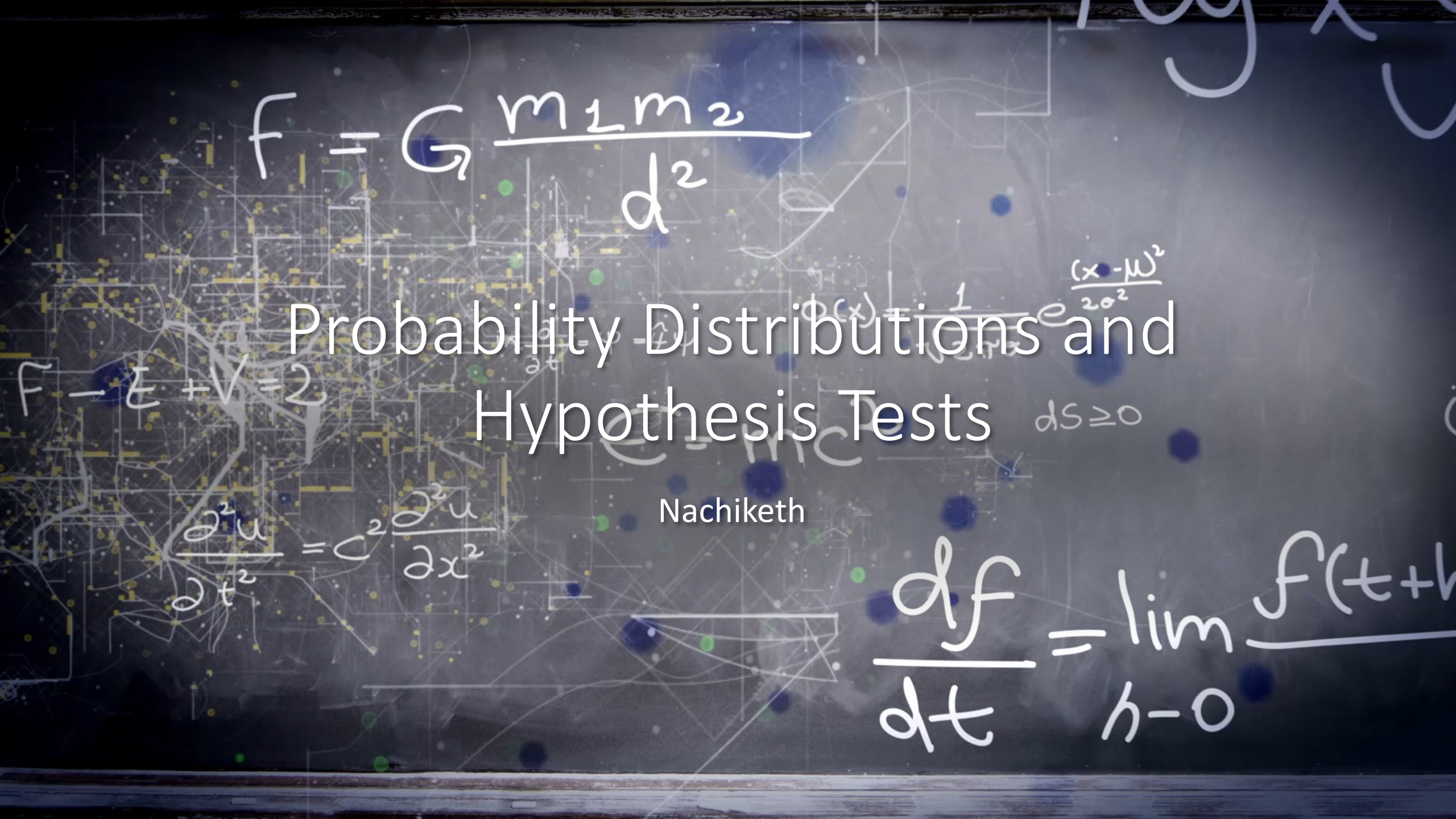


Probability Distributions and Hypothesis Tests

Nachiketh

Probability Theory : Terminology

Random Experiment:

- In ML we mostly deal with uncertain events. Random Experiment is an experiment in which outcome is not known with certainty

Sample Space

- It's a Universal set that consists of all possible outcomes of an experiment
- Example: Outcome of College Application: S = {admitted, not admitted}

Event:

- Subset of Sample space and probability is usually calculated with respect to an event.
- Example: Chances of getting head on a fair coin

Random Variables

Random Variables

A Function that maps every outcome in the sample space to a real number

It can be classified as discrete or continuous depending on the values it can take.

