



Statistics foundation for AI & ML

By AI&ML Community

AGENDA

- ▶ Descriptive Statistics
- ▶ Exploratory Data Analysis
- ▶ Inferential statistics

AGENDA

Descriptive Statistics & Exploratory Data Analysis

- ▶ Frequency Distribution - Histograms
- ▶ Cumulative Frequency Distribution
- ▶ Measures of Central Tendency - Mean Median Mode
- ▶ Measures of dispersion - Range, IQR , standard deviation, Coefficient of variation
- ▶ Normal Distribution , Empirical Rule,
- ▶ Five number summary , boxplots , scatter plt & How outliers can be identified, Pair plots and inferences
- ▶ Correlation Analysis (positive correlation or negative correlation with help of scatters etc)
- ▶ ROC curve
- ▶ Univariate and multivariate analysis

AGENDA

Inferential statistics

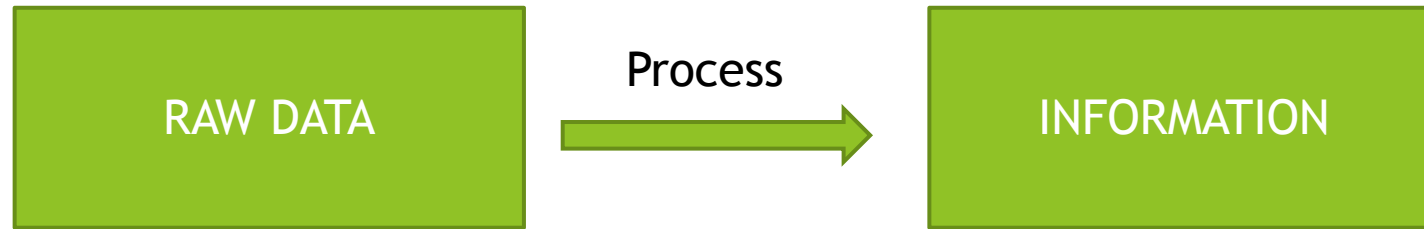
- ▶ Probability & Distributions
- ▶ Types of Probability
- ▶ Bayes theorem and its applications
- ▶ Hypothesis Testing
 - ▶ Null and Alternate Hypothesis
 - ▶ Hypothesis tests - Pvalues for level of significance , One tailed and 2 tailed, Paired T Test
 - ▶ ANOVA - One sample and 2 sample testing
 - ▶ Chi Square

Types of Statistics

- ▶ Descriptive Statistics - is concerned with Data summarization, Graphs / Charts and tables
- ▶ Inferential Statistics - is concerned with drawing conclusions about a population from a sample

Raw Data

- ▶ Raw Data represent numbers and facts in the original format in which the data have been Collected. You need to convert the raw data into information for managerial decision Making



Types of Data

- ▶ Categorical (Qualitative)

- ▶ E.g. Gender, brand name of TV owned, shirt size preferred (XS,S,M,L,XL etc)

- ▶ Numeric (Quantitative)

- E.g. Age, Height, Number of room at residence

- Depending on the values a variable can take, Numeric data type is further divided into

- Discrete

- Continuous

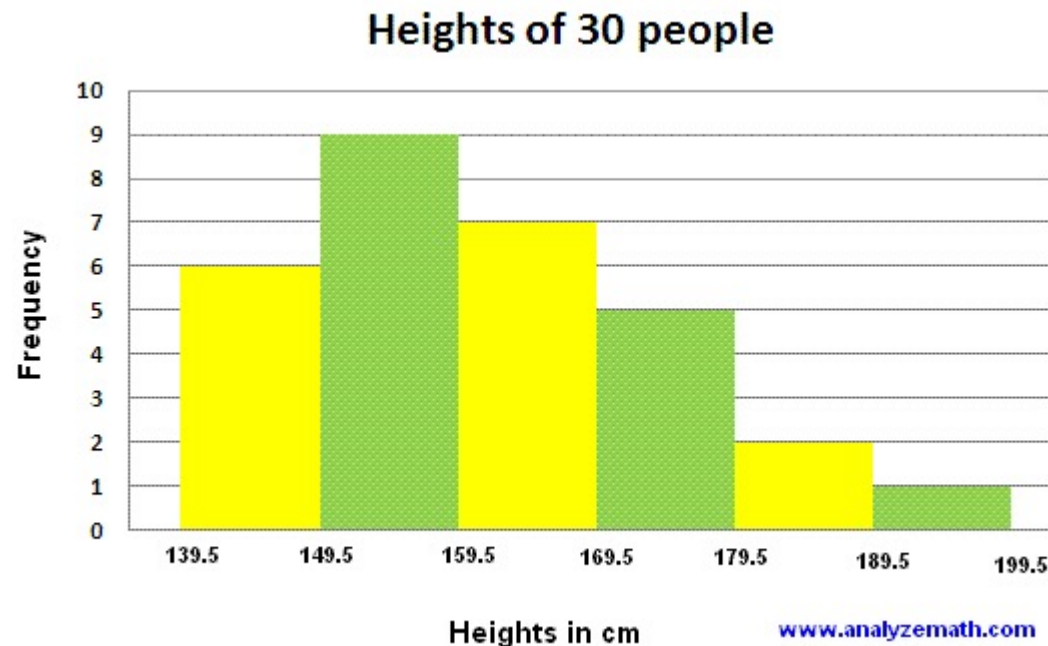
- ▶ Population: A population dataset contains ALL observation of characteristic under study

- ▶ Sample: Sample dataset is subset of a population

Frequency Distribution -(Univariate)

- ▶ In simple terms, frequency distribution is a summarized table in which raw data are arranged into classes and frequencies.
- ▶ Frequency distribution focuses on classifying raw data into information. It is the most widely used technique in descriptive statistics.

