# Univarite Statistics

In general the data you deal with in industry is too big to go through each and every observation. To understand the data you need to come up with a way to summarise this so that you can quickly go through the summary and understand your data faster.

In this module we are going to learn what kind of summaries to consider to understand your data in a wholesome manner. There are three facets of the data that we need to summarize in order to capture all aspects of the data:

- Central Tendency
- Variability
- Shape

We can create these summaries in two ways:

- Numeric Summary
- Data Visualisation

We'll learn how to summarise each aspect of the data numerically first. Also here onwards in this module , when we say data, we are talking about just one variable. [**Uni**variate Statistics]

### Central Tendency

Central tendency is average behaviour of the data. For example if somebody asks a student , "what is the age of a typical student in their class?", best bet is the average age of students in the class.

Central tendency is the value of your data which is most expected outcome. There are three measures for central tendency:

- Mean
- Median
- Mode

### Mean

Mean is defined as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

This is simply average of the values. Consider the average of these numbers: {1,2,3,4,5,6,7} . It is 4. Now if I include another value to this set of values say 10000, then the average becomes 1257, which is no where close to majority of the data {1,2,3,4,5,6,7,10000}.

This tells us that mean is not a good measure of central tendency if data contains extreme values or outliers.

### Median

Median is defined as the middle most value when your data is sorted. For example is your data was {1,10,8,7,9,12,3}, you'd sort it : {1,3,7,8,9,10,12} , your data has 7 elements, the middle most term will be $\frac{(7+1)}{2}^{th}$ term which is 8.

If your data has even number of elements then you'll have two middle most values. For example lets consider this data : {1,3,7,8,9,10,12,14000}. two middle most terms are : 8 & 9, median is taken as average of these two middle most values. which is 8.5

You might have noticed by now that median doesnt not depend on the values of elements , only on their order in the data which makes it not so sensitive to extreme values present in the data.

### Mode

Mode is simply defined as the value which occurs most frequently in the data. It is mostly used for categorical variables because numerical summaries do not make sense in context of categorical variables.

A dataset can have multiple modes , since highest frequency can be equal for many categories of the categorical variable. As opposed to this mean and median for a dataset take a single value.

### Variability

datasets having same average do not neccessarily behave in a same fashion. They might differ in terms of spread. Here is an example, both the data sets have their average as 10.

{7,8,9,10,11,12,13} : 10

{-100,-50,10,90,100} : 10

you can see that the second data is much more spread out in comparison to the first one, inspite of them having same mean. There are many ways in which this spread can be measured.

- Range
- Mean Absolute Deviation (MAD)
- Standard Deviation / Variance
- Inter Quartile Range

### Range

Range is simply defined as difference between minimum and maximum values in the data. It's a very primitive measure and is rarely used. You can observe that since it relies on extreme values in your data to start with, it is very sensitive to presence of extreme values in your data

### Mean Absolute Deviation [MAD]

Spread is defined as deviation of individual values from the mean of the data. If you simply try to add these deviations and take average, positive and negative deviations will cancel each other out . Average absolute values of these deviation is one way to avoid effect of signs in the deviations. MAD is defined as :

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

MAD is rarely used in practice because it is very tedious to manipulate it algebraically hence it hasnt been considered extensivley in theory.

### Standard Deviation and Variance

Standard Deviation is defined as :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

variance is simply square of the standard deviation: $\sigma^2$

### Inter Quartile Range [IQR]

All of the measures that we discussed for spread are sensitive to extreme values. IQR is one measure which is not sensitive to outliers. It is defined as difference between $q_1$ and $q_3$ where $q_1$ , $q_3$ are first and third quartiles respectively.

quartiles are the values which divide your data into 4 parts. Median is second quartile or $q_2$ , it divides into two equal parts. You can consider $q_1$ to be median of the first part and $q_3$ to be median of the second part.

IQR is not sensitive to outliers. why?

### Shape

Measures of central tendency and variability still do not tell how frequently a value occurs in your data. Or in other words , given a value what are the chances that it will be in the data.

To get an idea regarding the same we can plot frequency bar charts or histograms.The Shape depicted by these histograms can be of three types:

- Symmetric
- Positively Skewed
- Negatively Skewed

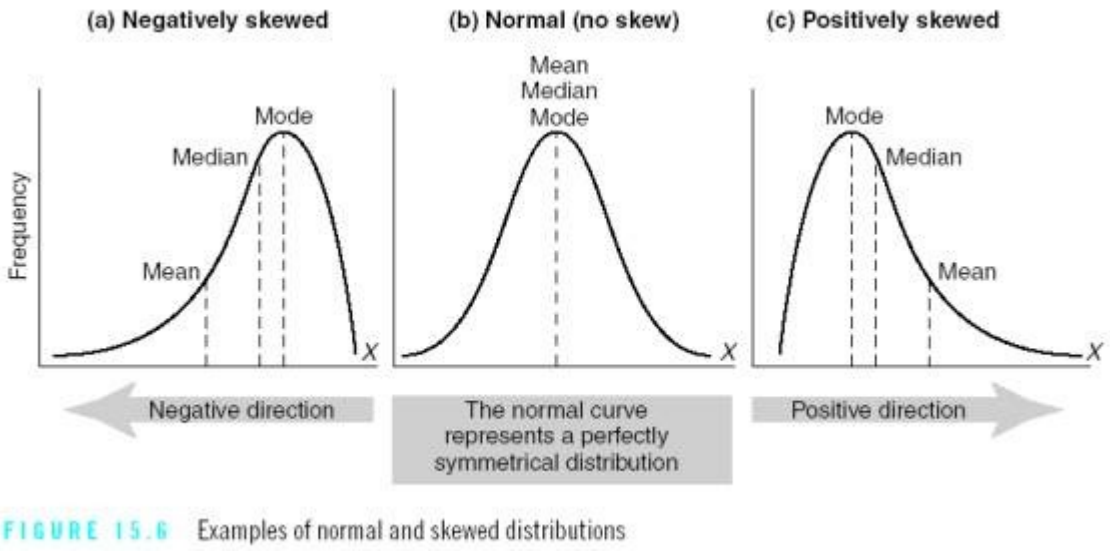


FIGURE 15.6 Examples of normal and skewed distributions

Figure 1: shape

Shape is also represented in terms of distributions which we discuss in detail in hypothesis testing module. For now you can understand symmetric shape to be where data values on either side of the mean are equally probable where as for positively and negatively skewed data has unbalanced properties.

For a symmetric shape

$$q_2 - q_1 = q_3 - q_2$$

where as for positively skewed shape

$$q_2 - q_1 < q_3 - q_2$$