



Artificial Intelligence and Machine Learning (ACCA6011)

Statistics and Probability

Dr. Seena Joseph

Statistics

- Statistics is the branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical data.
- Statistics and data override intuition, inform decisions, and minimize risk and uncertainty.
- Statistics refers to the numerical facts
 - Mark of a student, income of a family, pass rate, salary of a fresh graduate etc.
- Educated guess – decisions made by using statistical methods
- Pure guess - Decisions made without using statistical (or scientific) methods; unreliable

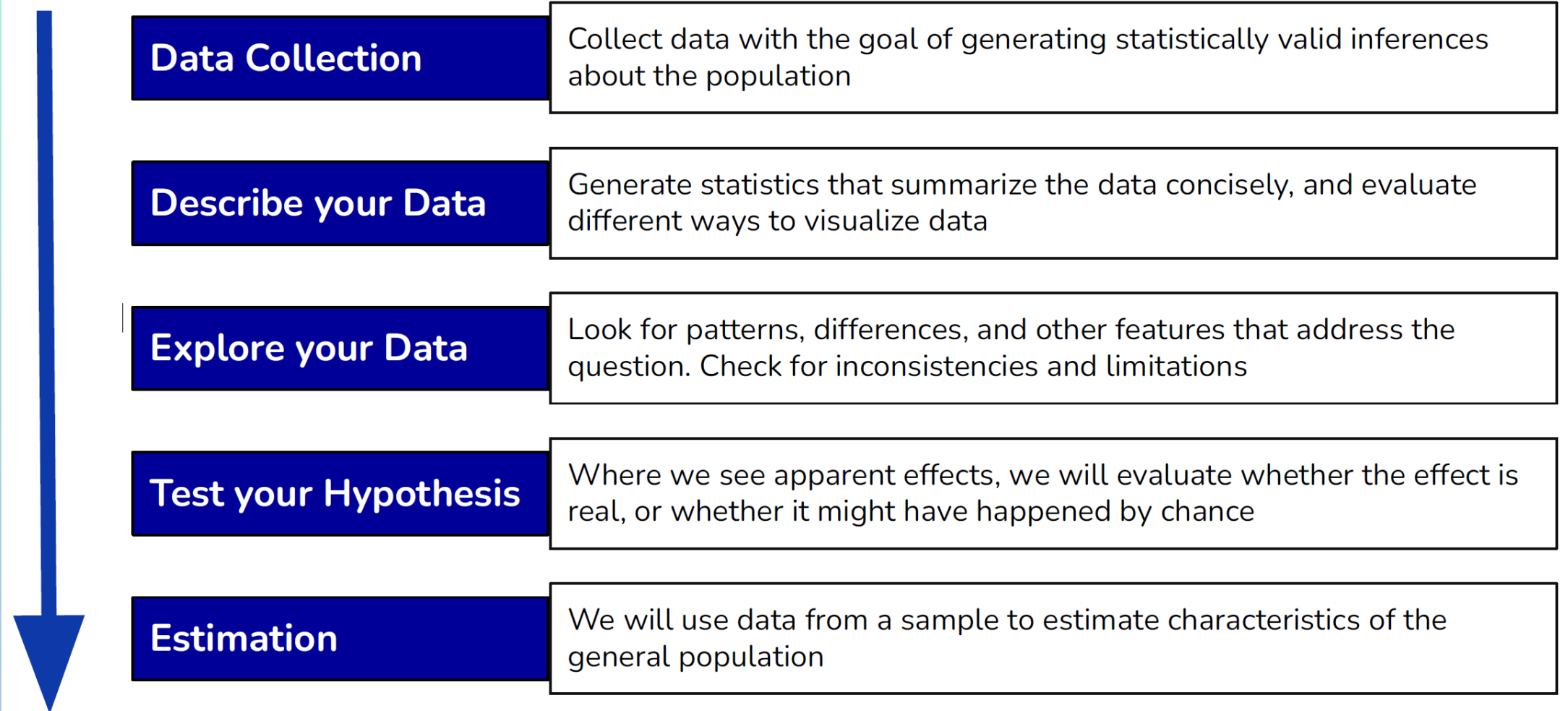
Statistical Thinking

- Statistical thinking is a way of analyzing data, making inferences, and solving problems using statistical methods.
 - **Anecdotal Evidence Vs Statistical Evidence**

Example:

- Anecdotal Evidence
 - A person says. “ I tried *Diet Z* and I lost 5 Kg in two weeks”;
 - A person says, “I used the *DebugTool XYZ* and it fixed the program in 10 minutes”.
 - Are these evidence reliable ?
- Statistical Evidence
 - A research study is conducted with 700 participants to test the effectiveness of Diet Z. The results showed that 75% of participants lost an average of 4 Kg in two weeks.
 - A study conducted with 300 programmers using different debugging tools including DebugTool XYZ over a period of six months. The results of the study showed that 85% of developers using DebugToolXYZ were able to reduce debugging time by an average of 25% compared to other tools.
 - Are these evidence reliable ?
 - What is the difference between these two evidences ?