Sample heights Data { 140,139.5,142,143,141,146,----- }



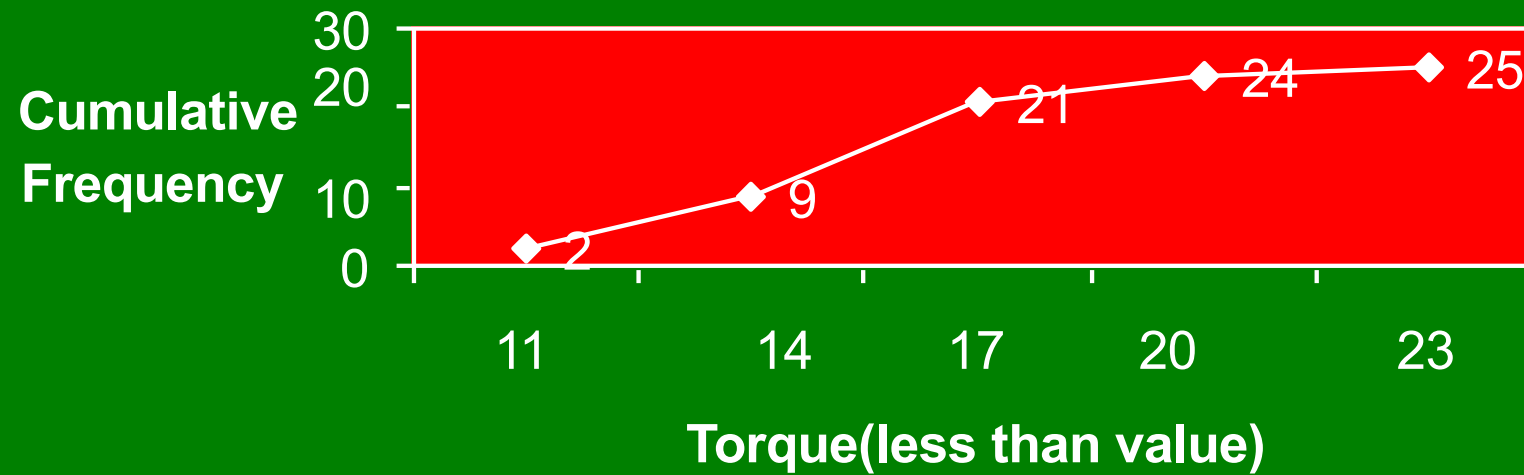
Cumulative Frequency Distribution

- A type of frequency distribution that shows how many observations are above or below the lower boundaries of the classes. You can formulate the following from the previous example of hose clamping force(torque).

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
8-11	2	0.08	2	0.08
11-14	7	0.28	9	0.36
14-17	12	0.48	21	0.84
17-20	3	0.12	24	0.96
20-23	1	0.04	25	1.00
Total	25	1.00		

Cumulative Distribution Function

Cumulative Distribution (Ogive Curve) for the Example



Central Tendency

- ▶ Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of "Central Tendency". The other terms that are used synonymously are "Measures of Location", or "Statistical Averages".
- ▶ Typically used to get the first impression / initial understanding of data
- ▶ Commonly used measures of central tendency are Arithmetic Mean, Median and Mode

Arithmetic Mean

Arithmetic Mean (called mean) is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:

In symbolic form mean is given by $\bar{X} = \frac{\sum X}{n}$

\bar{X} = Arithmetic Mean

$\sum X$ = Indicates sum all X values in the data set

n = Total number of observations(Sample Size)

Median

Median is the middle most observation when you arrange data in ascending order of magnitude. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.

Median is a very useful measure for ranked data in the context of consumer preferences and rating. It is not affected by extreme values (greater resistance to outliers)

Median = $\frac{n+1}{2}$ th value of ranked data

n = Number of observations in the sample

3, 5, 12

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

Mode

- ▶ Mode is the value that occurs most often. It has the maximum frequency of occurrence. Mode has resistance for outliers
- ▶ Example : Mode is a very useful measure when you want to keep in the inventory, the most popular shirt in terms of collar size during festive season.

3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29, 5

Arranged in order

3, 5, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56

Measures of Dispersion / Variation

- ▶ In simple terms, measures of dispersion indicate how large the spread of the distribution is around the central tendency.
- ▶ Range
 - ▶ Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.

$$\text{Range} = X_{\text{Maximum}} - X_{\text{Minimum}}$$

Example : 12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 8

$$\text{Range} = 18 - 8 = 10$$

There are several measures that help estimate the spread of observations around the center (i.e., the mean).

Inter-Quartile Range (IQR)

- ▶ IQR= Range computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. IQR is less affected by outliers.
- ▶ $IQR = Q3 - Q1$
 - ▶ Note: arrange the data in ascending order

Variance and Standard Deviation

- ▶ To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance. It is a measure of how spread out a data set is
- ▶ Example : The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate the sample standard deviation.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

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

Xbar is Mean , N is sample

Standard Deviation Calculation

- Consider observations: 20, 25, 18, 15, 21, 22, 19
- Arithmetic Mean = 20

X	X - XBar	(X - XBar)^2
20	0	0
25	5	25
18	-2	4
15	-5	25
21	1	1
22	2	4
19	-1	1
Total		60

Variance = 10

Standard Deviation = 3.16

Coefficient of Variation (Relative Dispersion)

- Coefficient of Variation (CV) is defined as the ratio of Standard Deviation to Mean.

In symbolic form

$$CV = \frac{S}{\bar{X}} \text{ for the sample data and } = \frac{\sigma}{\mu} \text{ for the population}$$

Example of CV

Consider two Sales Persons working in the same territory

The sales performance of these two in the context of selling PCs are given below. Comment on the results.

Sales Person 1

Mean Sales (One year average)

50 units

Standard Deviation

5 units

Sales Person 2

Mean Sales (One year average)

