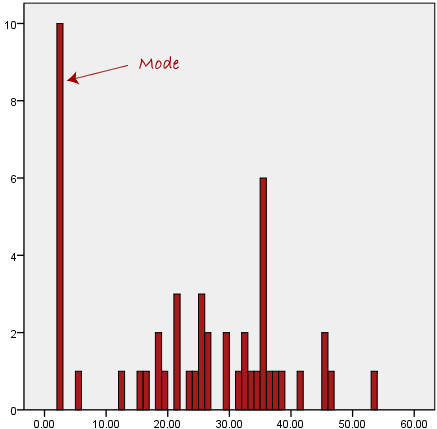


We can see above that the most common form of transport is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:



We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with **exactly** the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely - many people might be close, but with such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data.

Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:



In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

R does not have any inbuilt function to find out Mode. However, it does have a mode() function which returns the type or storage mode of an object instead. Let’s see how to find out the age with highest frequency.

#using inbuild mode function

mode(train$Age)

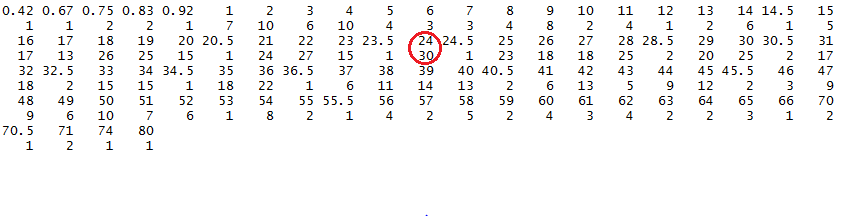
[1] "numeric"

#using a simple method

actual\_mode <- table(train$Age)    #creates a table of all age groups

#view the output of actual\_mode

actual\_mode                   #let's treat the age values here as names

[](http://www.analyticsvidhya.com/wp-content/uploads/2015/10/actual_mode1.png)

#find the most frequent value using max

names(actual\_mode)[actual\_mode == max(actual\_mode)]

[1] "24"

Interpretation: Most common age among passengers on Titanic was 24 years. As you can see, there were 30 passengers on board who are 24 years old (highest among all).

**Summary of when to use the mean, median and mode**

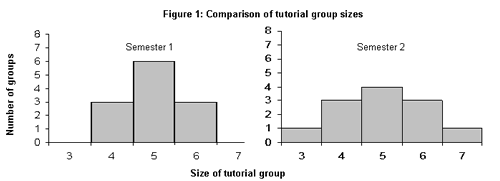
Please use the following summary table to know what the best measure of central tendency is with respect to the different types of variable.

|  |  |
| --- | --- |
| **Type of Variable** | **Best measure of central tendency** |
| Nominal | Mode |
| Ordinal | Median |
| Interval/Ratio (not skewed) | Mean |
| Interval/Ratio (skewed) | Median |

**Measure of Spread**

Measures of average such as the median and mean represent the typical value for a dataset. Within the dataset the actual values usually differ from one another and from the average value itself. The extent to which the median and mean are good representatives of the values in the original dataset depends upon the variability or dispersion in the original data. Datasets are said to have high dispersion when they contain values considerably higher and lower than the mean value.

In figure 1 the number of different sized tutorial groups in semester 1 and semester 2 are presented. In both semesters the mean and median tutorial group size is 5 students, however the groups in semester 2 show more dispersion (or variability in size) than those in semester 1.



Dispersion within a dataset can be measured or described in several ways including the range, inter-quartile range and standard deviation.

## Range

The range is the most obvious measure of dispersion and is the difference between the lowest and highest values in a dataset. In figure 1, the size of the largest semester 1 tutorial group is 6 students and the size of the smallest group is 4 students, resulting in a range of 2 (6-4). In semester 2, the largest tutorial group size is 7 students and the smallest tutorial group contains 3 students, therefore the range is 4 (7-3).

* The range is simple to compute and is useful when you wish to evaluate the whole of a dataset.
* The range is useful for showing the spread within a dataset and for comparing the spread between similar datasets.

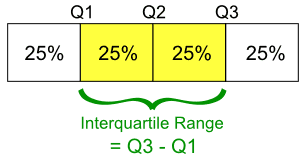
We can calculate range using this simple function:

#calculate range

range(train$Fare)

[1] 0.0000 512.3292

**InterQuartile Range:** It is the difference between the 75th and 25th percentile. The inter-quartile range provides a clearer picture of the overall dataset by removing/ignoring the outlying values.