If a random variable X can assume only finite/countable infinite set of values, we call its as *discrete random variable*

If a random variable X can take a value from an infinite set of values is called *continuous random variable*

Examples of Discrete Random Variable

- Credit Rating (finite categories)
- Number of orders received in an e-commerce website
- Customer churn (Yes or No)
- Fraud detection (Yes/No)

Examples of Continuous Random Variable

- Market share of a company (any value from an infinite set b/n 0 to 100)
- Percentage of Attrition of employees
- Time to failure of an Engineering system
- Time taken for an e-commerce company to complete the order

Discrete Random Variables

- Described *using probability mass function* (PMF) and *cumulative distribution function* (CDF)
- PMF is the probability that a random variable X takes a specific value *k*
  - Number of fraudulent txns at an e-commerce platform is 10
- CDF is the probability that a random variable X takes a value less than or equal to k
  - Number of fraudulent txns at an e-commerce platform is less than or equal to 10

Continuous Random Variables

- Described *using probability density function* (PDF) and *cumulative distribution function* (CDF)
- PDF is the probability that a continuous random variable X takes a value in a small neighbourhood
  - Price of house between 100k to 125k
- CDF is the probability that a continuous random variable X takes a value less than or equal to k
  - Price of house less than or equal to 100k
- PPF is the probability that a continuous random variable X takes a specific value equal to k
  - Price of house is equal to 100k

Distribution of Data

Binomial Distribution

- Discrete probability distribution
- Random variable has only two outcomes
- Objective to find probability of getting x successes out of n-trials
- Prob of success is p, hence prob of failure is (1-p)
- Example: Loan Repayment Default
- PMF/CDF

Example Problem

Example Problem

We have a company that's called as fashion trends online . Its an ecommerce company and it sells women apparel. It is observed that 10% of their customers is going to return the items for many reasons (it could be due to size colour and material mismatch). Now on a specific day 20 customers has purchased from this website . Now as a data scientist we want to solve this problem :