Statistical Thinking Approach



Why statistics is important in ML ?

- Statistics is at the core of all sophisticated ML Algorithms
- Data scientists use statistics to gather, review, analyze and draw conclusions from data.

[Statistics for Machine Learning: Analysing Unseen Patterns \(turing.com\)](#)

Descriptive and Inferential statistics

Descriptive Statistics involves describing, summarizing and organizing the data so it can be easily understood.

Inferential statistics is a branch of statistics that allows you to make predictions or generalizations about a population based on data collected from a sample. Instead of analyzing an entire population, inferential statistics uses sample data to draw conclusions and make inferences about population parameters.

Descriptive statistics

Descriptive statistics allows to make sense of the data and it helps in exploring and making conclusions about the data in order to make rational decisions.

How to describe your data ?

- 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

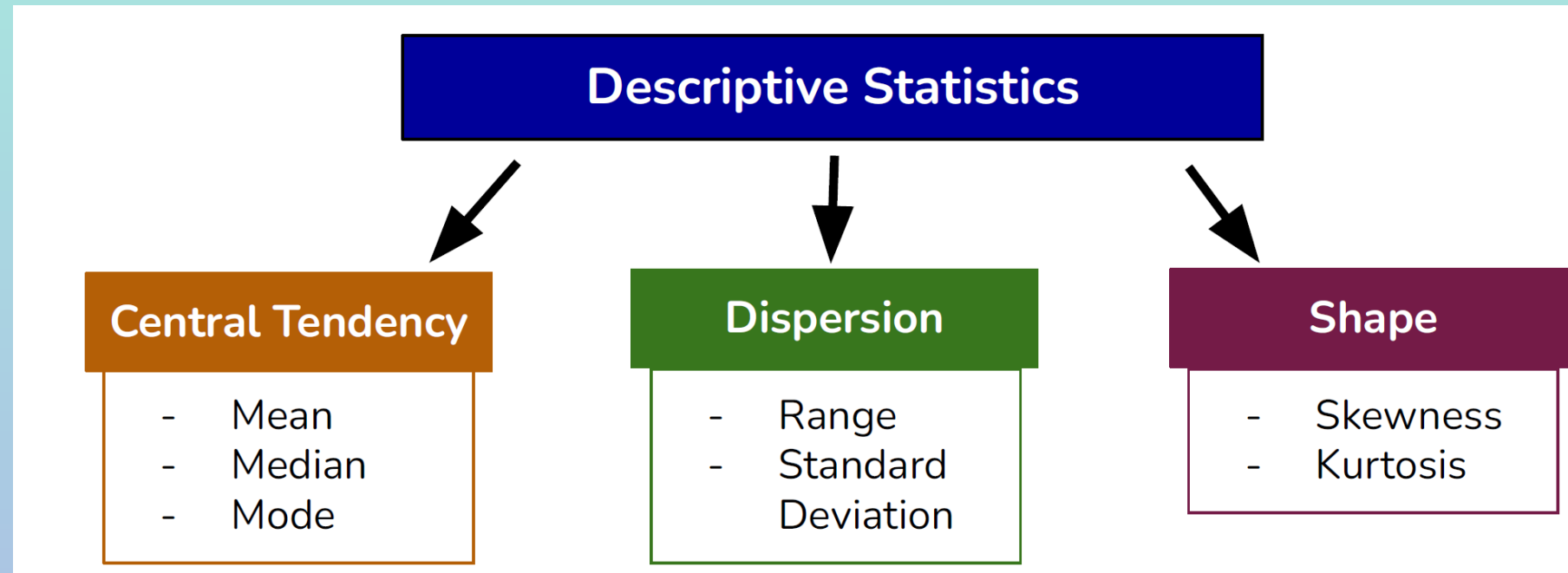
Descriptive statistics

A data set is describe using the following measures.

