

STATISTICS



GROUND RULES

- Come prepared for these sessions by watching the video lectures.
 - Concepts will be covered in the videos.
 - Hands-On Application will be covered in Mentor Sessions.
- Submit all assignments on time.
- Let's be punctual & respect each others time.



LEARNING OBJECTIVE OF THIS MODULE

- Descriptive Statistics
- Inferential Statistics
- Hypothesis Testing

LEARNING OBJECTIVES OF THIS SESSION -

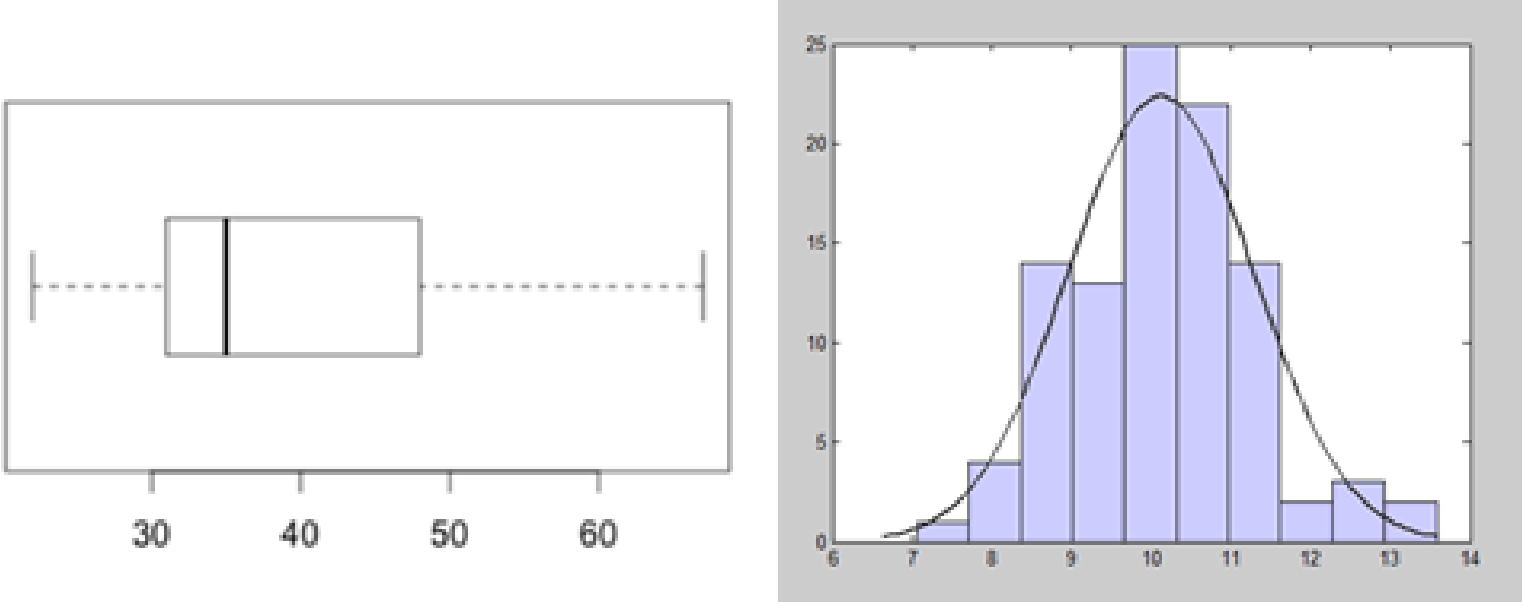
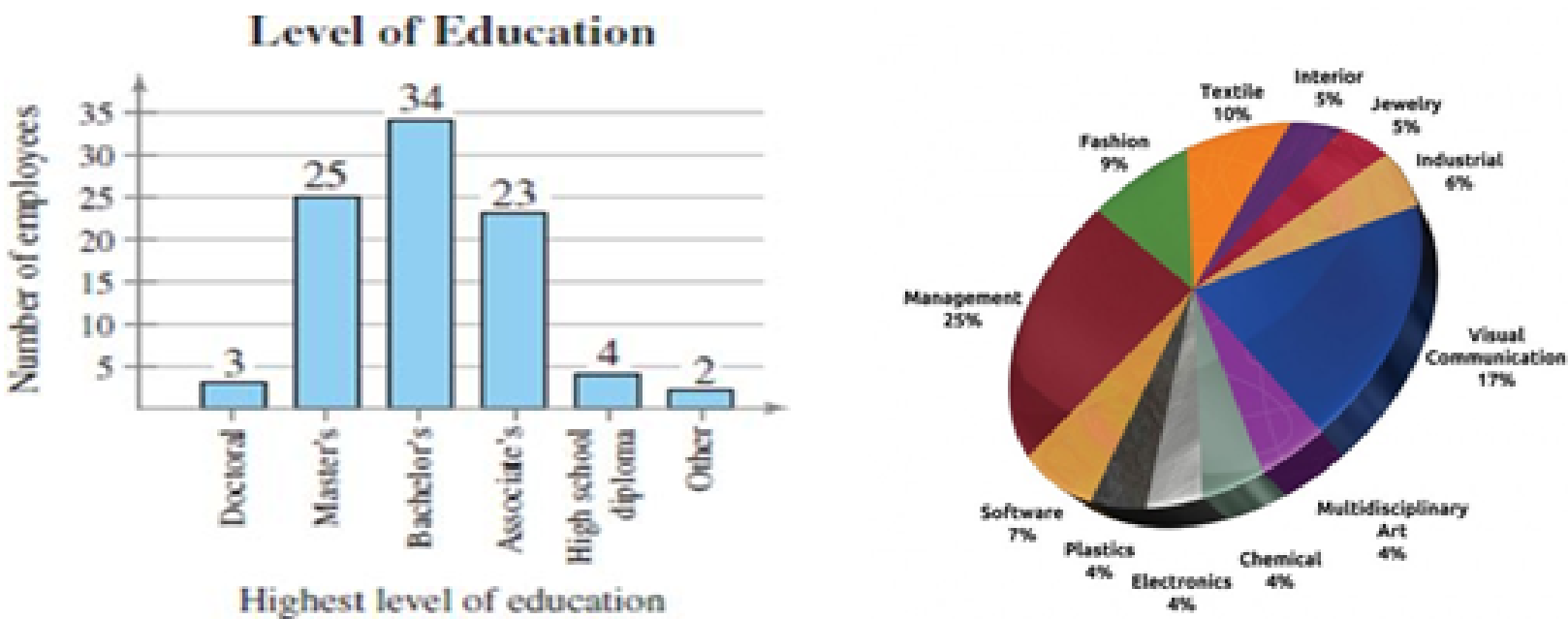
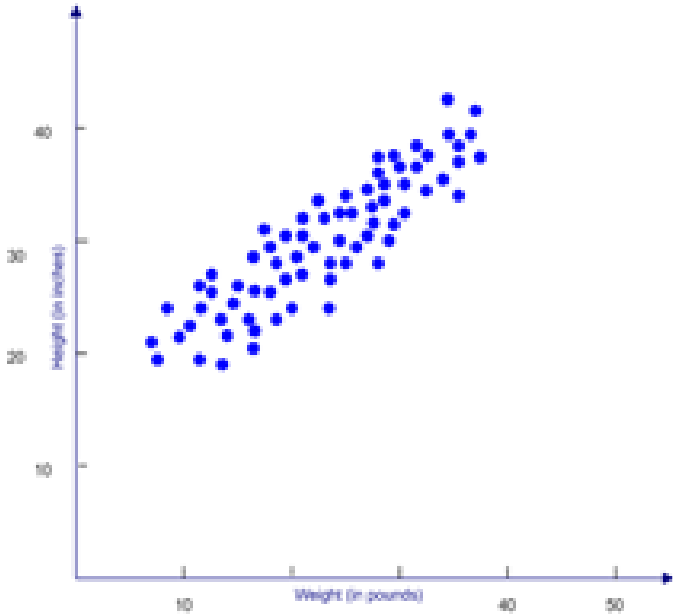
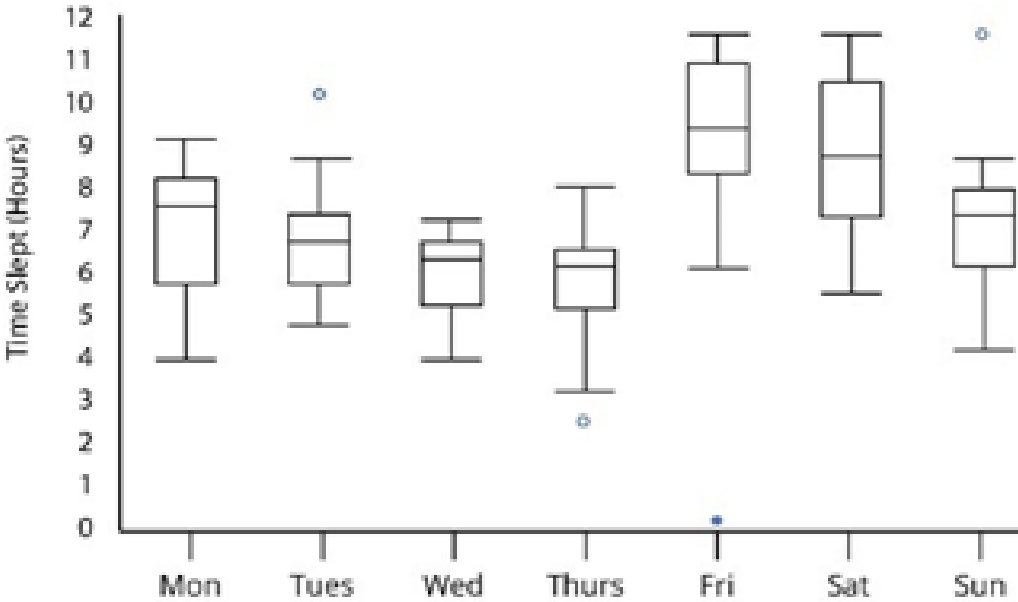
- Measures of Central Tendency
- Measures of Dispersion
- Graphical Display of Data
- Distribution and idea of Skewness
- Correlation