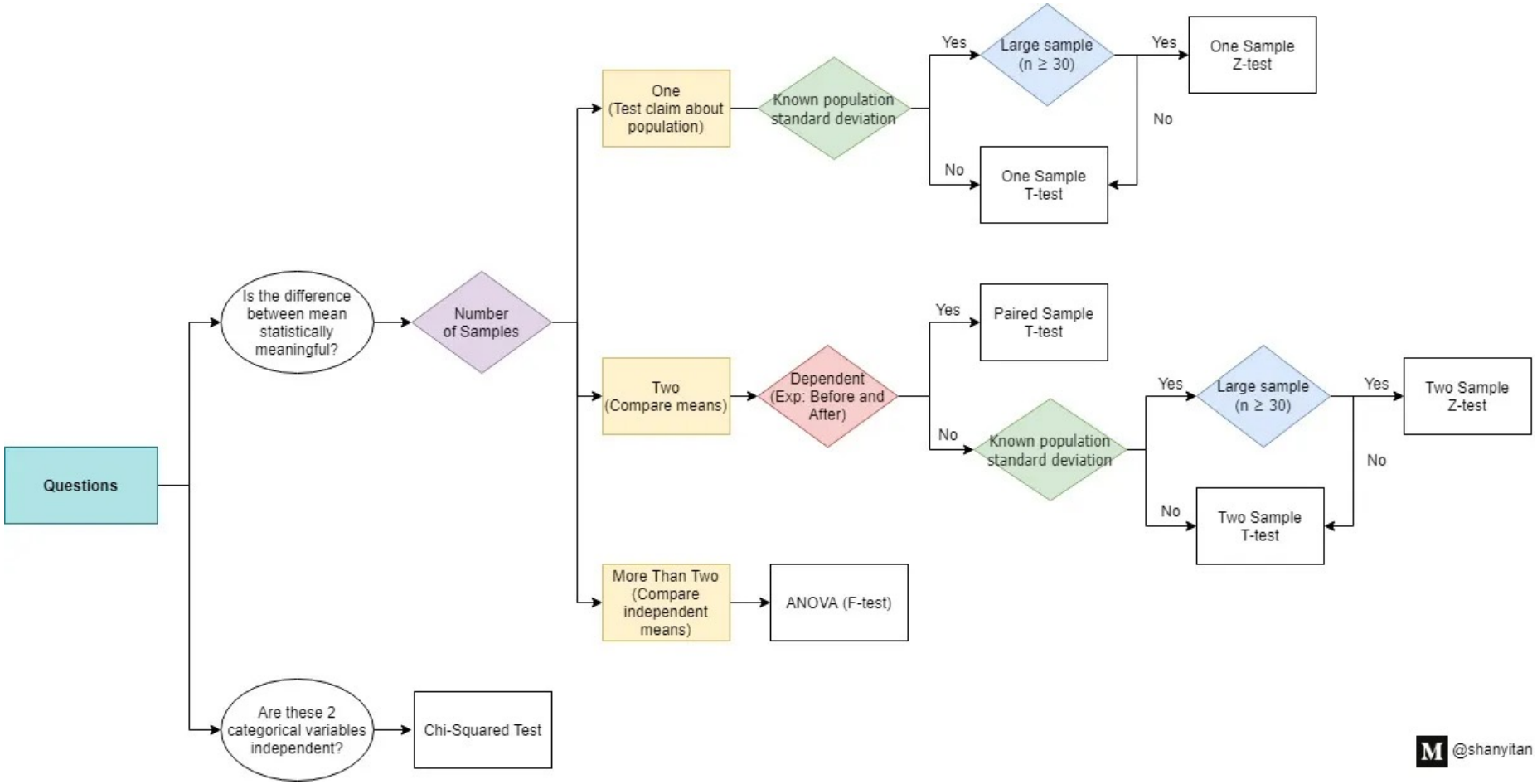
1. what is the probability that exactly 5 customers will return the items
2. probability that the maximum of five customers will return the items
3. more than five customers will return the items purchased by them
4. average number of customers who are likely to return the items and the variance and the standard deviation of the number of returns

Central Limit Theorem

- Let’s say if I have samples S1, S2, S3,…. SK with each sample is of size ‘n’. If the population standard deviation is **Sigma** and the mean is **mu.