75 units

Standard deviation

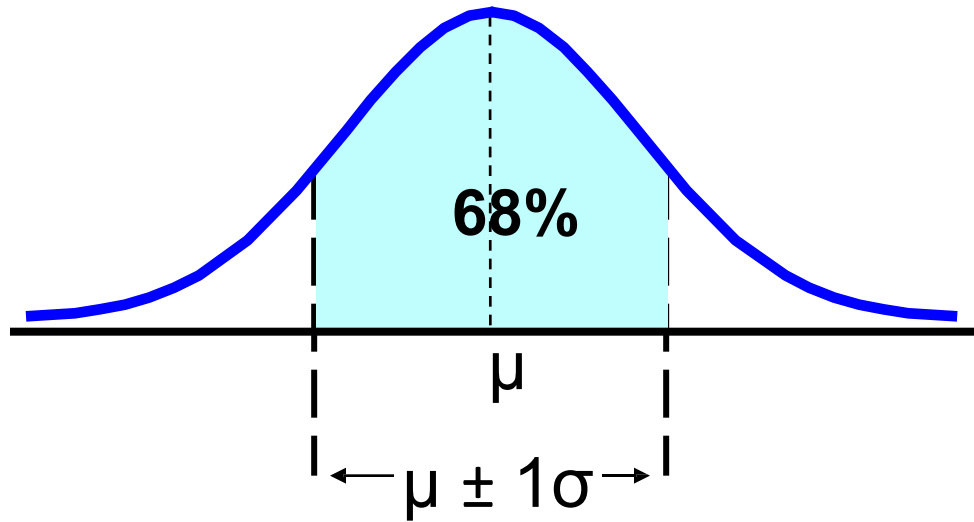
25 units

The CV is $5/50 = 0.10$ or 10% for the Sales Person1
and $25/75 = 0.33$ or 33% for sales Person2.

The moral of the story is "don't get carried away by averages. Consider variation ("risk").

The Empirical Rule

- ▶ The empirical rule approximates the variation of data in bell-shaped distribution.
- ▶ Example : Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$



Symmetric distribution - > No Skew

Five number summary, Box and other plots

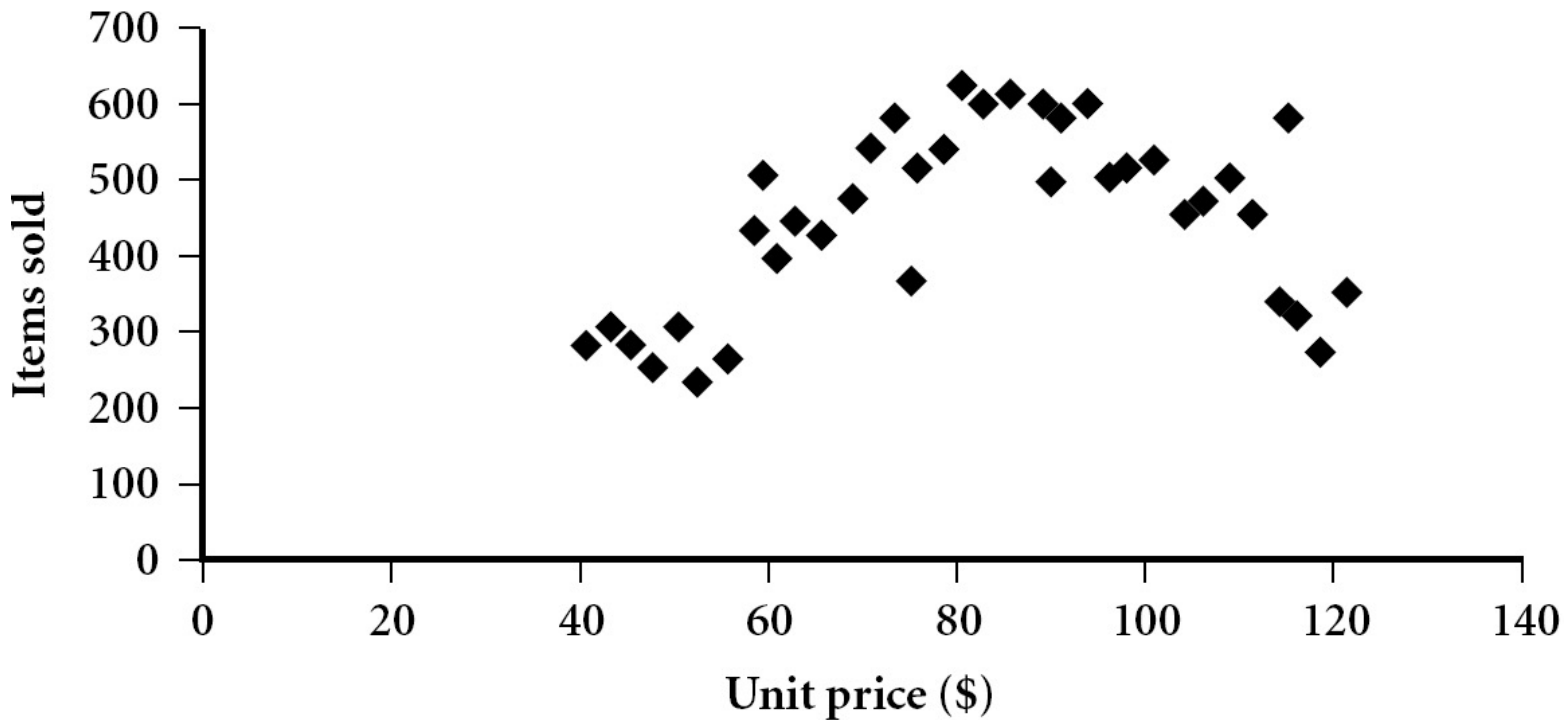
The five numbers that help describe the center, spread and shape of data are:

- X_{smallest}
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- X_{largest}

- ▶ IQR - Inter Quartile Range
- ▶ $\text{IQR} = Q_3 - Q_1$ (resistant to outliers)