Measures of Position (referred to as central tendency or location measures)

Measures of spread (referred to as variability or dispersion measures)

Measures of shape



Descriptive statistics

Measure of position– Mean

To calculate the arithmetic mean of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n). Since we obtain the following formula for the mean ($\bar{x}$).

$$\bar{x} = \frac{\sum x}{n}$$

- Why do we need mean?
 - The mean gives us an idea of where the “center” of a dataset is located
 - The mean is also important because it carries a piece of information from every observation in a dataset.
 - Going forward you will notice how mean is used in Machine learning to minimize the prediction errors

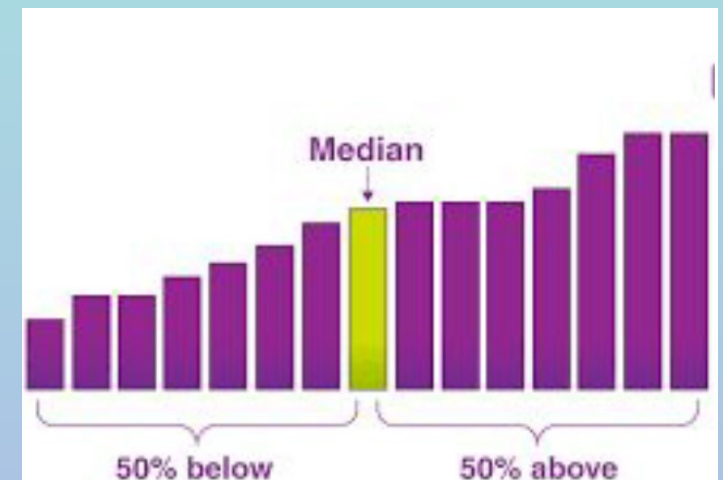
Descriptive statistics

Measure of position– Median

- Median is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not. Median is robust and is less affected by the outliers.
- Why do we need median?
 - The median also represents the 50th percentile of a dataset. That is, exactly half of the values in the dataset are larger than the median and half of the values are lower.

The median is a more useful metric than mean in the following circumstances:

- When the distribution is skewed.
- When the distribution contains outliers



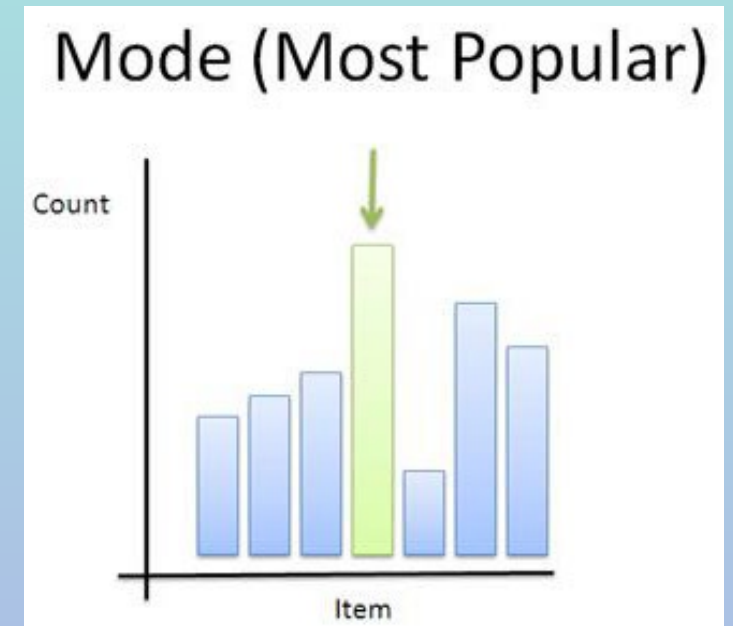
Descriptive statistics

Measure of position– Mode

The value that occurs the most often in a data set. The mode is best used when you want to indicate the most common response or item in a data set.

Why do we need mode?