** let X1, X2, X3, up to Xk be the means of my individual samples , and S1,S2 ,S3 upto Sk are the standard deviation of individual samples. Then according to the central limit theorem the distribution of this X1 X2 X3 up to Xk , follows the normal distribution with mean **mu** and standard deviation **Sigma by square root of n**

Hypothesis Test

How to Choose Hypothesis Testing



Hypothesis Test

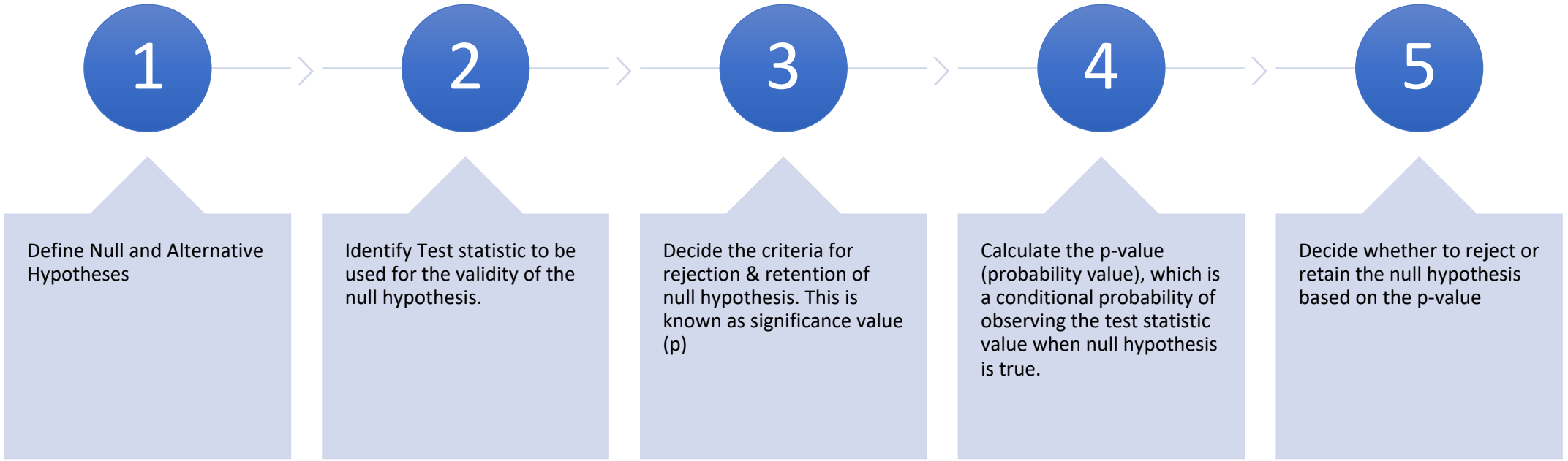
Hypothesis is a claim, and the objective of hypothesis testing is to either reject or retain a *null hypothesis (current belief)* with the help of data. Normally Hypothesis consists of Two complementary statements called :