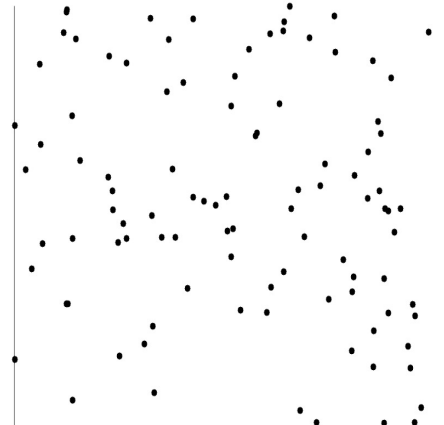
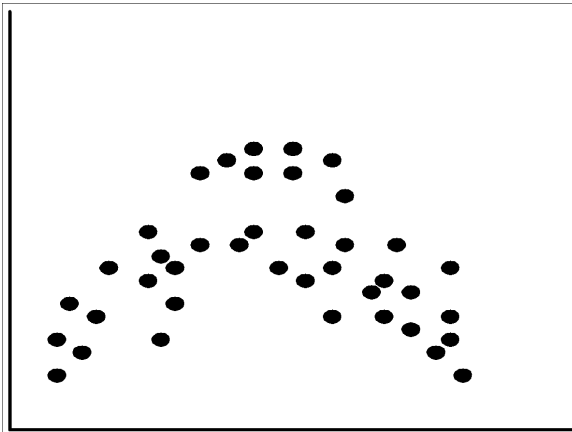
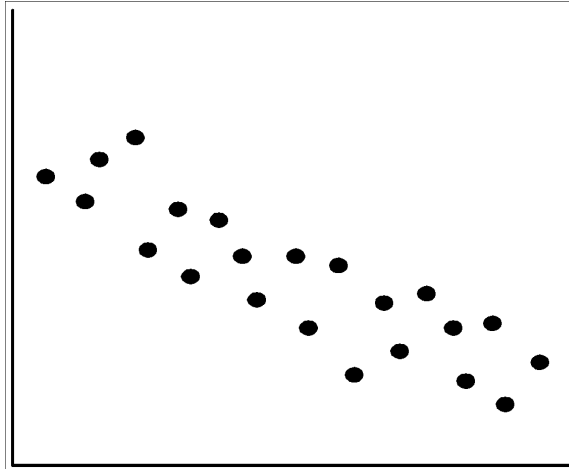
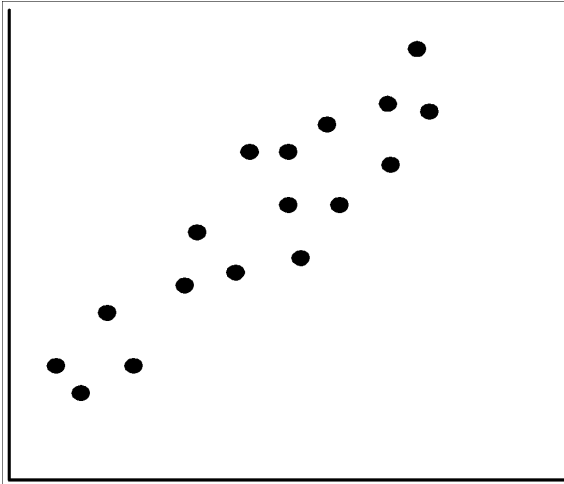
Scatter Plot

- Provides a first look of BIVARIATE data to see clusters of points, outliers etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



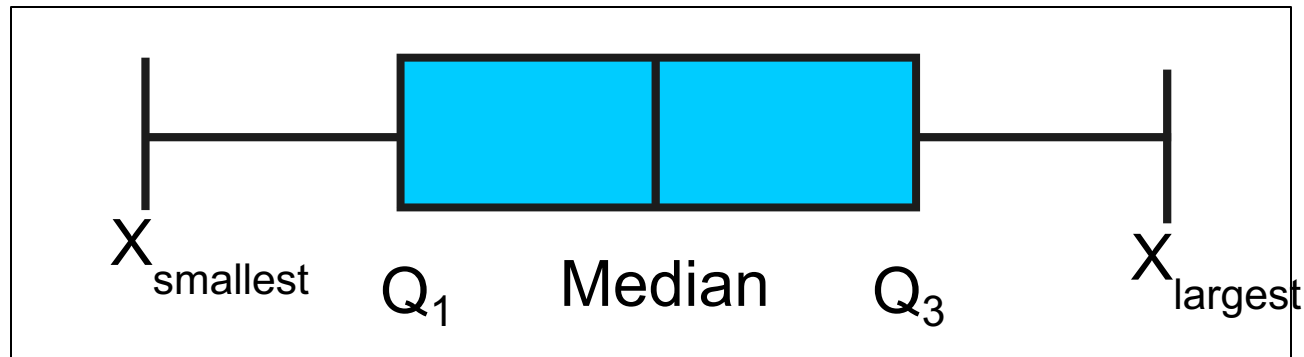
Correlations with Scatter Plots

(Relationship between 2 variables)

