Basic Statistics

- The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

Statistics: “a bunch of mathematics used to summarize, analyze, and interpret a group of numbers or observations

Independent and Dependent Variables

- Variables are defined as the properties or kinds of characteristics of certain events or objects.
  - **-Independent variables** are variables that are manipulated or are changed and whose effects are measured and compared. The other name for independent variables is Predictor(s).
  - The independent variables are called as such because independent variables predict or forecast the values of the dependent variable.
  - **-Dependent variables** refer to that type of variable that measures the affect of the independent variable(s). We can also say that the dependent variables are the types of variables that are completely dependent on the independent variable(s). The other name for the dependent variable is the Predicted variable(s).

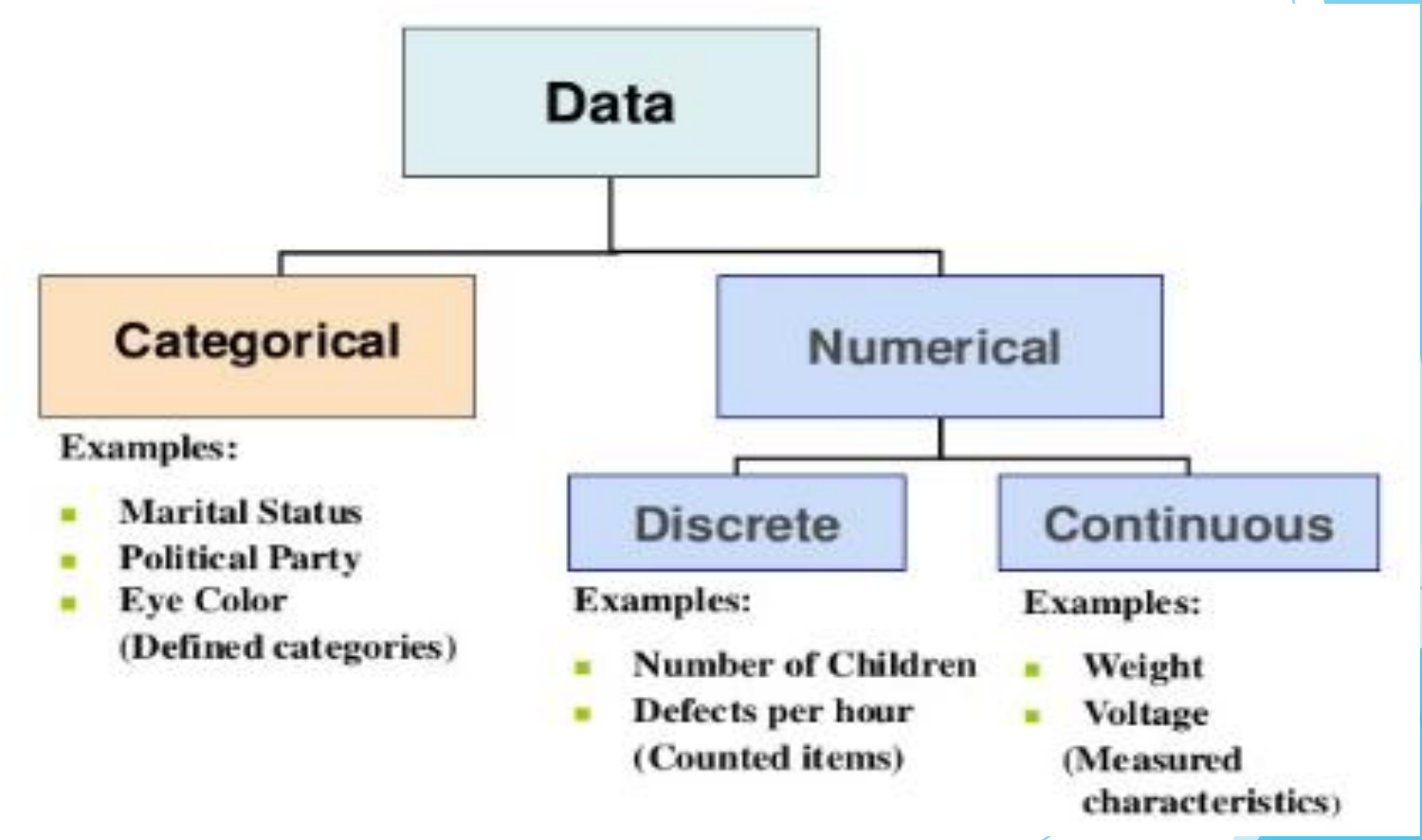
Independent and Dependent Variables contd..

- Independent variables are also called
  - controlled variable
  - explanatory variable
  - input variable
- dependent variables are also called
  - response variable,
  - measured variable,
  - observed variable,
  - outcome variable,
  - output variable.

