Descriptive Statistics

Dr. Praveen R. Barapatre

Descriptive Statistics

- Methods of describing the characteristics of a data set.
- Useful because they allow you to make sense of the data.
- Helps exploring and making conclusions about the data in order to make rational decisions.
- Includes calculating things such as the average of the data, its spread and the shape it produces.

Descriptive Statistics

- Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood.
- **Graphical displays** are often used along with the quantitative measures to enable clarity of communication.

Describing data

- **Qualitative data**-

the variable which yield non numerical data.

  - E.g.- education, marital status, eye colour
  - **Frequency**- number of observations falling into particular class/ category of the qualitative variable.
  - **Frequency distribution**- table listing all classes & their frequencies.
  - Graphical representation- **Pie chart, Bar graph**.

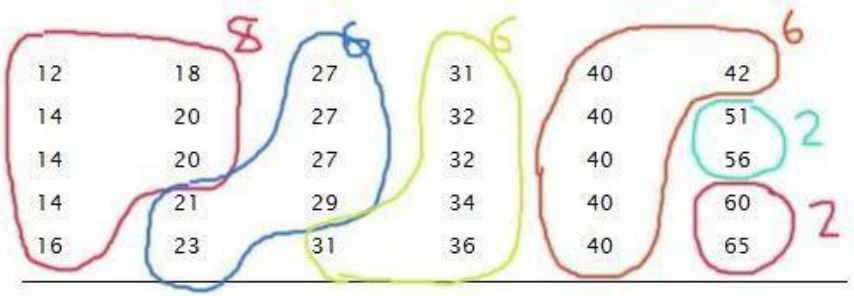
Describing data

- **Quantitative data**-
  - Can be presented by a frequency distribution.
  - If the discrete variable has a lot of different values, or if the data is a continuous variable then data can be grouped into classes/ categories.

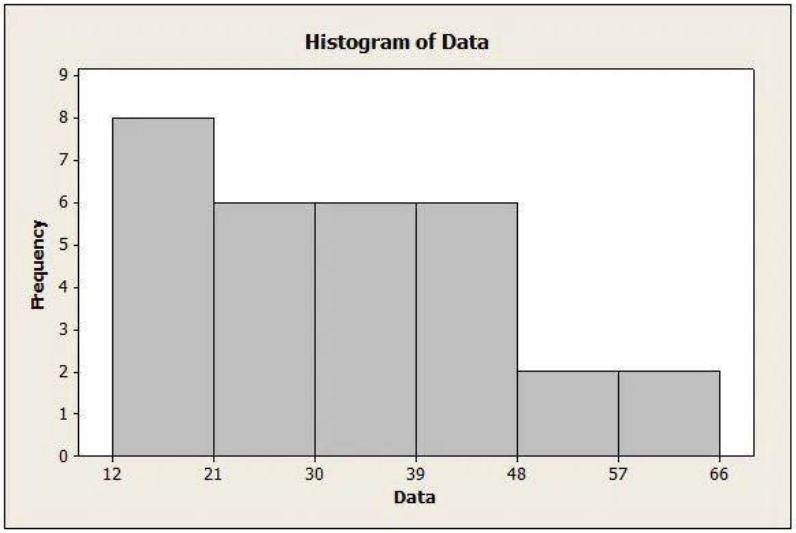
  - **Class interval / BINS**- covers the range between maximum & minimum values.
  - **Class limits**- end points of class interval.
  - **Class frequency**- number of observations in the data that belong to each class interval.

  - Usually presented as a **Histogram** or a **Bar graph**.