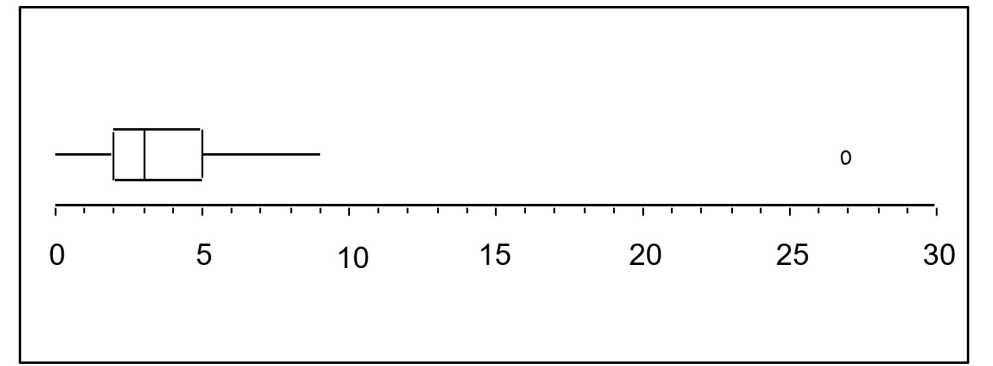


Whisker Box Plots

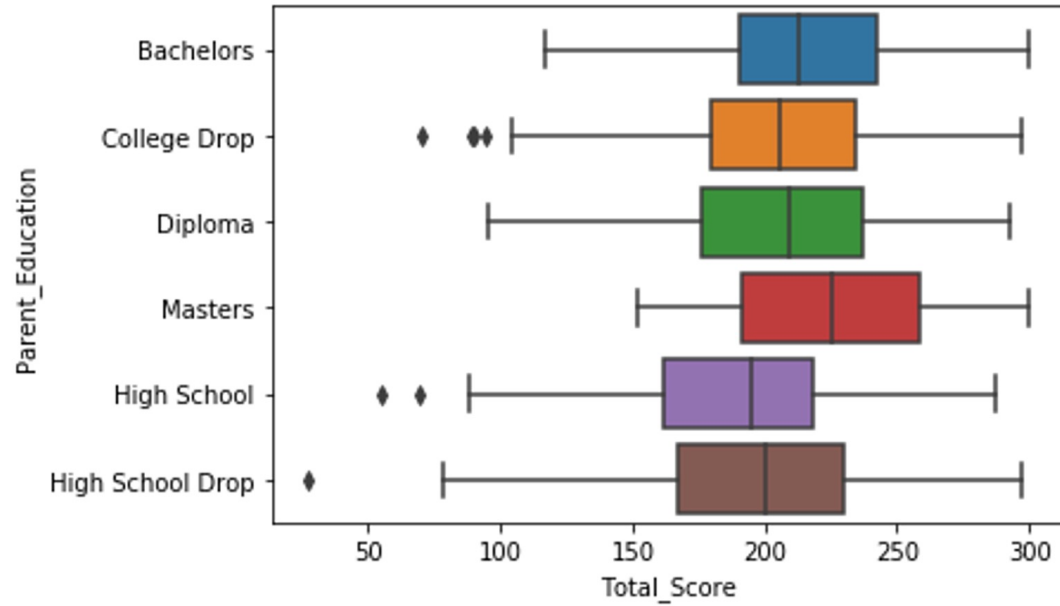
- ▶ The Boxplot : A graphical display of data based on five-number summary
- ▶ If data is symmetric around the median then the box and central line are centered between the end points
- ▶ A boxplot can be shown in either a vertical or horizontal orientation



Boxplot cont



Boxplot of Total score
in an exam Vs
Education level of
parents



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.

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

- ▶ where m = number of ways that are favorable to the occurrence of A
- ▶ and n = the total number of outcomes of the experiment (all possible outcomes)

Probability

Emperical Probability

- ▶ Based On actual evidence
- ▶ These are based upon how likely an event has proven in the past. Thus, they are always estimates.

Example : Cricket India - WWWLLLWWWW , $P(W) = 7/10 = 0.7$

Mutually Exclusive Events

Two events A and B are said to be mutually exclusive if the occurrence of A precludes the occurrence of B.

Example : Pickup a card from the shuffled pack of cards , of you pickup one card at random and would like to know whether it is king or Queen. It cannot be both King or Queen

Toss a coin - Can be only head or tails

Roll a dice - can be only one number at once

Probability

Independent Events

Two events A and B are said to be independent if the occurrence of A is in no way influenced by the occurrence of B. Likewise occurrence of B is in no way influenced by the occurrence of A.

Example - rolling 2 dies

Conditional Probability (Based on Dependent Events)

is the probability of one event occurring with some relationship to one or more other events.

$P(A|B)$ => Probability of A given B

Everything in ML is conditional probability

$$P(B|A) = P(A \text{ and } B) / P(A)$$

which can also rewrite as: $P(B|A) = P(A \cap B) / P(A)$

Rules for computing probability


$P(A \cup B) = P(A) + P(B)$ - Mutually exclusive events

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (- Events are not mutually exclusive)

Probability

Marginal Probability

The term marginal is used to indicate that the probabilities are calculated using a contingency table (also called joint probability table).

Family 	Income below Rs 10 Lakhs	Income of Rs. ≥ 10 lakhs	Total
Buyer of Car	38	42	80
Non-Buyer	82	38	120
Total	120	80	200

Probability

What is the probability that a randomly selected family is a buyer of the Car?

What is the probability that a randomly selected family is both a buyer of car and belonging to income of Rs. 10 lakhs and above?

A family selected at random is found to be belonging to income of Rs 10 lakhs and above. What is the probability that this family is buyer of car?

Bayes' Theorem / Bayes' Formula

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayes' Theorem / Bayes' Formula