Like the range however, the inter-quartile range is a measure of dispersion that is based upon only two values from the dataset. Statistically, the standard deviation is a more powerful measure of dispersion because it takes into account every value in the dataset.

We can calculate interquartile range using this simple function:

#calculate range

IQR(train$Fare)

[1] 0.0000 512.3292

**Variance and Standard Deviation:** Variance and Standard Deviations are closely related. In simple words:

SQRT(variance) = Standard Deviation

Variance is the average of squared difference from mean. When the data points are clusted around mean the variance is small.

The standard deviation is a measure that summarizes the amount by which every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. It is the most robust and widely used measure of dispersion since, unlike the range and inter-quartile range, it takes into account every variable in the dataset. When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large.

**Population and sample standard deviations**

There are two different calculations for the Standard Deviation. Which formula you use depends upon whether the values in your dataset represent an entire population or whether they form a sample of a larger population. For example, if all student users of the library were asked how many books they had borrowed in the past month then the entire population has been studied since all the students have been asked. In such cases the population standard deviation should be used. Sometimes it is not possible to find information about an entire population and it might be more realistic to ask a sample of 150 students about their library borrowing and use these results to estimate library borrowing habits for the entire population of students. In such cases the sample standard deviation should be used.

**Formulae for the standard deviation**

The standard deviation of an entire population is known as σ (sigma) and is calculated using:

var6.gif

Where x represents each value in the population, μ is the mean value of the population, Σ is the summation (or total), and N is the number of values in the population.

The standard deviation of a sample is known as S and is calculated using:

var7.gif

Where x represents each value in the population, x is the mean value of the sample, Σ is the summation (or total), and n-1 is the number of values in the sample minus 1.

In R, you don’t need to get into mathematics of these measures since we have inbuilt functions to calculate these values. Let’s find out the variance & standard deviation for Fare variable.

#variance of fare

var(train$Fare)

[1] 2469.437

#Standard Deviation of Fare

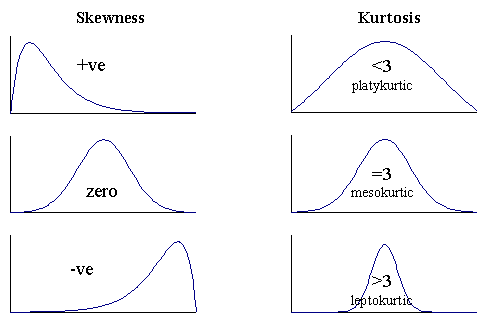
sqrt(var(train$Fare))

[1] 49.69343

**Shape of Distribution**

#### ****Measures to describe shape of distribution:****

* **Skewness** – Skewness is a measure of the asymmetry. Negatively skewed curve has a long left tail and vice versa.
* **Kurtosis** – Kurtosis is a measure of the “peaked ness”. Distributions with higher peaks have positive kurtosis and vice-versa

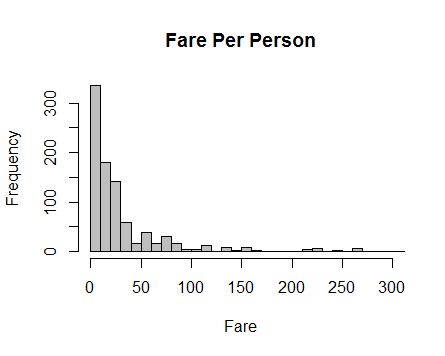
[](http://www.analyticsvidhya.com/blog/wp-content/uploads/2014/07/skewness-and-kurtosis.gif)

 Let’s look at the distribution of Fare using a bar chart:

#create bar chart of Fare

hist(train$Fare, main = "Fare Per Person", xlab = "Fare", col = "grey", breaks = 40,

xlim = c(0,300))

[](http://www.analyticsvidhya.com/wp-content/uploads/2015/10/Positive-Skewed.png)

**Question:** How is this distribution skewed? Positive or Negative?

Some modeling techniques (like regressions) make an assumption about the underlying distributions of the population. In case your population does not fulfill them, you may need to make some transformations to improve the results of your models.