Example 1

- Steve loves gardening. He decides to experiment to help him determine which fertilizer would be ideal for faster plant growth. He added different brands of fertilizer to different plans and observed their growth over time.
- The dependent variable Steve will be working with is the increase in the height of each plant (from each separate brand). The independent variable would be the fertilizer brand.
- Example2
  - **Response variable**: Yes or No to the Question "Do you smoke?"
  - **Explanatory variable**: Gender (Female or Male)
- Example3
  - **Response variable**: Weight
  - **Explanatory variable**: Diet

Types of variables/Data



1. Numerical Data

- Numerical data can be subdivided into two types:
  1.1)Discrete data
  Discrete data refers to the measure of things in whole numbers (integers). For example, the number of purchases made by a customer in a year. Since the number of things that a person buys cannot be three and a half, or four and a third – it must be a whole number like four or five things – this kind of data falls under the discrete category.
  1.2) Continuous data
  In contrast to discrete data, continuous data includes all numbers possible between any two integers or whole numbers. For example, the height of something. It could be 9.2345 inches or 9.7219 inches, or any other fraction between the two whole numbers nine and ten.

2.Categorical/Nominal Data

- This type of data is non-numeric.
- We use it to quantify things in categories like gender, ethnicity, nationality, political party, etc.
- Gender ---Male ,Female
- Country --- US,UK,UAE
- We can assign numbers to the categories, but the numbers would not, in that case, represent their value per say.
- They will only separate one type from the other – type one from type two or three.
  - For example, while calculating India's population, Bangalore could be city number one, Mumbai number two, and so on

3. Ordinal Data

- Ordinal data is an amalgamation of numerical and categorical data. Simply put, this data type consists of categories that are in order.
- The intervals between categories are not known.
- movie or music ratings that use stars to denote quality.
- Numbers simply represent the good and bad categories.
- A movie with a 5-star rating is obviously very good as opposed to a movie with only 1-star, which, very likely, is terrible.
- Note that the numbers in this example do denote value.
- Mathematically speaking, 5 is greater than 1. This difference in value is used to differentiate good films from bad. Good films receive a higher rating of 4 or 5, while bad films only get a lower rating of 1 or 2.

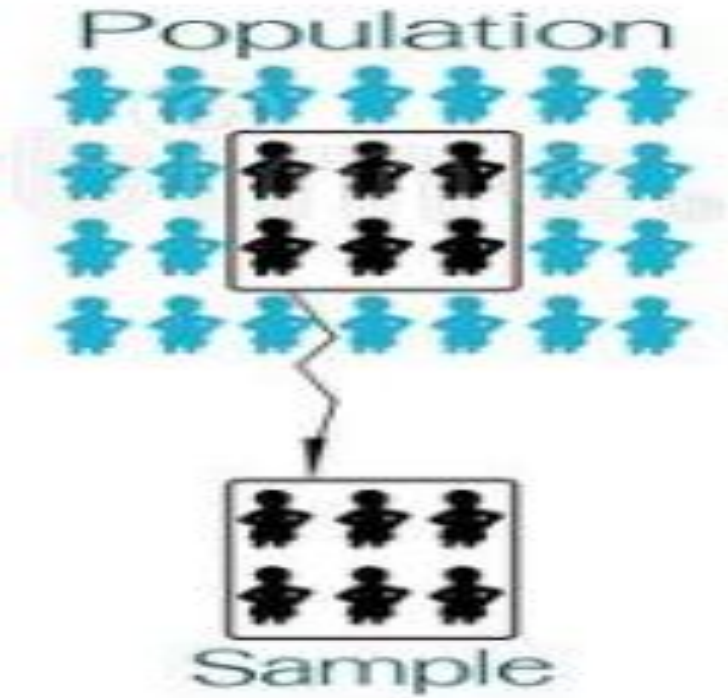
Interval Data

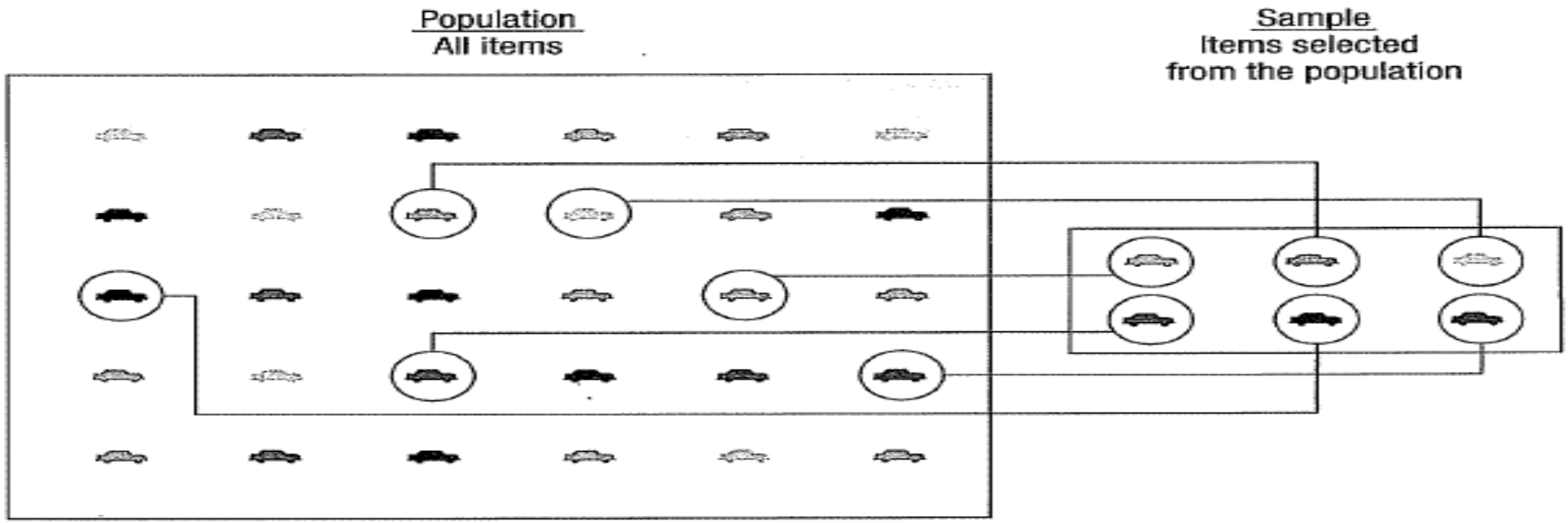
- Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth