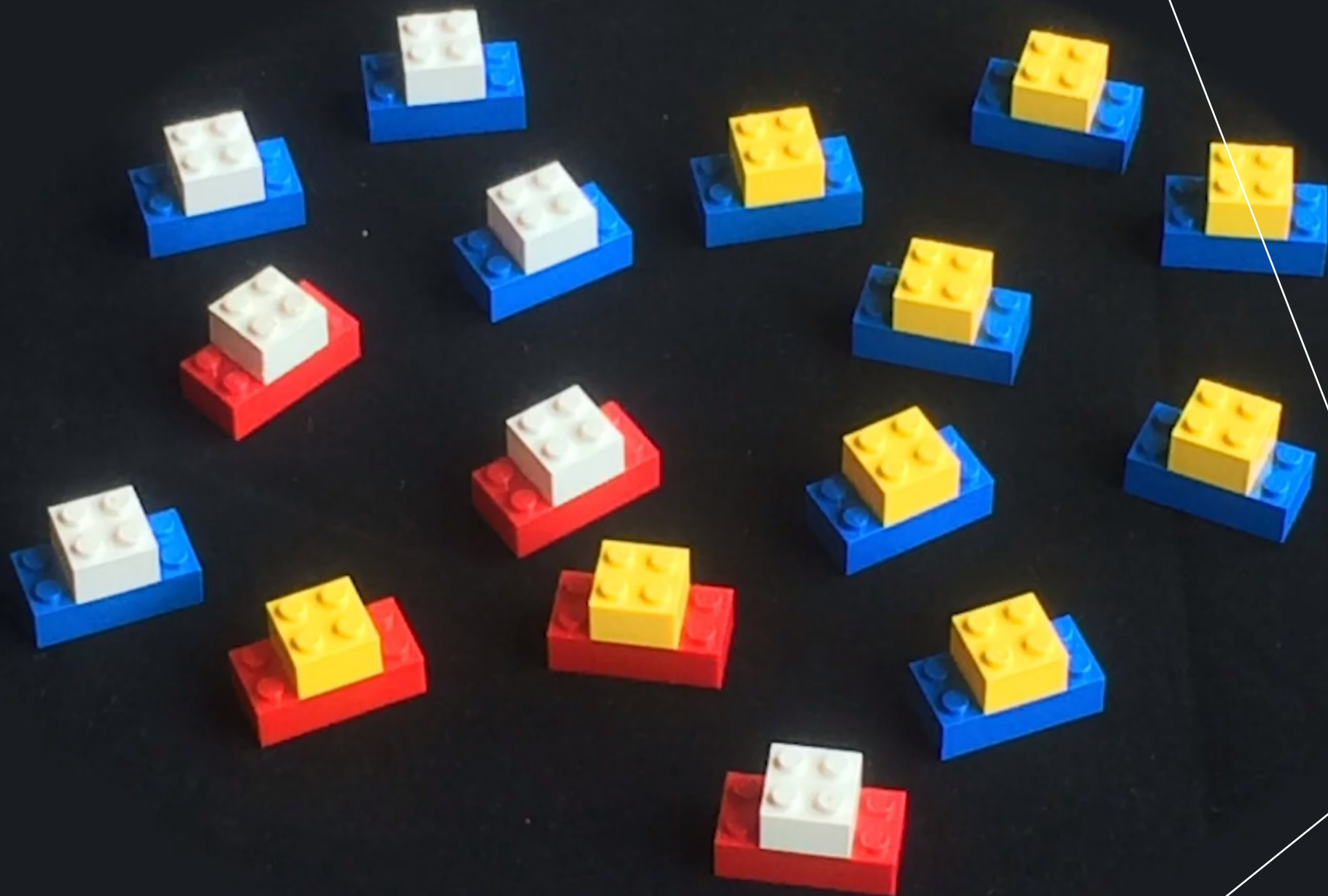
The probability of
B given A

The probability
of A

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

The probability of
A given B

The probability
of B



Bayes Theorem

- Given a hypothesis H and evidence E , Bayes theorem states that the relationship between the probability of the hypothesis $P(H)$ before getting the evidence and the probability $P(H|E)$ of the hypothesis after getting the evidence is

$$P(H | E) = \frac{P(E | H)}{P(E)} P(H).$$

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_k)P(B_k)}$$

Probability distribution and Binomial distributions

- ▶ **Probability distribution** is a total listing of the various values the random variable can take along with the corresponding probability of each value. A real life example could be the pattern of distribution of the machine breakdowns in a manufacturing unit.
- ▶ The random variable in this example would be the various values the machine breakdowns could assume.
- ▶ The probability corresponding to each value of the breakdown is the relative frequency of occurrence of the breakdown.
- ▶ **Binomial distribution** to compute probabilities for a process where only one of two possible outcomes may occur on each trial
- ▶ The Binomial Distribution is a widely used probability distribution of a discrete random variable
- ▶ It plays a major role in quality control and quality assurance function. Manufacturing units do use the binomial distribution for defective analysis.
- ▶ Reducing the number of defectives using the proportion defective control chart (p chart) is an accepted practice in manufacturing organizations.

Hypothesis Testing

Concepts of Sample Distribution

- ▶ Why do we need sampling?
- ▶ Analyse the sample and make inferences about the population
- ▶ Sample statistic vs population parameter
- ▶ Sampling distribution - distribution of a particular sample statistic of all possible samples that can be drawn from a population - sampling distribution of the mean

Sample Distribution - Central Limit Theorem

--

Assume a dice is rolled in sets of 4 trials and the faces are recorded. This is repeated for a month (30 days)

Sample	Throw 1	Throw 2	Throw 3	Throw 4	Mean
1	4	1	6	2	3.25
2	1	2	3	2	2
3	5	6	4	6	5.25
4	4	3	6	1	3.5
5	2	2	4	3	2.75
6	4	2	1	6	3.25
7	3	6	6	4	4.75
8	2	4	2	5	3.25
9	2	1	5	6	3.5
10	1	3	6	6	4
11	4	3	3	3	3.25
12	6	5	4	1	4
13	3	3	3	1	3.25
14	2	5	2	6	3.75
15	1	3	1	6	2.75

Sample	Throw 1	Throw 2	Throw 3	Throw 4	Mean
16	6	4	5	5	5
17	3	2	3	6	3.5
18	1	3	2	1	1.75
19	6	1	3	3	3.25
20	5	2	5	6	4.5
21	1	2	1	6	2.5
22	3	2	6	2	3.25
23	3	1	3	4	2.75
24	3	2	6	4	3.75
25	6	1	1	5	3.25
26	1	5	2	2	2.5
27	4	2	2	3	2.75
28	4	6	2	5	4.25
29	4	2	3	5	3.5
30	3	1	4	1	2.25

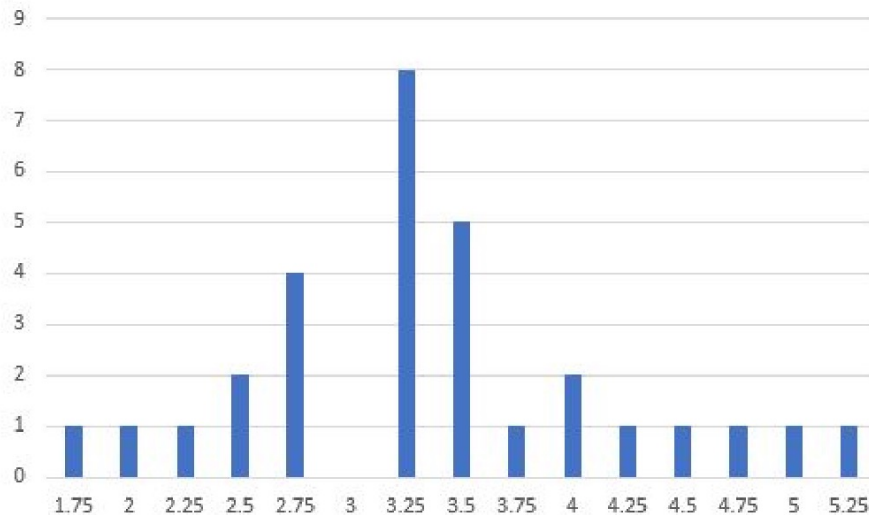
Sample Distribution - Central Limit Theorem