Frequency Distribution and Histogram

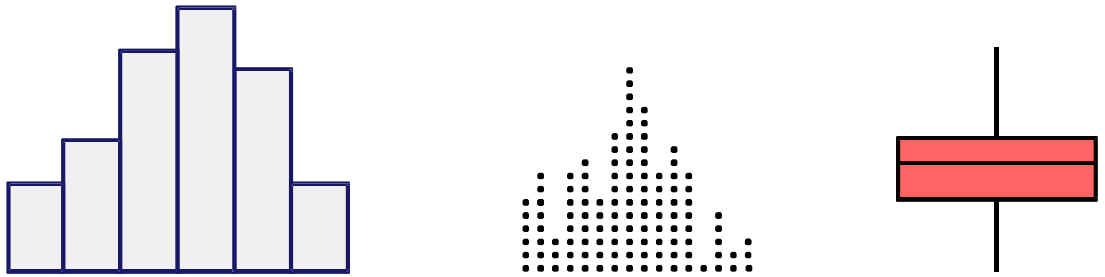


CLASSES	FREQUENCY
12 – 21	8
21 – 30	6
30 – 39	6
39 – 48	6
48 – 57	2
57 – 66	2



Descriptive Statistics

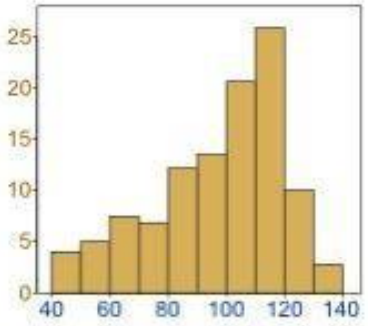
- When analyzing a graphical display, you can draw conclusions based on several characteristics of the graph.
- **You may ask questions such ask:**
  - Where is the approximate middle, or center, of the graph?
  - How spread out are the data values on the graph?
  - What is the overall shape of the graph?
  - Does it have any interesting patterns?



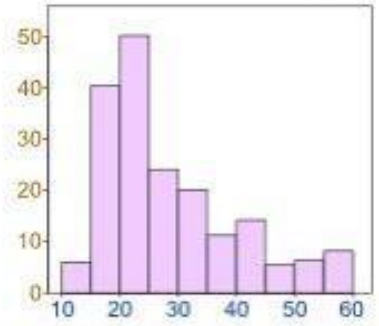
Normal Distribution

Data can be "distributed" (spread out) in different ways.

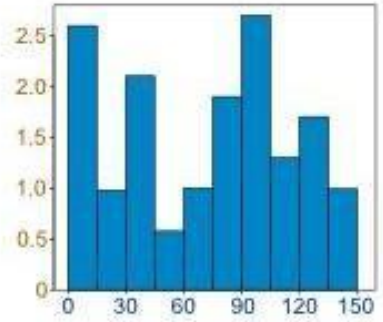
It can be spread out more on the left



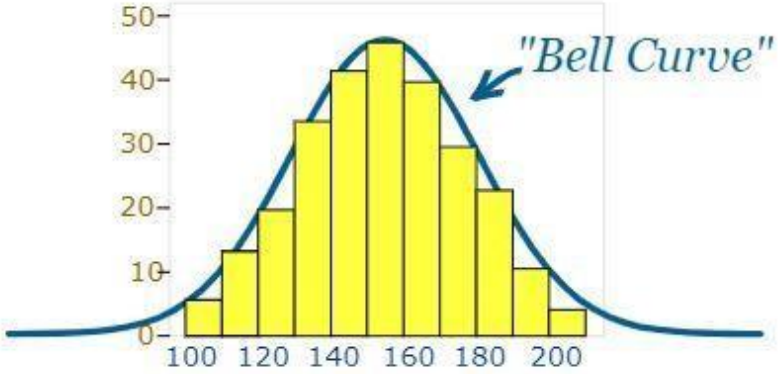
Or more on the right



Or it can be all jumbled up



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



A Normal Distribution

The "Bell Curve" is a Normal Distribution.
And the yellow histogram shows some data that follows it closely, but not perfectly (which is usual).

It is often called a "Bell Curve" because it looks like a bell.

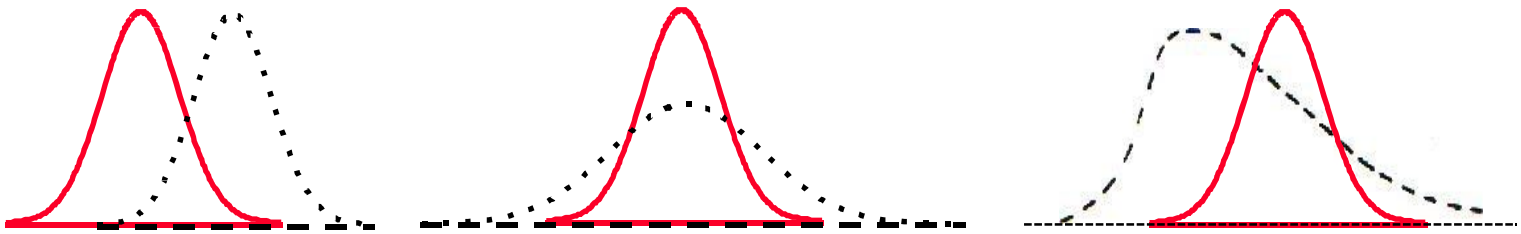
Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