Population: any group of interest or any group that researchers want to learn more about.

–Population parameters (unknown to us): characteristics of population

• Sample: a group of individuals or data are

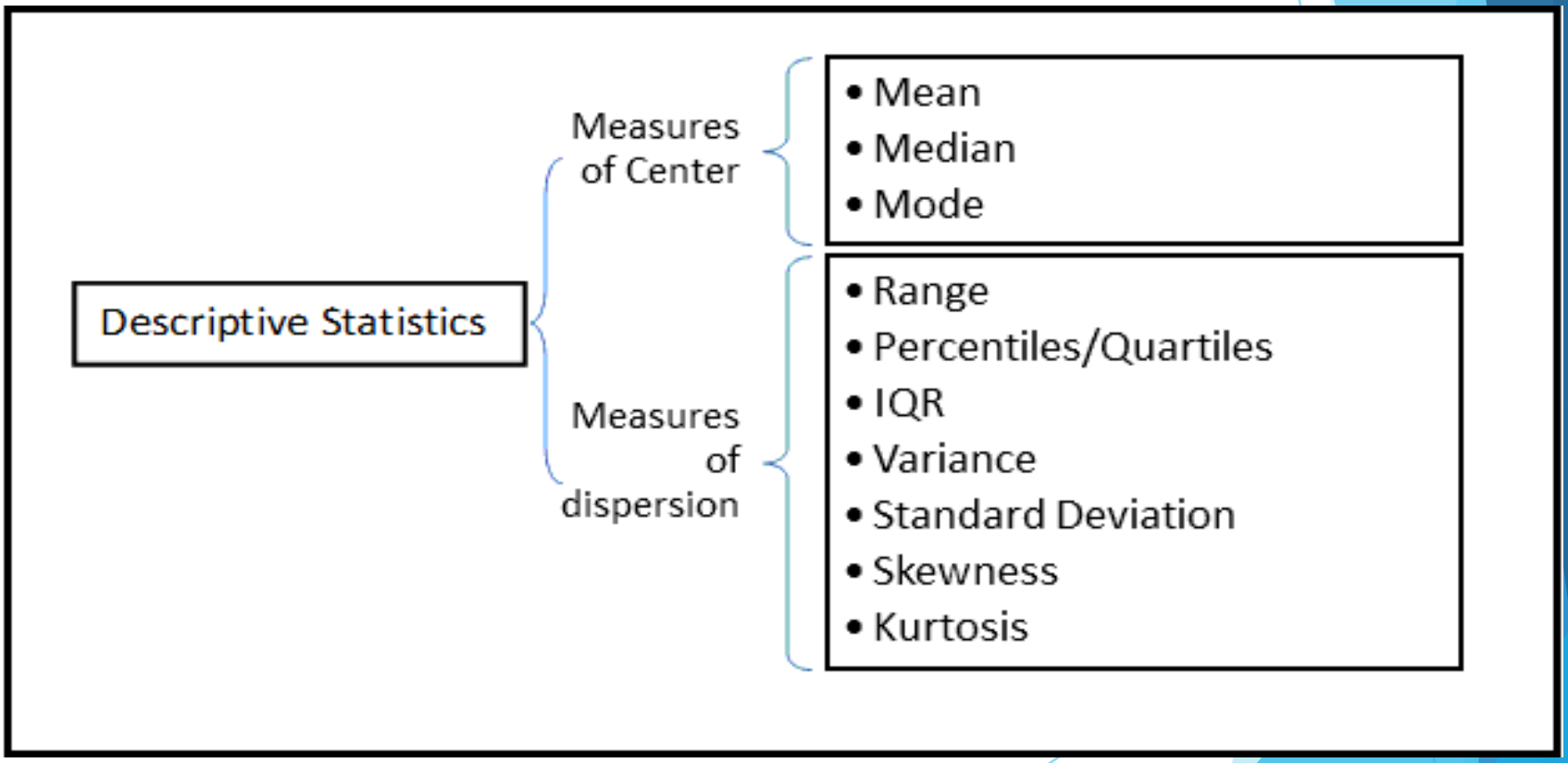




Types of statistics

- **Descriptive statistics** - are procedures used to summarize, organize, and make sense of a set of scores or observations."
- **Inferential statistics** – procedures used that allow researchers to infer or generalize ,observations made with samples to the larger population from which they were selected."