↵

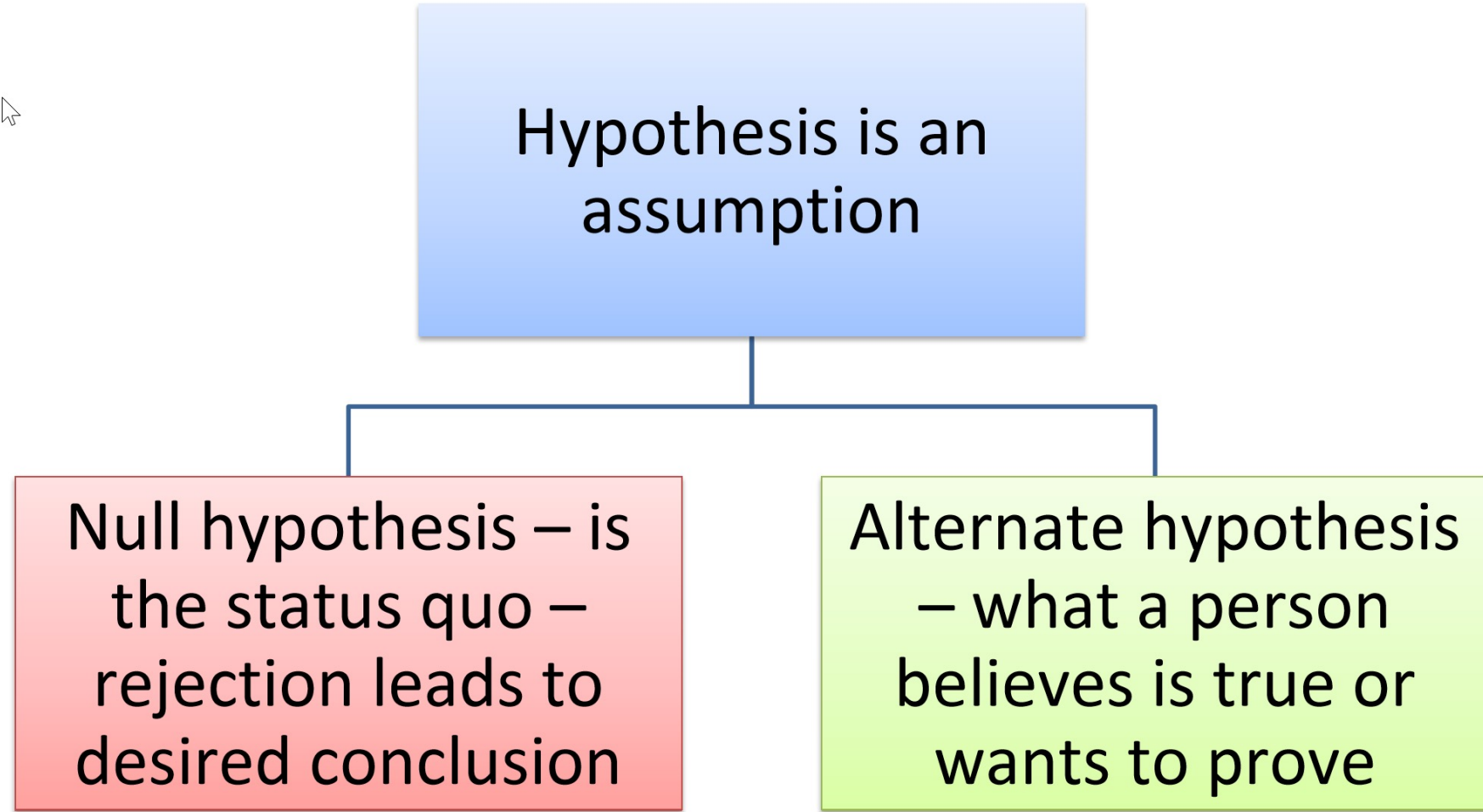
The means of the 30 samples are obtained are recorded in a frequency distribution table:

Mean	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4	4.25	4.5	4.75	5	5.25
Frequency	1	1	1	2	4	0	8	5	1	2	1	1	1	1	1

Plotting the sample distribution of the sample mean, the following curve is obtained:



Hypothesis



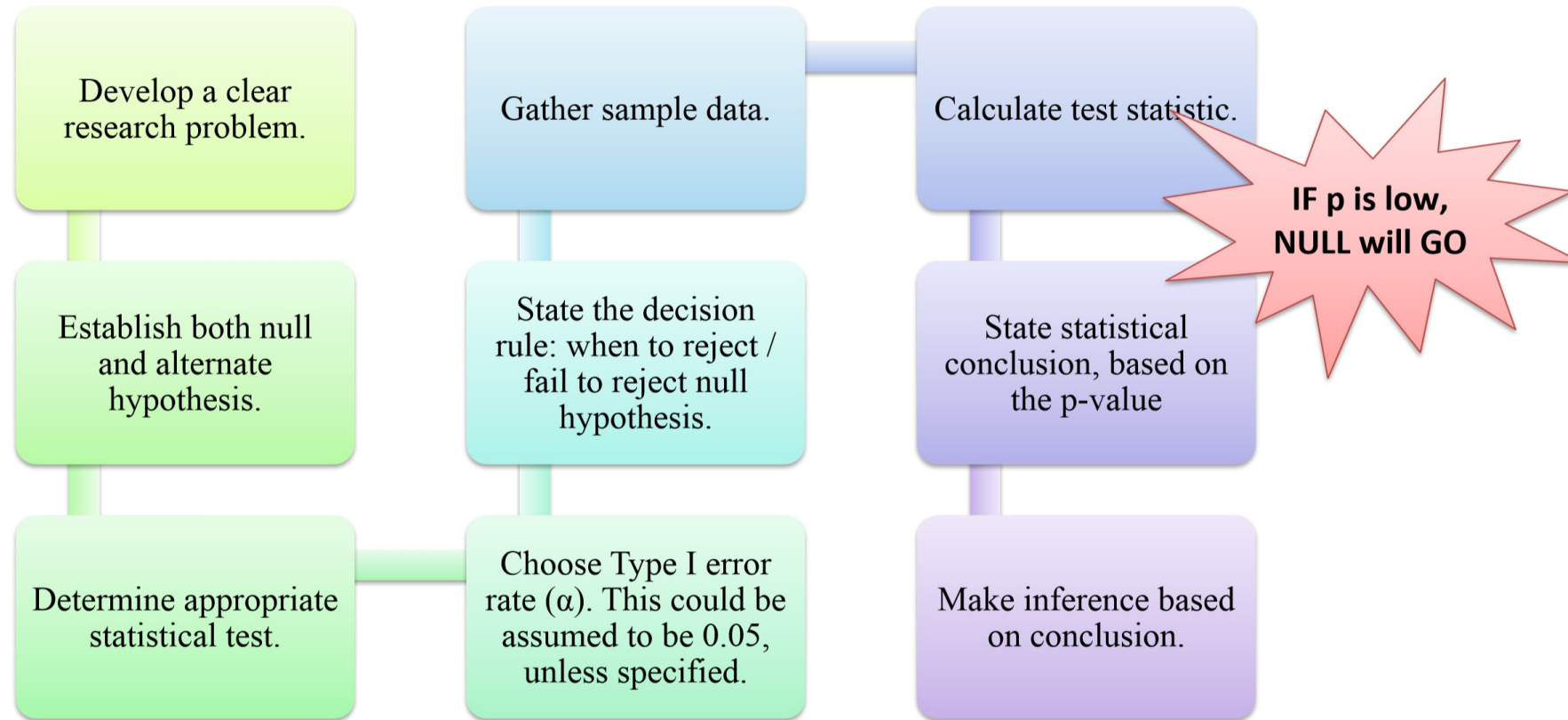
Hypothesis

- ▶ Null and Alternative Hypothesis
- ▶ All statistical conclusions are made in reference to the null hypothesis.
- ▶ We either reject the null hypothesis or fail to reject the null hypothesis; we do not accept the null hypothesis. From the start, we assume the null hypothesis to be true, later the assumption is rejected or we fail to reject it.
 - ▶ When we reject the null hypothesis, we can conclude that the alternative hypothesis is supported.
 - ▶ If we fail to reject the null hypothesis, it does not mean that we have proven the null hypothesis is true.
 - ▶ Failure to reject the null hypothesis does not equate to proving that it is true.
 - ▶ It just holds up our assumption or the status quo.

Level of significant

Alpha (α) => Level of significance = 5% default

Steps in Hypothesis testing



Types of Hypothesis Testing

- ▶ Single sample or two or more samples
- ▶ One tailed or two tailed
- ▶ Tests of mean, proportion or variance
 - ▶ (Handling categoriacal data , and continuous numeric data)

One tailed vs two tailed test

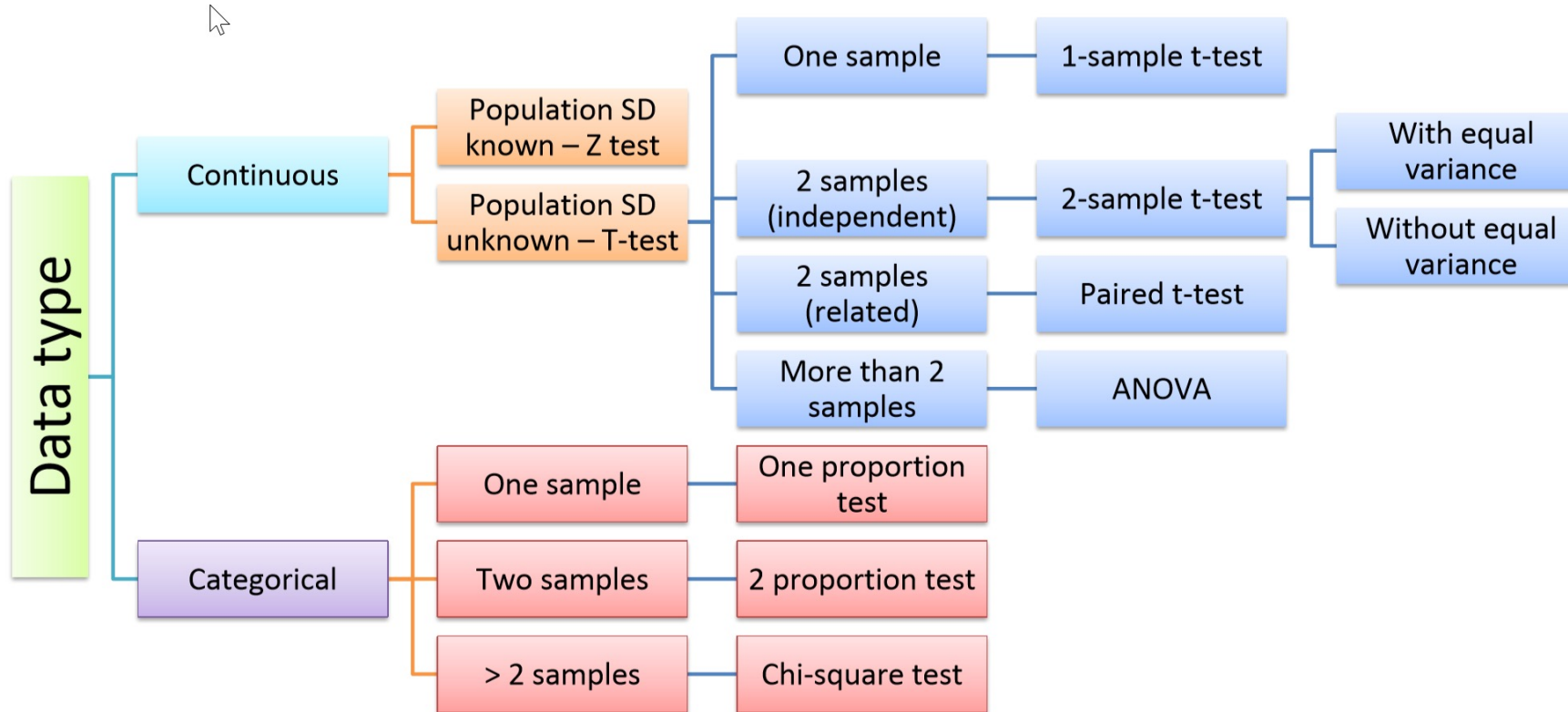
Case 1

- $H_0: \mu = 600\text{ml}$
- $H_a: \mu \neq 600\text{ml}$
- Two-tailed test

Case 2

- $H_0: \mu \leq 600\text{ml}$
- $H_a: \mu > 600\text{ml}$
- One-tailed test

Hypothesis testing roadmap



HYPOTHESIS TESTING

Approaches

		O/P	
		Num	Cat
I/P	Num	PEARSON	T-Test ANOVA
	Cat	T-Test ANOVA	CHI SQUARE