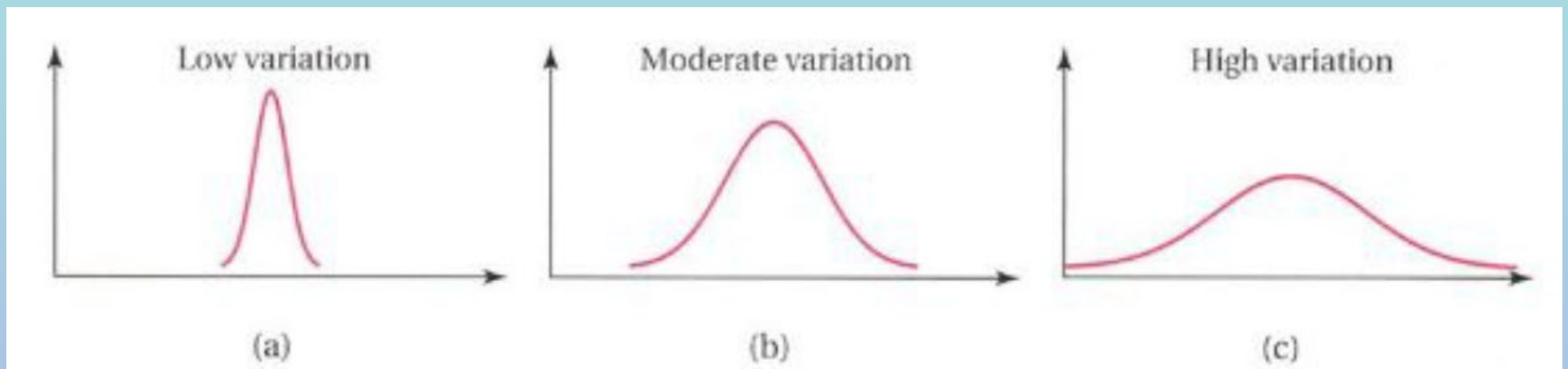
- *It can be used when the data is not numerical.*
- It is more useful to distinguish between unimodal and multimodal distributions
 - When data has more than one peak



Descriptive statistics

Measures of spread

- Measures of spread describe how similar or varied the set of observed values are for a particular variable. It gives an indication of the amount of variation in the process. It is an important indicator of quality of the obtained data.
- Range
- Standard deviation
- Variance

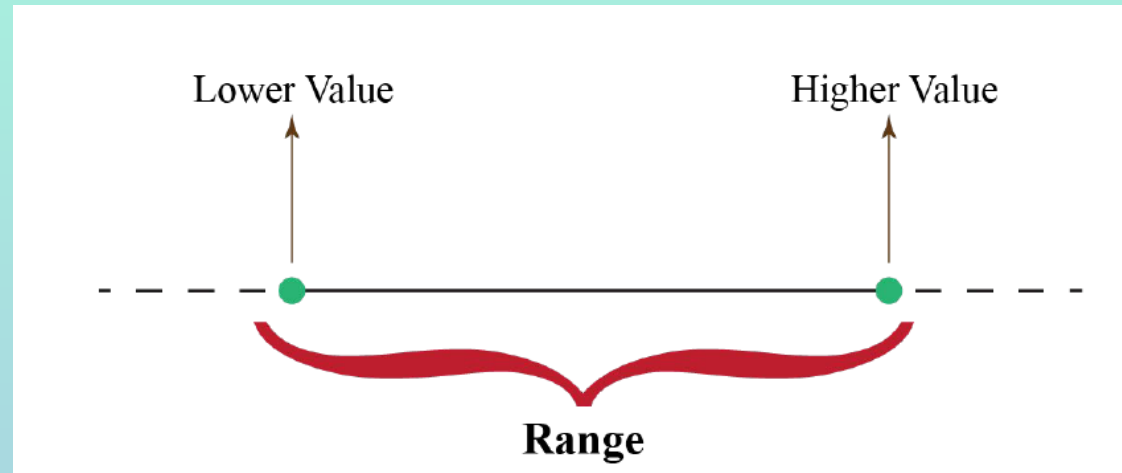


Descriptive statistics

Measures of spread - Range

The difference between the highest and the lowest value

It is the simplest measure of variability. It is often denoted by 'R'. It does not make full use of the available data, but still gives us a fair idea of spread or dispersion of the data.



It can be misleading when the data is skewed or in the presence of outlier.

Just one outlier will increase the range dramatically.

Descriptive statistics

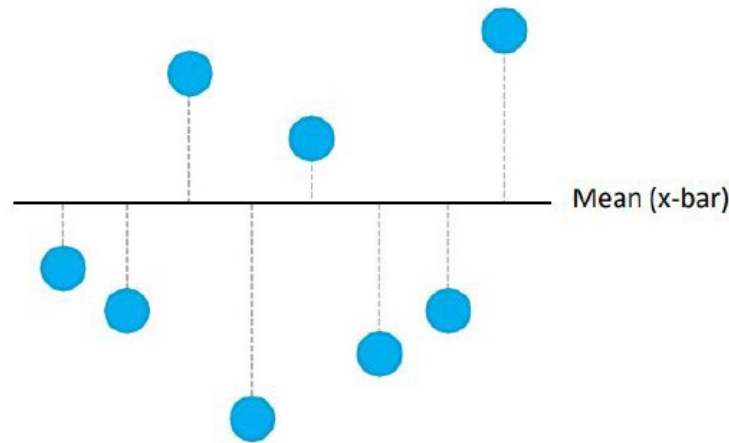
Measures of spread – Standard deviation

The average distance of the data points from their own mean.