Descriptive statistics contd……



Mean

- …the average of the given values. To compute mean, sum all the values and divide the sum by the number of values.

$$\text{MEAN} = \bar{X} = \frac{\sum_{i}^{n} X_i}{n}$$

Median:

- The median is the middle value for a dataset that has been arranged in order of magnitude.
- If the total number of values is odd then->

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

If the total number of values is even then

$$\text{Median} = \left(\frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2}\right)^{th} term$$

Median = (1 + 2) / 2
= 1.5
Median is 1.5.

Mode

The value in a data set that occurs most often or most frequently.

- Example: 2,3,3,3,4,4,4,4,7,7,8,8,8

  Mode=4

Descriptive statistics

- Dispersion

–Range –

largest value and smallest value.

–Variance

–Standard deviation

Variance

- Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.
- Variance is the average of all squared deviations.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \text{ for populations}$$

Standard deviation is…

- …the square root of variance.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set,
N = number of values in the data set.

For eg: {3,5,6,9,10} are the values in a dataset.

$$\text{Mean} = \frac{3+5+6+9+10}{5} = 6.6$$

$$\text{Variance} = \frac{(3-6.6)^2+(5-6.6)^2+(6-6.6)^2+(9-6.6)^2+(10-6.6)^2}{5}$$

$$= \frac{12.96+2.56+0.36+5.76+11.56}{5} = \frac{33.2}{5} = 6.64$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{6.64} = 2.576$$

Percentile

- In statistics, Percentile is used to indicate the value below which the group of percentage of data falls below. For example, consider if your score is 75th percentile, which you scored far better than 75% of people who took part in the test.
- It is most commonly applicable in indicating the scores from the norm-referenced tests such as SAT, GRE and LSAT.

Percentile Example 1:

Learn how to calculate percentile for the given example:

There are 25 test scores such as:

72,54, 56, 61, 62, 66, 68, 43, 69, 69, 70, 71,77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.

Find the 60th percentile?

Solution:

Step 1: Arrange the data in the ascending order.

Ascending Order = 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.

Step 2: Find Rank, Rank = Percentile / 100 = 60 / 100 k = 0.60

Step 3: Find 60th percentile, 60th percentile = 0.60 x 25 = 15

Step 4: Count the values in the given data set from left to right until you reach the number 15.

From the given data set, 15th number is 79.

Now take the 15th number and the 16th number and find the average:

79 + 85 / 2 = 164 / 2 = 82

Hence, 60th percentile of given data set = 82.

Probability

- In real-life situations what the probabilities are of some event occurring, such as winning the lottery, the victory of your soccer team or a discount on your favorite pair of shoes. "What are the chances..." is an expression you probably use very often. Determining the chances of an event occurring is called "probability".
- **Probability is a measure of uncertainty of various phenomena.**
- **Probability theory is mainly concerned with predicting the likelihood of future events**

- **Examples**
  - Flipping coins
    - Head /Tail - ½ -.5 or 50 percent chance
  - Rolling dice

    1/6
  - Deck of cards -- 4 Aces 52 cards
    - 4 /52 = .08 -- > 8 % chance

What is a random variable?

- Suppose we flip a fair coin three times and record if it shows a head or a tail. The outcome or sample space is S={HHH,HHT,HTH,THH,TTT,TTH,THT,HTT}. There are eight possible outcomes and each of the outcomes is equally likely. Now, suppose we flipped a fair coin four times. How many possible outcomes are there? There are 24=16. How about ten times? 1024 possible outcomes! Instead of considering all the possible outcomes, we can consider assigning the variable X, say, to be the number of heads in n flips of a fair coin. If we flipped the coin n=3 times (as above), then X can take on possible values of 0,1,2, or 3. By defining the variable, X, as we have, we created a random variable.

Random Variable

- A **random variable** is a variable that takes on different values determined by chance. In other words, it is a numerical quantity that varies at random.

Types of Random Variables

- **Discrete Random Variable**
  - When the random variable can assume only a countable, sometimes infinite, number of values.
- **Continuous Random Variable**
  - When the random variable can assume an uncountable number of values in a line interval.

Probability functions

- **Probability Mass Function (PMF)**
  - If the random variable is a discrete random variable, the probability function is usually called the probability mass function (PMF)
- **Probability Density Function (PDF)**
  - If the random variable is a **continuous random variable**, the probability function is usually called the **probability density function (PDF).**

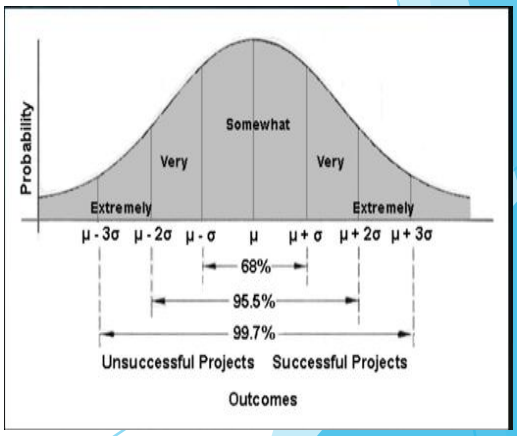
Independent versus Dependent Events

- **Independent events**
  - are events that don't impact the probability of the other event(s).
- **Dependent events,**
  - are events that have an impact on the probability of the other event(s).