Descriptive Statistics

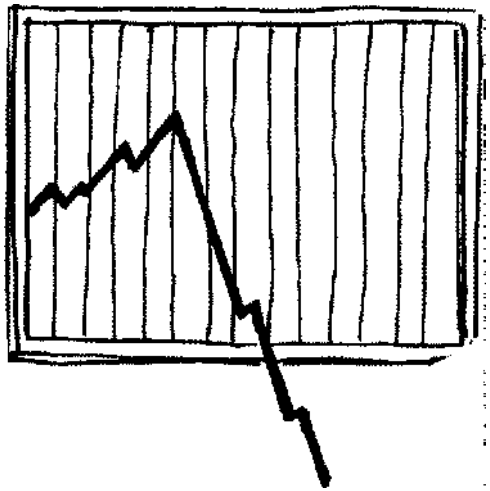
The following measures are used to describe a data set:

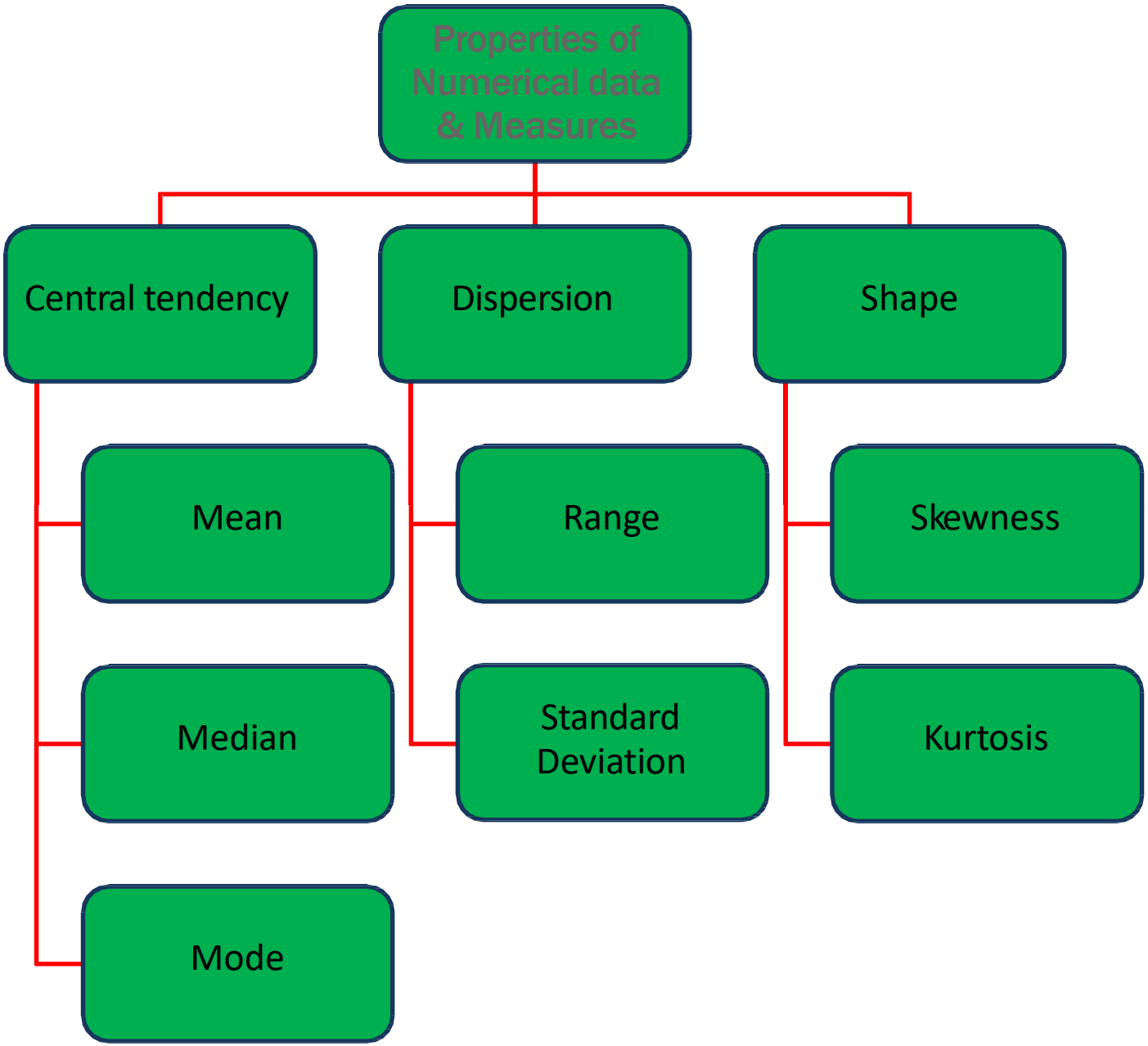
- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.