- A low standard deviation – data points are clustered around the mean.
- A Large standard deviation – data points are widely scattered around the mean

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

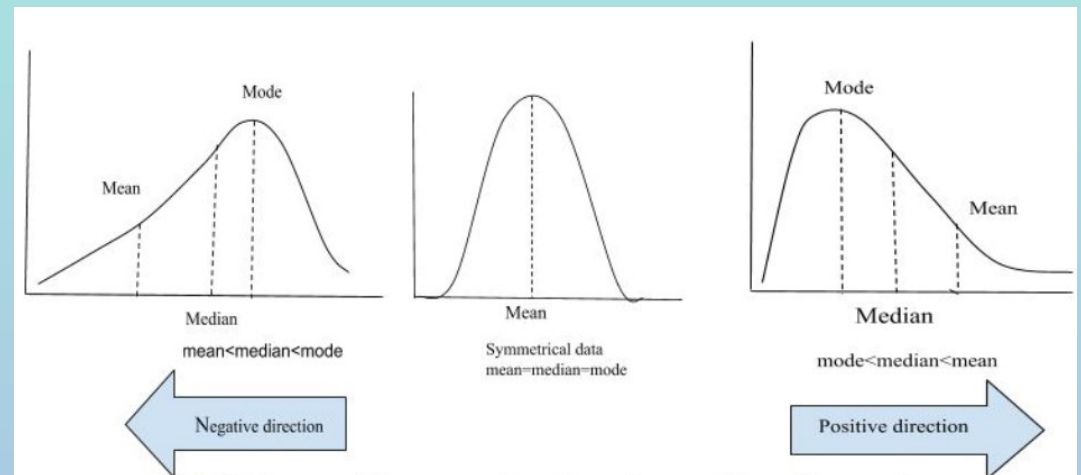
s = standard deviation
 $\bar{x}$ = mean
x = values of the data set
n = size of the data set



Descriptive statistics

Measures of shape – Skewness

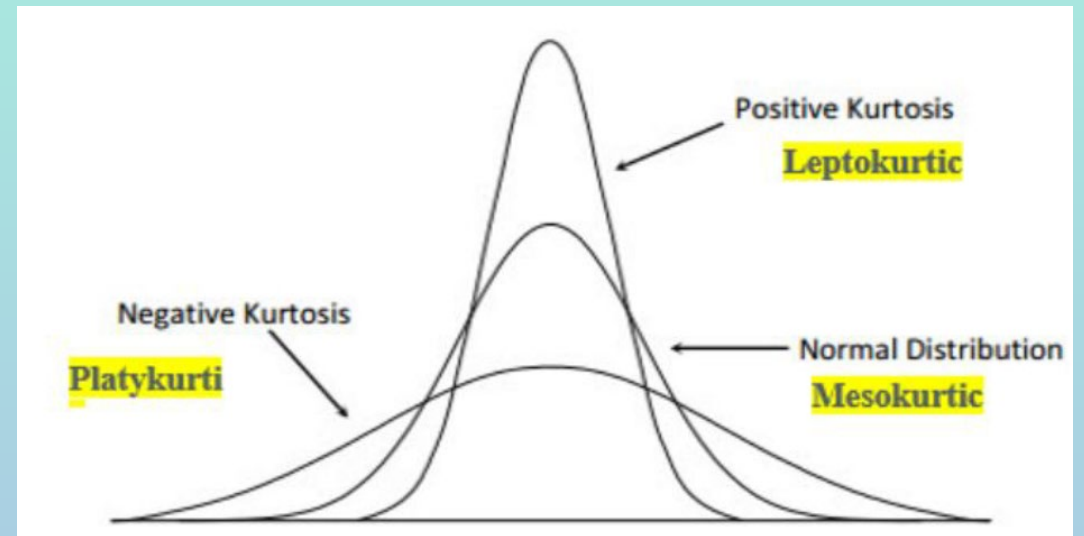
- A skewness describes whether the data is distributed symmetrically around the mean.
- A skewness value of zero indicates perfect symmetry.
- A negative value – implies left skewed data
- A positive value – implies right skewed data



Descriptive statistics

Measures of shape – Kurtosis

- This defines how heavily the tails of a distribution differ from the tails of a Symmetrical distribution. It identifies whether the tail of a distribution contains extreme values.