Types of Distribution

- Discrete data –
  - Binomial distribution
  - Poisson distribution
    - Exponential distribution
      - **Exponential growth (e.g. prices, incomes, populations)**

- Continuous data –
  - Normal distribution
    - Gaussian distribution/normal curve
      - the amount of rainfall in inches in a year for a city.
      - the weight of a newborn baby.
      - the height of a randomly selected student.



Types of Probability Distributions

- The **binomial distribution** evaluates the probability of an event occurring several times over a given number of trials and given the event's probability in each trial.
- Use a fair coin and figuring the probability of that coin coming up heads in 10 straight flips. A binomial distribution is *discrete*, as opposed to continuous, since only 1 or 0 is a valid response.

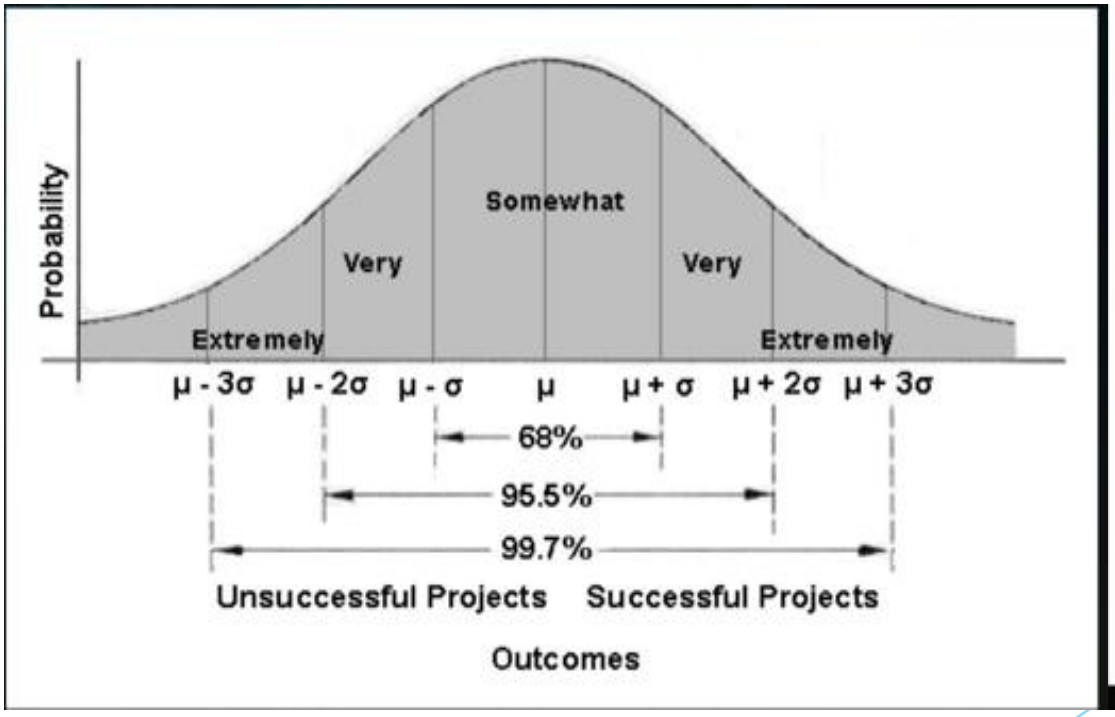
- The **normal distribution** is fully characterized by its mean and standard deviation. This makes the distribution symmetric and it is depicted as a bell-shaped curve when plotted. A normal distribution is defined by a mean (average) of zero and a standard deviation of 1.0.
- In a normal distribution, approximately 68% of the data collected will fall within +/- one standard deviation of the mean; approximately 95% within +/- two standard deviations; and 99.7% within three standard deviations.

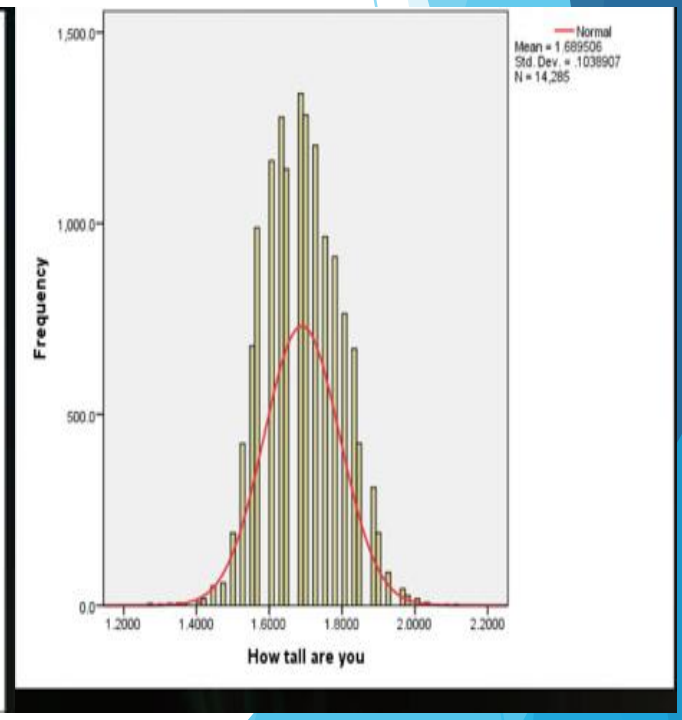
Poisson probability distribution

- **Poisson probability distribution** is a discrete probability distribution that represents the probability of a given number of events happening in a fixed time or space if these cases occur with a known steady rate and individually of the time since the last event. The Poisson distribution can also be practised for the number of events happening in other particularised intervals such as distance, area or volume. Some of the real-life examples are:
  - A number of patients arriving at a clinic between 10 to 11 AM.
  - The number of emails received by a manager between the office hours.
  - The number of apples sold by a shopkeeper in the time period of 12 pm to 4 pm daily.