TRY ANSWERING THE FOLLOWING

- Find Mean, Median, Mode and Range of the below data :-
40,20,20,60,80,60,20,60
- What are the 5 points in a box-plot?
- What is the formula of Coefficient of Variation(CV)?



BROAD OVERVIEW

	Numerical	Categorical
Univariate	<div><p>The univariate numerical section displays two charts. On the left is a box plot with a median around 35, an IQR from approximately 32 to 48, and whiskers extending from 25 to 65. On the right is a histogram with a normal distribution curve overlaid, centered at 10, with the x-axis ranging from 6 to 14 and the y-axis from 0 to 25.</p></div>	<div><p>The univariate categorical section contains two charts. The 'Level of Education' bar chart shows the number of employees for each education level: Doctoral (3), Master's (25), Bachelor's (34), Associate's (23), High school diploma (4), and Other (2). The pie chart shows the distribution of various fields: Management (25%), Visual Communication (17%), Textile (10%), Fashion (9%), Software (7%), Plastics (4%), Electronics (4%), Chemical (4%), Multidisciplinary Art (4%), Industrial (6%), Jewelry (5%), and Interior (5%).</p></div>
Multivariate	<div><p>The multivariate numerical section features a scatter plot showing a positive correlation between 'Weight (in pounds)' on the x-axis (ranging from 10 to 50) and 'Height (in inches)' on the y-axis (ranging from 20 to 40). The data points are blue dots.</p></div>	<div><p>The multivariate categorical section displays a box plot of 'Time Slept (Hours)' across the days of the week. The y-axis ranges from 0 to 12. The boxes represent the median, quartiles, and range of sleep hours for each day, with outliers shown as small circles.</p></div>

Application of Measures of Central Tendency

Which measure will be used in case of following scenarios ?

- When you try to search cool game app in play store, what measure will you look for? Is it the average rating?
- Suppose in your class there are 11 students and one of them is son of Bill Gate. How will you calculate average pocket money?
- How will you decide which color is most popular among citizens ?

Application of Measures of Dispersion

For example, Rohit Sharma has a batting average of 47 runs per innings and the standard deviation is 30 runs. So more or less he scores between 17–77 runs on an average. Now if you look at scores of Shikhar Dhawan. He has an average of 45 runs per innings and his standard deviation is also 15 runs. So he scores somewhere between 30–60 runs on an average.

Now since both have nearly similar batting average, whom can we count on to score more consistently?