1. Null Hypothesis
2. Alternative Hypothesis

Examples for Null Hypothesis (Current Belief)

1. Children who drinks Boost (A health drink produced by the company Nestle) are likely to grow taller and stronger.
2. Women are more addicted to Instagram than men.
3. Vegetarians miss few flights.
4. Smokers make more sales than non-smokers.

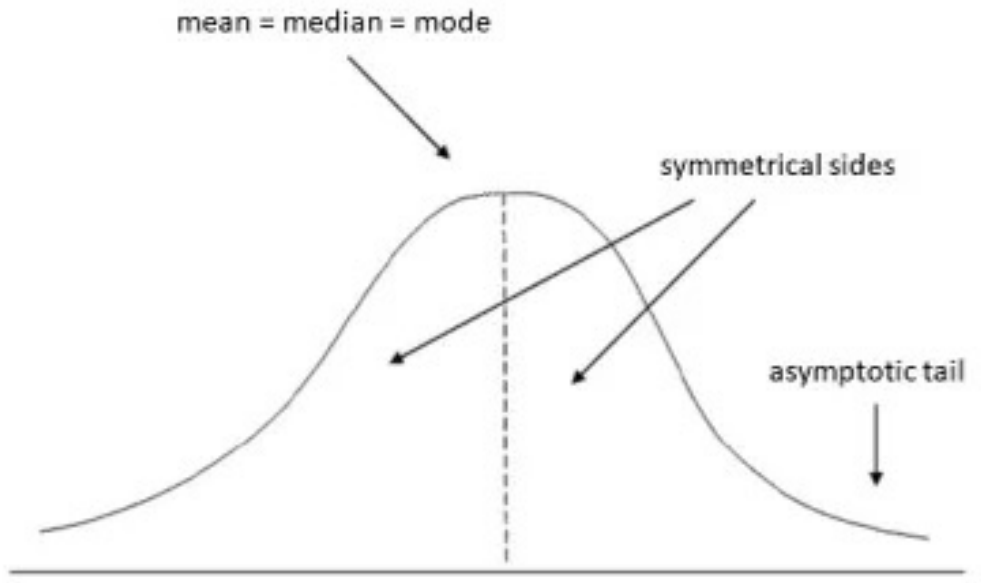
Steps for Hypotheses Test



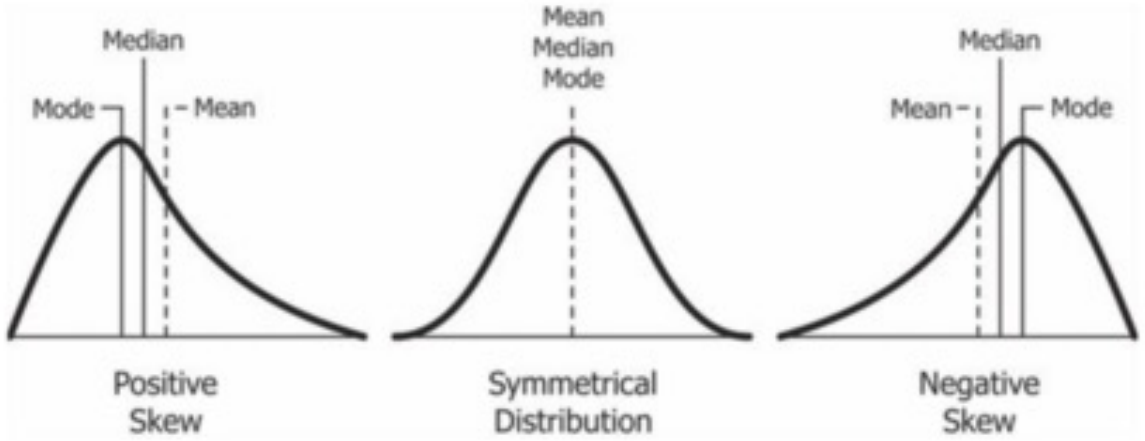
Normal Distribution

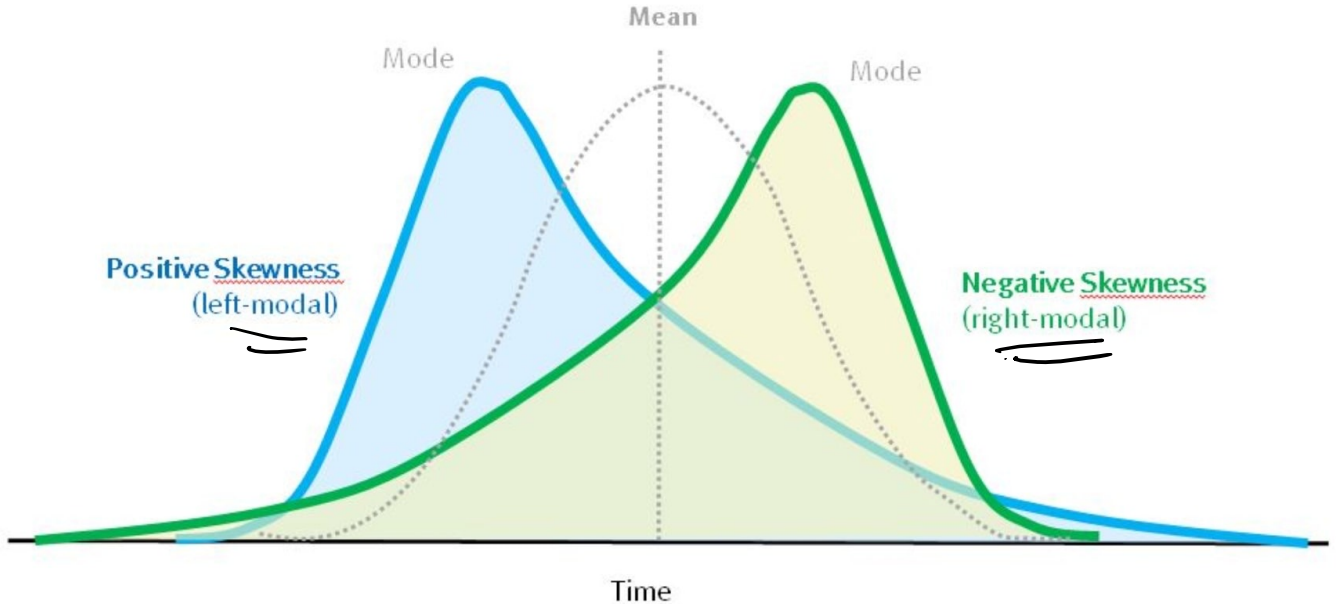
Normal distribution is actually known as “gaussian distribution”. It's one of the most popular continuous distribution in the field of analytics.

This normal distribution is observed across various natural occurring measures such as age salary, sales volume, birth weight, height etc. it is popularly known as bell curve .