Normal distribution/Normal curve

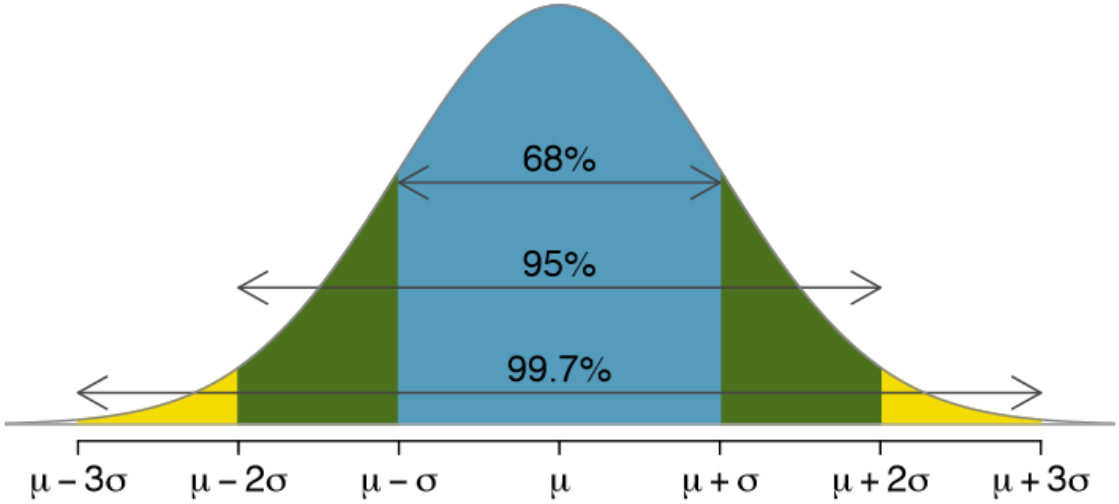
Data are symmetrically distributed around mean, median, and mode or bell-shaped distribution. The form of a normal distribution is determined by its mean and standard deviation. Mean =0 and Standard deviation=1

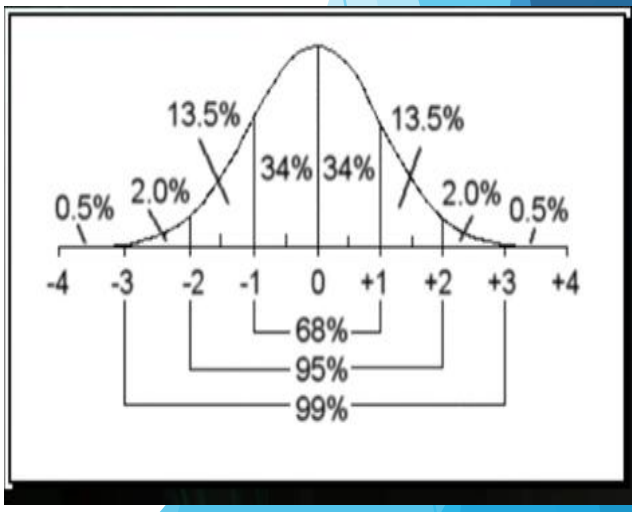




Three Sigma Rule/Empirical rule 68-95-99.7 rule

- It is an expression of how many of our observations fall within a certain distance of the mean. The standard deviation (a.k.a. “sigma”) is the average distance an observation in the data set is from the mean.
- The Three Sigma rule dictates that **given a normal distribution**, 68% of your observations will fall between one standard deviation of the mean. 95% will fall within two, and 99.7% will fall within three. Less than 5% scores are far from the mean.(Not normal scores).