Descriptive Statistics

- If assignable causes of variation are affecting the process, we will see changes in:
  - Position.
  - Spread.
  - Shape.
  - Any combination of the three.




Properties of Numerical data & Measures
Central tendency
Dispersion
Shape
Mean
Median
Mode
Range
Standard Deviation
Skewness
Kurtosis

Descriptive Statistics

Measures of Position:

- Position Statistics measure the data central tendency.
- Central tendency refers to where the data is centered.
- You may have calculated an average of some kind.
- Despite the common use of average, there are different statistics by which we can describe the average of a data set:
  - Mean.
  - Median.
  - Mode.

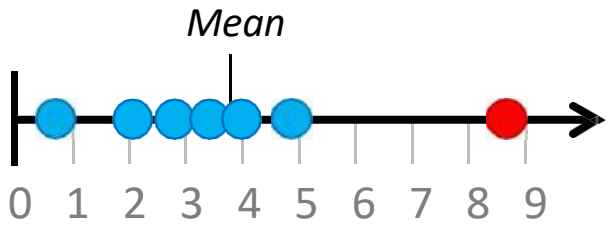
Measures of center

- **Central tendency**- In any distribution, majority of the observations ***pile up, or cluster around*** in a particular region.
- **Mean**- sum of observed values in a data divided by the number of observations
- **Median**- observation in the data set that divides the data set into half.
- **Mode**- value of the data set which occurs with greatest frequency
- Mean & Median can be applied only to Quantitative data
- Mode can be used either to Qualitative or Quantitative data.
- **Outlier**- observation that falls far from the rest of the data. Mean gets highly influenced by the outlier.

Descriptive Statistics

Mean:

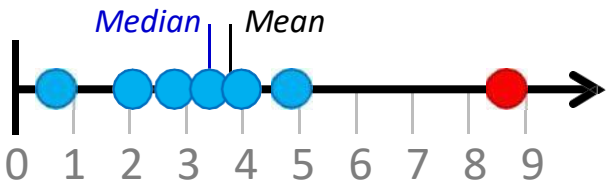
- The total of all the values divided by the size of the data set.
- It is the most commonly used statistic of position.
- It is easy to understand and calculate.
- It works well when the distribution is symmetric and there are no outliers.
- The mean of a sample is denoted by '**x-bar**'.
- The mean of a population is denoted by '**μ**'.



Descriptive Statistics

Median:

- The middle value where exactly half of the data values are above it and half are below it.
  - Less widely used.
  - A useful statistic due to its robustness.
  - It can reduce the effect of outliers.
- Often used when the data is nonsymmetrical.
- Ensure that the values are ordered before calculation.
- With an even number of values, the median is the mean of the two middle values.



Descriptive Statistics

Median Calculation:

23
33
34
36
38
40
41
41
44

12
30
31
37
38
40
41
41
44
45

Median = 38 + 40 / 2 = 39

Example

1,2,1,1,3,4,100

Mean = 16

median = 2

mode = 1

Assume 100 is an outlier

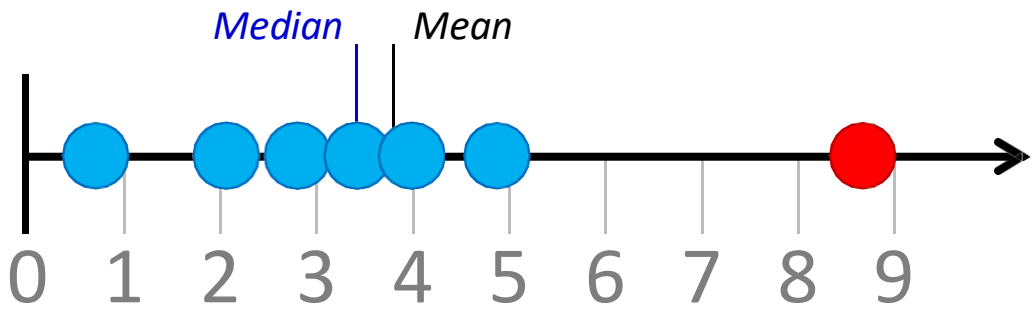
Mean =2

median = 1.5

mode = 1

Descriptive Statistics

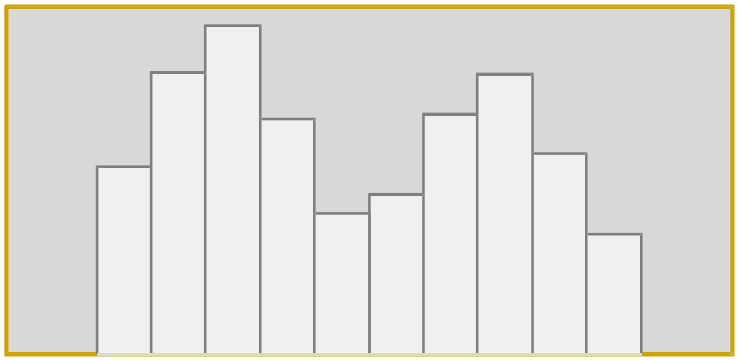
- Why can the mean and median be different?



Descriptive Statistics

Mode:

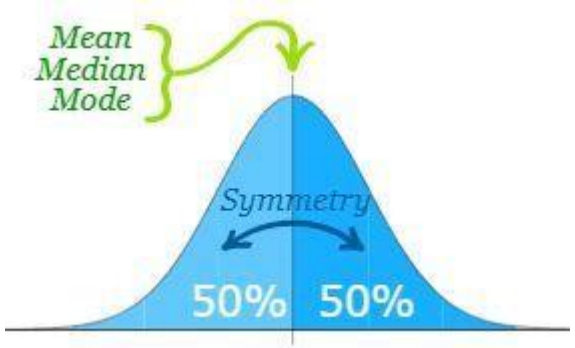
- The value that occurs the most often in a data set.
- It is rarely used as a central tendency measure
- It is more useful to distinguish between unimodal and multimodal distributions
  - When data has more than one peak.



Normal distribution

- Bell shaped symmetric distribution.
- Why is it important?
  - Many things are normally distributed, or very close to it.
  - It is easy to work with mathematically
  - Most inferential statistical methods make use of properties of the normal distribution.

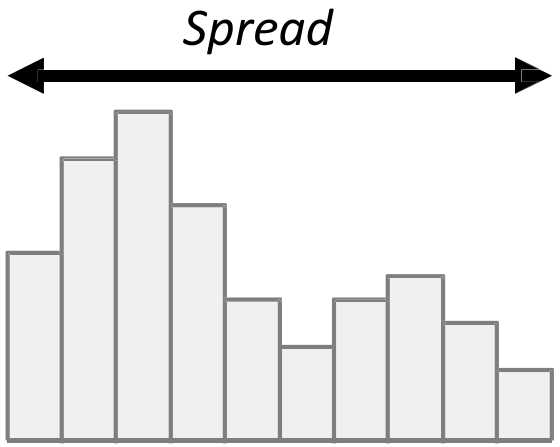
- Mean = Median = Mode



Descriptive Statistics

Measures of Spread:

- The **Spread** refers to how the data deviates from the position measure.
- It gives an indication of the amount of variation in the process.
  - An important indicator of quality.
  - Used to control process variability and improve quality.
- All manufacturing and transactional processes are variable to some degree.
- There are different statistics by which we can describe the spread of a data set:
  - Range.
  - Standard deviation.