Industry Application - Personality Assessment

In order for one to make meaningful statements about psychological events, the variable or variables involved must be organized, measured, and then expressed as quantities. Such measurements are often expressed as measures of central tendency and measures of variability.

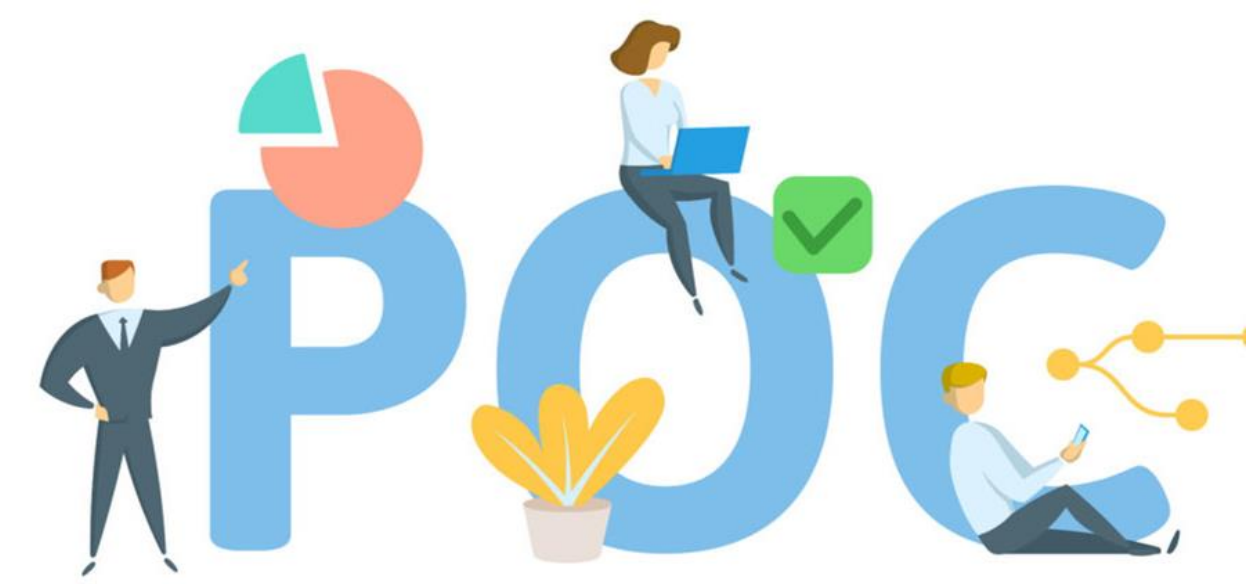
In psychological testing, Descriptive statistics give a general picture of the scores in a given group. They include the measures of central tendency and the measures of variability. Central tendency involves different kinds of averages: the mean, median, and mode. Variability involves the standard deviation, which indicates how far scores in a group are likely to be from the average.



CASE STUDY - HEALTH INSURANCE

Most companies are now recognizing the power of data in making crucial business decisions. For an Insurance company, it becomes more important to study various attributes about their customers. Leveraging this customer information to make business decisions can provide a competitive edge to the company over other players in the market

We are provided with some customer data of an Insurance company like age, gender, BMI and medical charges billed by insurance company. We need to explore this data to see if we can derive some meaningful insights from this data.



Proof of concept

ALUM TALKS - Proof of Concept

“For every Licensed user, our Company is charged for different software used by us. I manage the details of the number of licenses for different software being used in various accounts. After completing the SMDM course, I analyzed the cumulative costs, maximum and minimum costs of software licenses in different regions using various visualization techniques.

This made us realize that we were using 6 CRM software licenses where only 1 was sufficient at a particular location. I presented this to our management team and this saved us **4080 USD** annually. I am really looking forward to upcoming modules in the Program.”