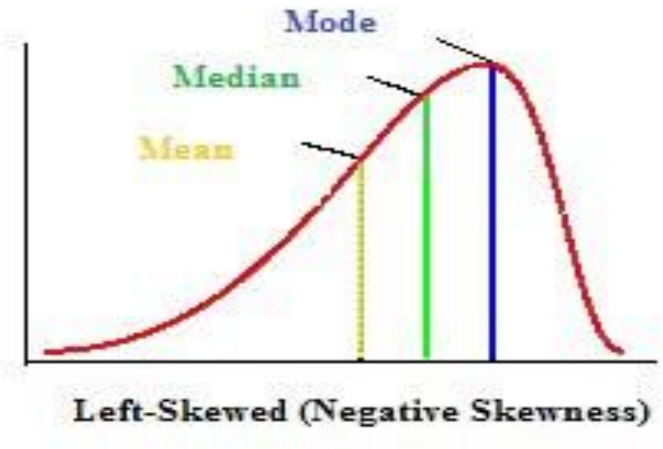




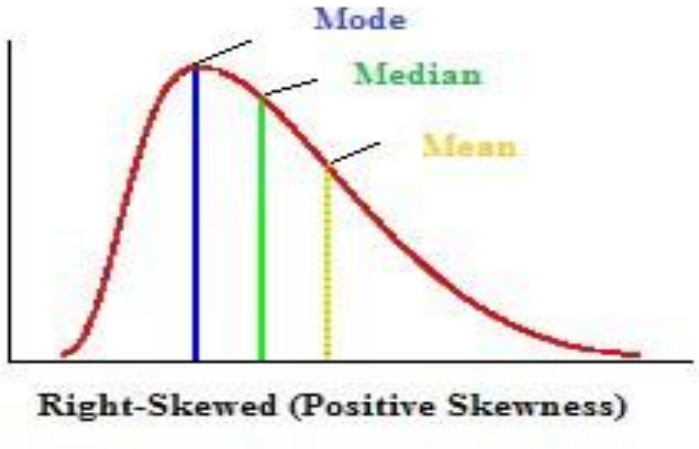
Measures of Asymmetry -

Exponential distribution

- **Skewness:** Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right.
- Skewness indicates whether the data is concentrated on one side.



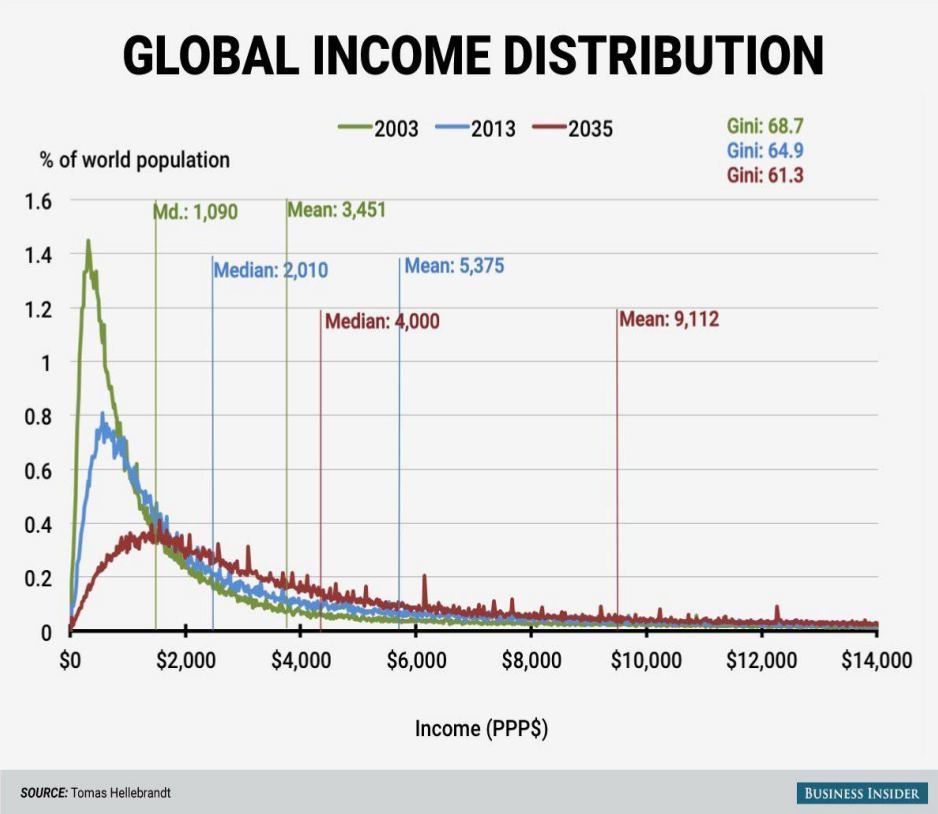
Left-Skewed (Negative Skewness)