Probability

- Probability refers to chance or likelihood of a particular event-taking place.
- An event is an outcome of an experiment.
- An experiment is a process that is performed to understand and observe possible outcomes.
- Set of all outcomes of an experiment is called the sample space.
- Probability of an event A is defined as the ratio of two numbers m and n .
In symbols,

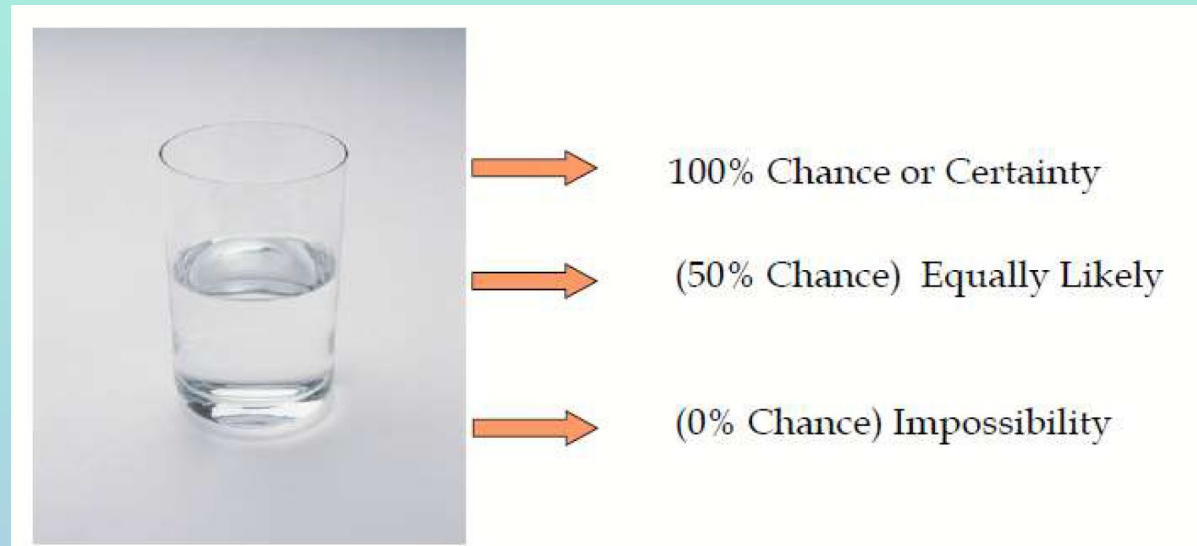
$$P(A) = \frac{m}{n}$$

- m – the number of ways that are favourable to the occurrence of A
- n - the total number of outcomes of the experiment.

Probability, $P(A)$ is always ≥ 0 and ≤ 1 .

Probability

- Example – the figure shows three stages – Empty, half-full, and full to explain the properties of probability.



Probability

- Mutually exclusive events
 - Events that cannot occur at the same time. If one event happens, the other cannot. In probability terms, for events A and B, this means: $P(A \cap B) = 0$
 - This indicates that the probability of both events occurring together is zero. For example, when flipping a coin, getting "heads" and "tails" are mutually exclusive events because both outcomes cannot happen on the same flip.

Probability

- Independent events
- Independent events are two or more events where the occurrence of one event does not affect the occurrence of the other
- Example:
 - Drawing two cards from a deck *with replacement*. After drawing a card, if it's returned to the deck before the next draw, the two draws are independent. The outcome of the first draw does not affect the probability of the second.
 - Event A: Drawing a King on the first draw.
 - Event B: Drawing a Queen on the second draw.
 - Since the card is replaced, the probability of drawing a King or Queen on either draw remains the same, and the events are independent.

Probability

- **Addition rule for mutually exclusive events**

- For two mutually exclusive events A and B (events that cannot happen simultaneously)

$$P(A \cup B) = P(A) + P(B)$$

- **Addition rule for not mutually exclusive events**

- For any two events A and B , the probability that A or B occurs is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability

- **Multiplication Rule for Independent Events**
- For two independent events A and B, the probability that both occur is: $P(A \cap B) = P(A) \times P(B)$

Probability

- Examples – Addition Rules

From a pack of well-shuffled cards, a card is picked up at random.

- 1) What is the probability that the selected card is a King or a Queen?
- 2) What is the probability that the selected card is a King or a Diamond?

- More Examples ?