Call for Action - Please go back and think how you can use the concepts learned in this SMDM module, in your present role in your organization. - **Satish Kumar**

- APPLICATION OF INFERENCE STATISTICS

- Probability
- Bayes' Theorem
- Various Probability Distributions

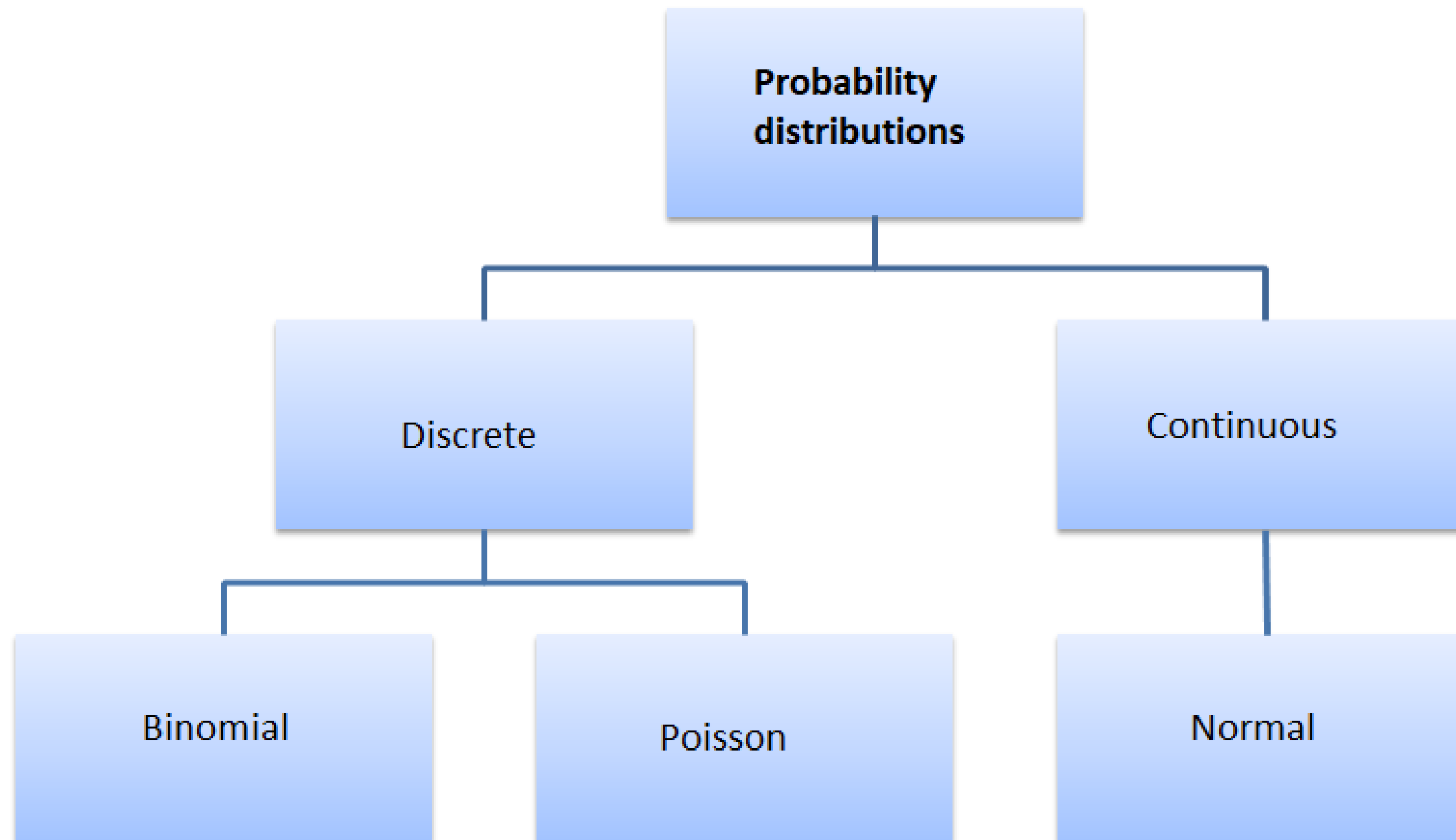
TRY ANSWERING THE FOLLOWING

- If a dice is thrown twice, what is the probability of getting 6 in both throws
- What is the formula to find mean and standard deviation for Normal Distribution
- In case of Normal Distribution what percentage of data will lie in the range of $\mu \pm 2\sigma$?