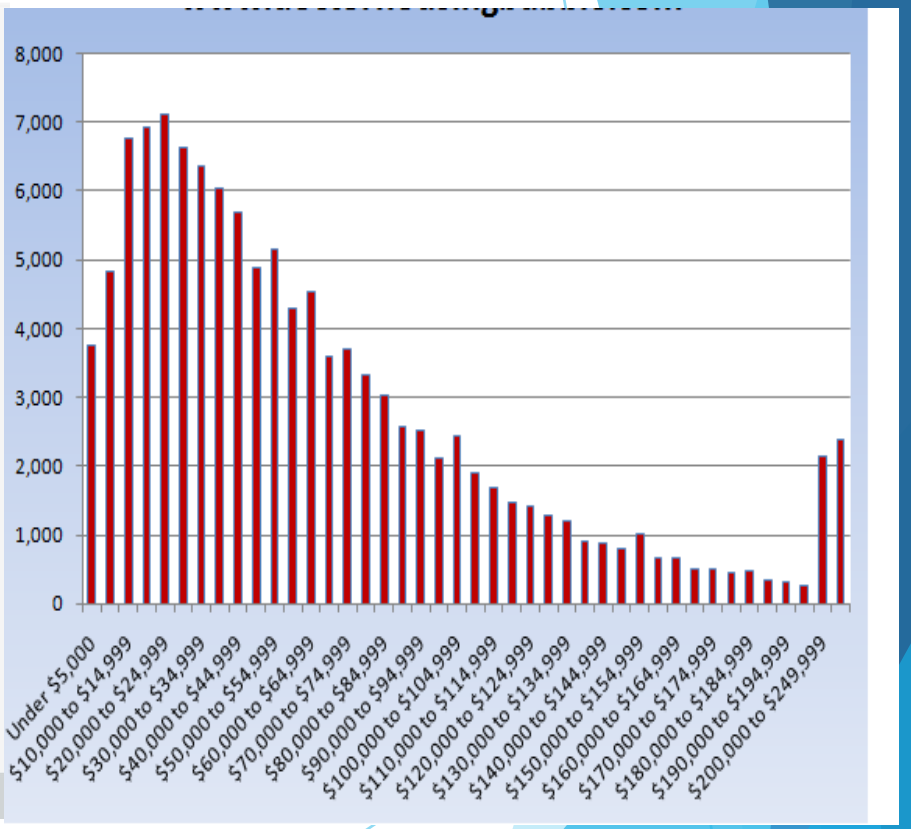


Right-Skewed (Positive Skewness)

For eg: Global Income Distribution in 2003 is highly right-skewed. We can see the mean $3,451 in 2003(green) is greater than the median $1,090. It suggests that the global income is not evenly distributed. Most individuals

incomes are less than $2,000





Box Plot

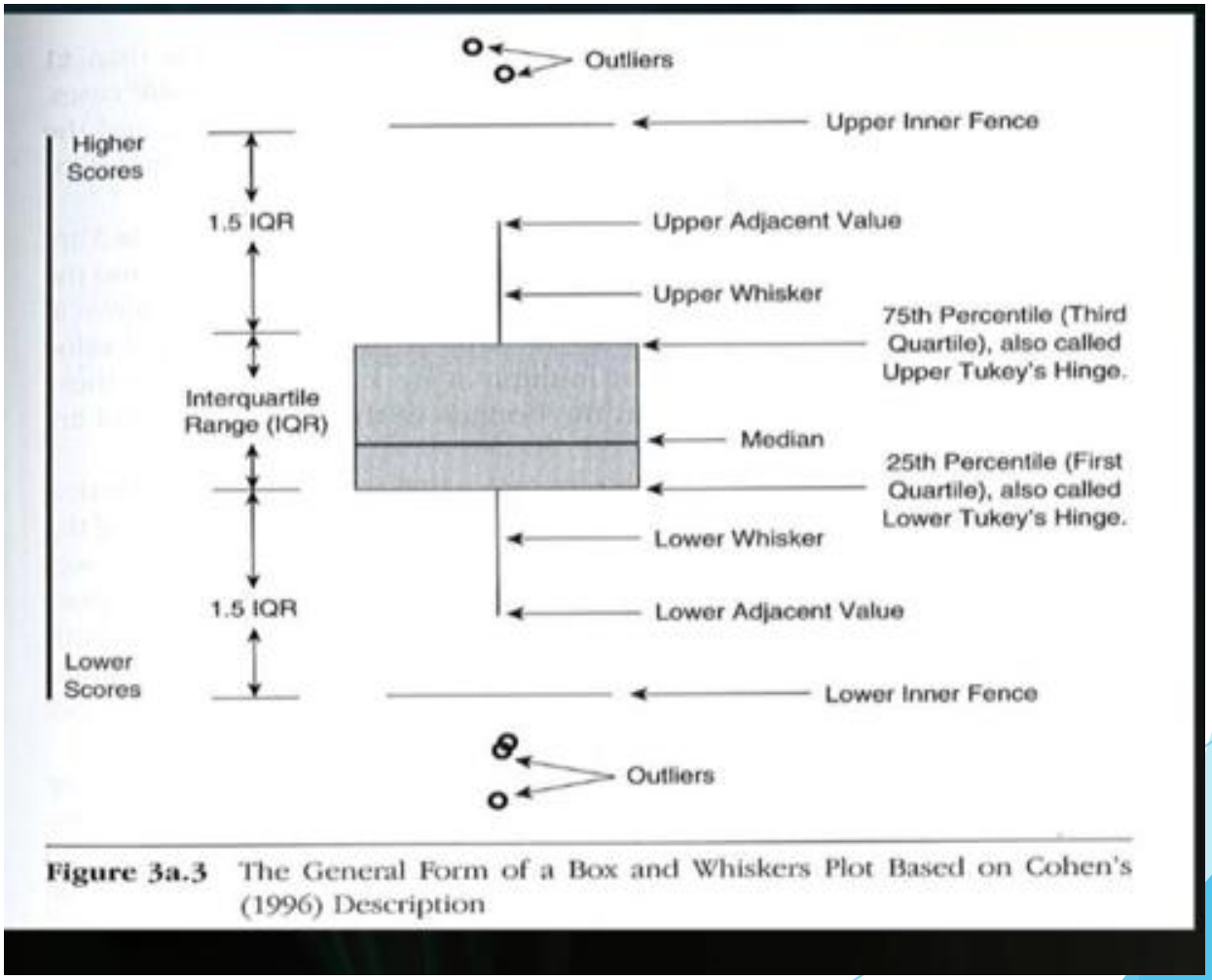


Figure 3a.3 The General Form of a Box and Whiskers Plot Based on Cohen's (1996) Description

Inferential Statistics contd ….