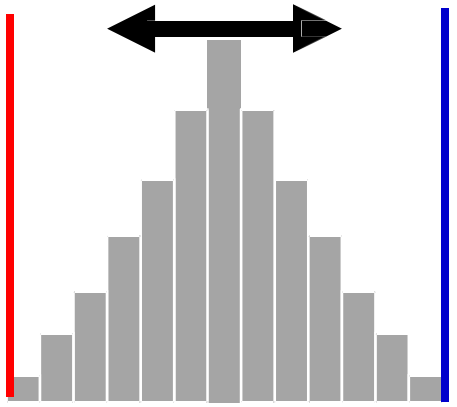


- **Range**- difference between the largest observed value in the data set and the smallest one.
  - So, while considering range great deal of information is ignored.
- **Standard deviation**- it is a kind of average of the absolute deviation of observed values from the mean of the variable.
  - It is defined using the sample mean & values get strongly affected by few extreme observations.
- **Variance**- square of standard deviation

Descriptive Statistics

Standard Deviation:

- The average distance of the data points from their own mean.
- A low standard deviation indicates that the data points are clustered around the mean.
- A large standard deviation indicates that they are widely scattered around the mean.
- The standard deviation of a sample is denoted by '**s**'.
- The standard deviation of a population is denoted by "**μ**".



Descriptive Statistics

Standard Deviation:

- Perceived as difficult to understand because it is not easy to picture what it is.
- It is however a more robust measure of variability.
- Standard deviation is computed as follows:

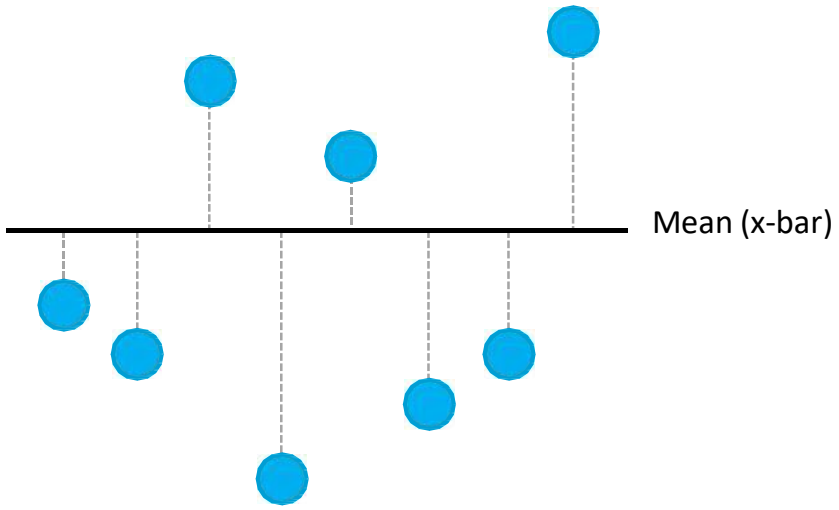
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

s = standard deviation

$\bar{x}$ = mean

x = values of the data set

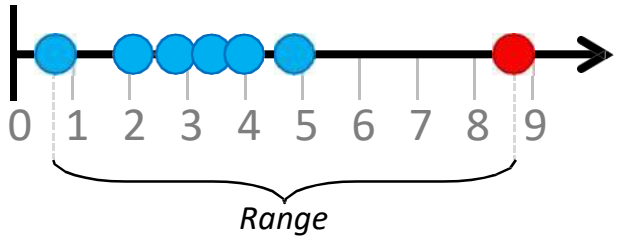
n = size of the data set



Descriptive Statistics

Range:

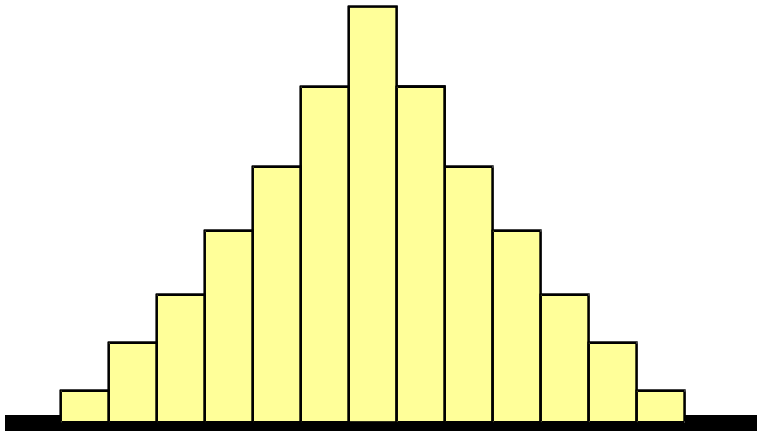
- The difference between the highest and the lowest values.
- The simplest measure of variability.
- Often denoted by '**R**'.
- It is good enough in many practical cases.
- It does not make full use of the available data.
- It can be misleading when the data is skewed or in the presence of outliers.
  - Just one outlier will increase the range dramatically.