Skewness


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew


Mean
Mode
Mode
Positive Skewness
(left-modal)
Negative Skewness
(right-modal)
Time

Z-Test

population mean = μ
population std = σ
Sample mean = $\bar{X}$
Sample std = s
Sample size = n

① Z-test:

(a) to test pop. mean given pop variance

(b) I have information on a Sample.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Example 1
Z Test

A passport office it claims that the passport applications **are processed within 30 days of submitting the application** form and all the necessary documents. Now we have a CSV file and it contains the processing time of **40 passport applicants**, the population **standard deviation of the processing time is given by 12.5 days.** As a part of our activity we have to conduct a hypothesis test at the significance level of **0.05 (p-value)** to verify the claim made by the passport office .

- Null Hypothesis : Passport office processes application greater than 30 days
- Alternate Hypothesis: Passport office process application within 30 days

One Sample t-Test

One Sample t-test

① When pop Std. is unknown & 'S' is known

$$t\text{- Statistic} = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$$

One Sample t-Test

Arvind productions is a newly found movie production company based out of India, and this production company is interested in understanding the production cost which is required for producing a Bollywood movie. The industry believes that production house will require around 50 crore rupees on an average to produce a movie. It is assumed that Bollywood movie production cost follows the normal distribution. The production cost of 40 Bollywood movies in millions of rupees are given in our CSV file. Now conduct an appropriate hypothesis test at alpha equal to 0.05 to check whether the belief about average production cost is correct or not .

Two-Sample t-Test

We uses two sample t test when we want to find the difference between two population means where standard deviations are unknown. In this scenario the parameters are estimated from the samples .

Example 3 – Two Sample t-Test

A company claims that children who drink their health drink will grow taller than the children who do not drink that health drink . We have a CSV file where it shows about the change in height over one year period from two groups : one group drinking the health drink, and the another drink who is not drinking health drink. Now at alpha equal to 0.05. We want to test whether the increase in height for children who drank the health drink is it different than those who didn’t drink health drink

Null Hyp – H0 – There is no difference in height between children drinking Horlicks with children not drinking Horlicks

Alt. Hyp - The increase in height for children who drank the health drink is it different than those who drink health drink

When we want to analyse weather in intervention or a treatment such as an event, training program, marketing promotion, treatment of specific illness and lifestyle changes, have had any significant changes in the population parameters → mean

Before and after the intervention -> in such scenarios we make use of this paired sample t test.

Paired Sample t-test

Example 4

I have a data set which is called as breakups CSV file and it contains the alcohol consumption before and after breakup. Now conduct a paired sample t-test to check whether the alcohol consumption is more after the breakup at alpha is equal to 0.05

Chi-Square Goodness of Fit Test

This chi square goodness of fit test is actually a nonparametric test and we use it for comparing the observed distribution of data with the expected distribution of data, to decide whether there is any statistically significant difference between the observed distribution and a theoretical distribution .

Example 5:

Indigo airlines which operates daily flights to several Indian cities . One of the problem this airline faces is the food preference by the passengers. The Captain Cook the operation managers of indigo airlines, believes that 35% of their passengers prefer vegetarian food, 40% of the passengers prefer non-vegetarian food, 20% of people low calorie food and 5% request for diabetic food.

We have a sample of 500 passengers which was chosen to analyse the food preference and the observed frequencies are as follows :

- Vegetarians 190
- Non vegetarian 185
- Low calorie 90
- diabetic 35

Perform chi square test to check whether captain cooks belief is true or not .

ANOVA

So we use this ANOVA testing when we want to conduct an hypothesis test to compare the mean values simultaneously for more than 2 groups and these 2 groups are created from the same samples using a factor .

Problem Statement

So we've got a brand manager Surf Excel which is a detergent company and they are going to offer the discounts.

The discounts are 0%, 10% and 20% on randomly selected days .

Now we have the data that contains the quantity sold in kilograms of the detergent powder on each day. As part of the hypothesis testing we want to perform one way ANOVA to check whether the discount has any impact on the quantity of detergent sold .

Comparing Population and a Sample

- Pop_mean, Pop_std is known → Z-test
- Pop_mean is known, pop_std is unknown → t-test

Comparing Samples

- When comparing two independent samples → 2 sample t-test
- When comparing 2 related samples (breakups) → pair sampled t-test

Non-parametric test (frequency of categorical data)

- Chi Square Goodness of Fit test

Compare Multiple Samples together

- Anova test