BROAD OVERVIEW

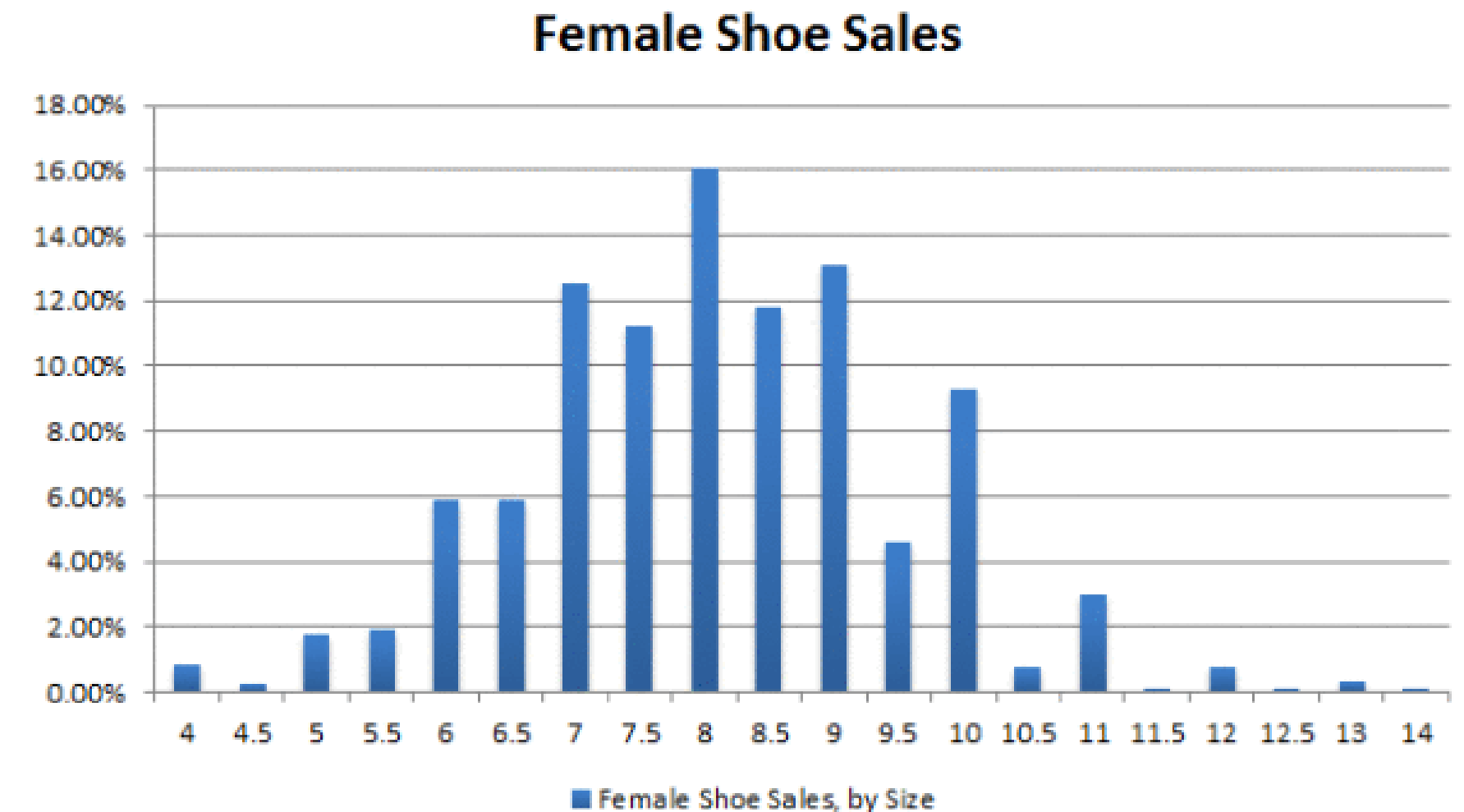
Various Probability Distributions



Real Life Example - Female Shoe Sales

This data present on the right shows the distribution of Female Shoe Sales in USA in 1998.