Descriptive Statistics

Measures of Shape:

- Data can be plotted into a histogram to have a general idea of its shape, or distribution.
- The shape can reveal a lot of information about the data.



Shape

- **Skewness**- Lack of **symmetry** in distribution. It can be interpreted from frequency distribution.

- Properties-
  - Mean, median & mode fall at different points.
  - Curve is not symmetrical but stretched more to one side.

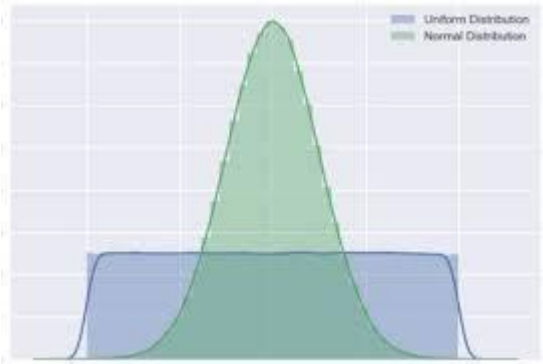
- Distribution may be **positively or negatively skewed**. Limits for coefficient of skewness is ± 3.

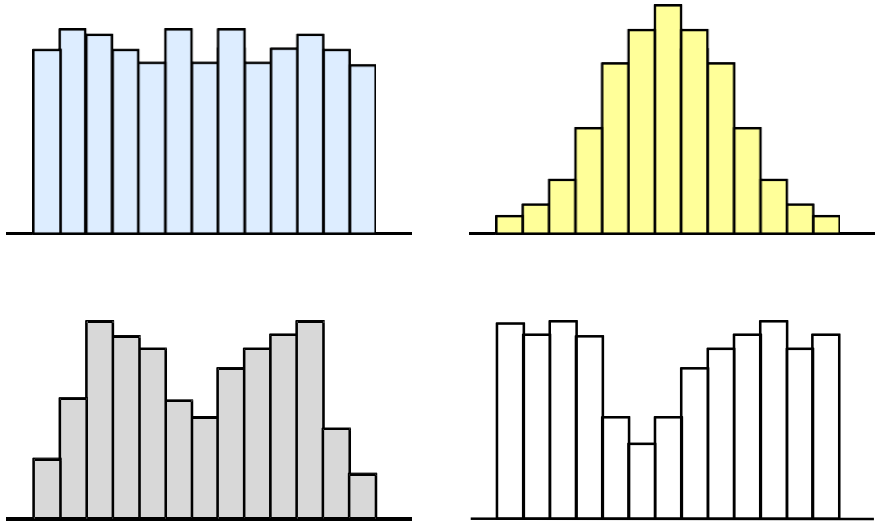
- **Kurtosis**- convexity of a curve.
  - Gives an idea about the **flatness/ peakedness** of the curve.
  - Gives an idea about how much weights are at the tail end of the distribution

Descriptive Statistics

Measures of Shape:

- It may be symmetrical or nonsymmetrical.
- In a symmetrical distribution, the two sides of the distribution are a mirror image of each other.
- Examples of **symmetrical** distributions include:
  - Uniform.
  - Normal.
  - Camel-back.





Descriptive Statistics

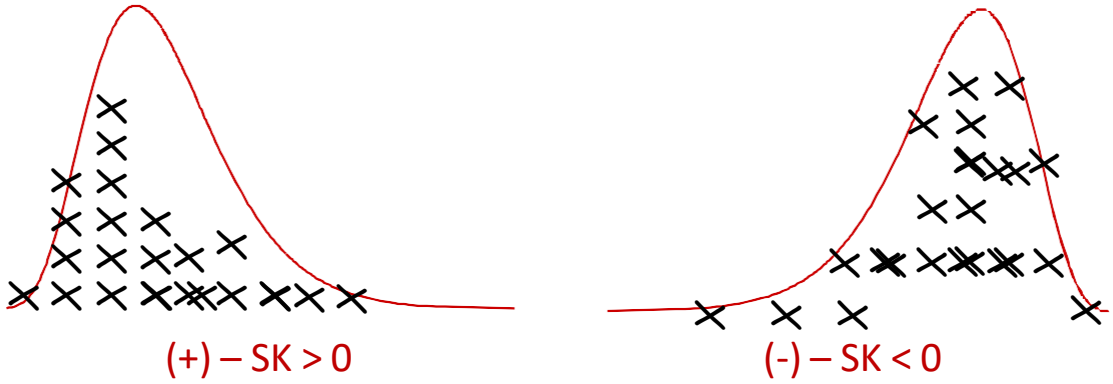
Measures of Shape:

- The shape helps identifying which descriptive statistic is more appropriate to use in a given situation.
- If the data is symmetrical, then we may use the mean or median to measure the central tendency as they are almost equal.
- If the data is skewed, then the median will be a more appropriate to measure the central tendency.
- Two common statistics that measure the shape of the data:
  - Skewness.
  - Kurtosis.

Descriptive Statistics

Skewness:

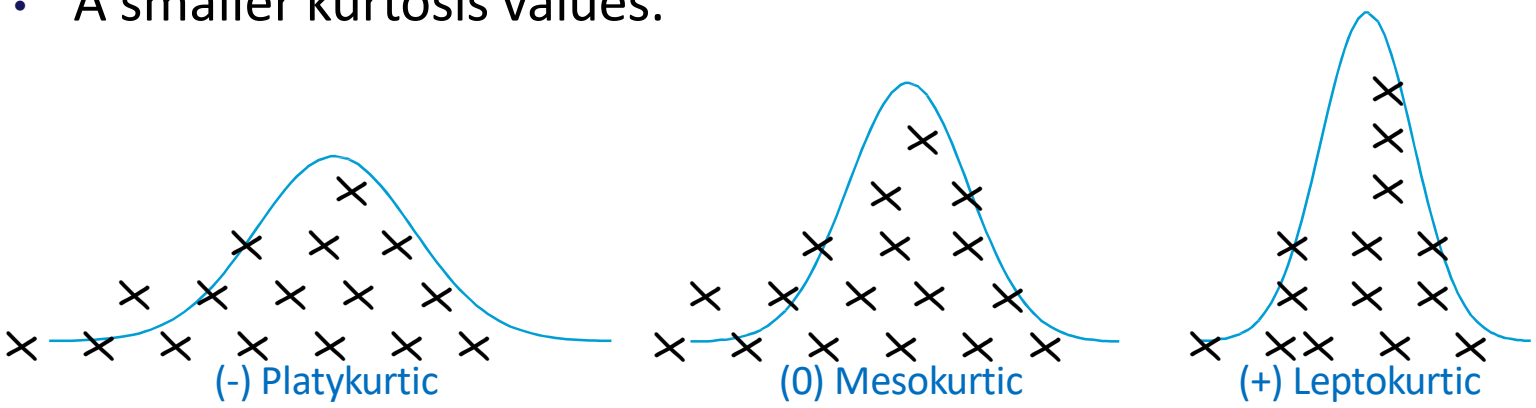
- Describes whether the data is distributed symmetrically around the mean.
- A skewness value of zero indicates perfect symmetry.
- A negative value implies left-skewed data.
- A positive value implies right-skewed data.



Descriptive Statistics

Kurtosis:

- Measures the degree of **flatness** (or **peakness**) of the shape.
- When the data values are clustered around the middle, then the distribution is more peaked.
  - A greater kurtosis value.
- When the data values are spread around more evenly, then the distribution is more flatted.
  - A smaller kurtosis values.



Descriptive Statistics

Further Information:

- **Variance** is a measure of the variation around the mean.
- It measures how far a set of data points are spread out from their mean.
- The units are the square of the units used for the original data.
  - For example, a variable measured in meters will have a variance measured in meters squared.
- It is the square of the standard deviation.

Variance = s^2

Some Formulas

Mean/Average

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Skewness =

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Standard Error

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$kurtosis = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

Standard Error of Means vs Standard Deviation

- The **standard error** (SE) of a **statistic** is the approximate **standard** deviation of a **statistical** sample population.
- the mean and standard deviation are descriptive statistics, whereas the standard error of the mean is descriptive of the random sampling process.
- the **standard error** of the sample mean is an estimate of how far the sample mean is likely to be from the population mean, whereas the **standard deviation** of the sample is the degree to which individuals within the sample differ from the sample mean.