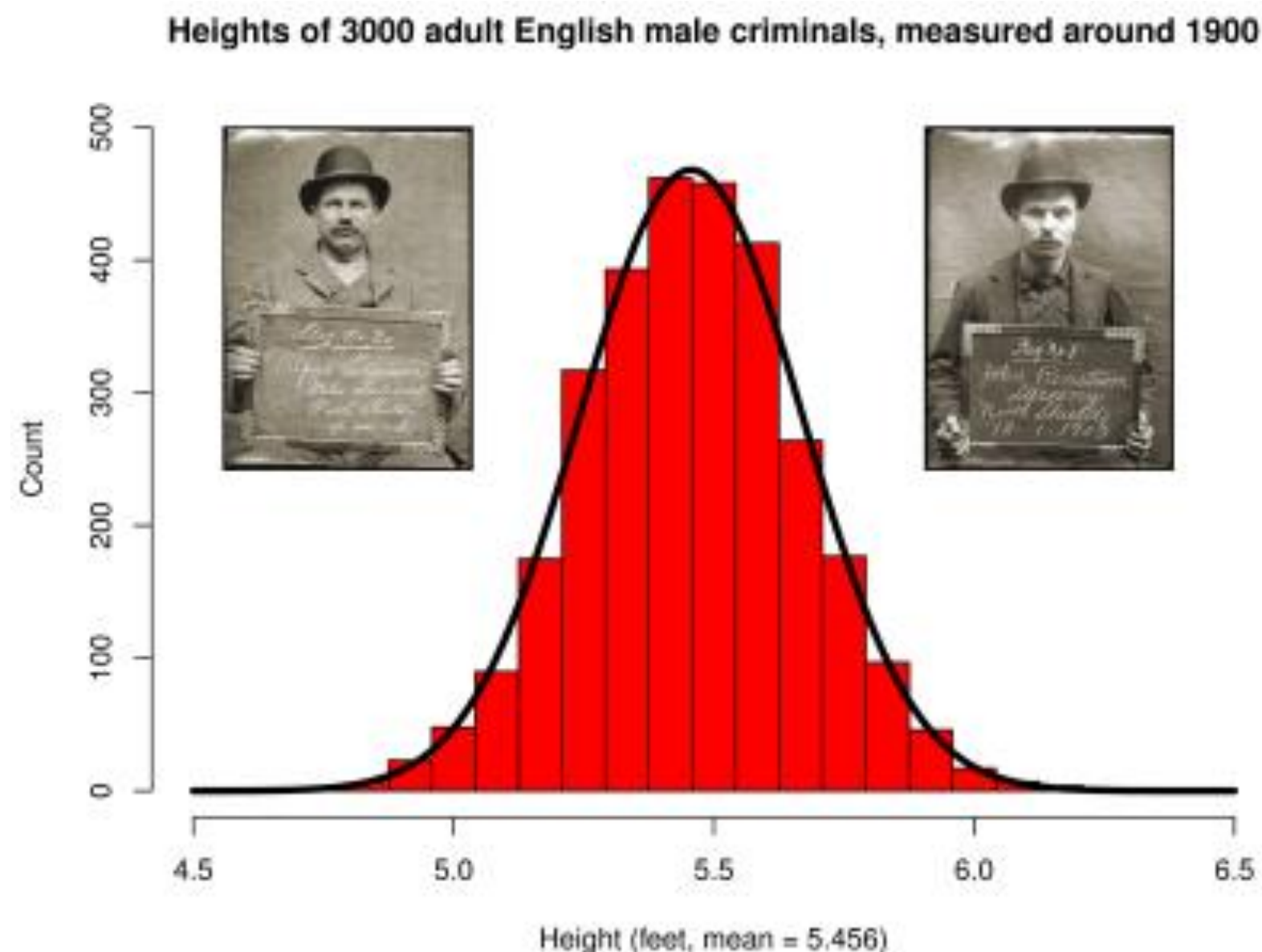
It can be used by footwear companies to produce the footwears in similar proportions and hence minimize inventory and maximize profits.



<https://thoughtburner.org/tag/normal-distribution/>

School Shootings – Can Potential Shooter Profiles be Identified?

The use of statistics has long been important in the human sciences. An early example is an analysis by William Sealy Gosset (alias “Student”) of biometric data obtained by Scotland Yard around 1900. The heights of 3,000 male criminals fit a bell curve almost perfectly.



<https://igorscience.org/category/research/>

CASE STUDIES

Probability

- 1) HR Employee Satisfaction
- 2) ATM Usage

Probability Distribution

- 1) Automobile Pollution
- 2) HR Appraisal
- 3) Labor Union Selection criteria
- 4) Average monthly cellphone bill
- 5) Campus Recruitment
- 6) ATM usage during night hours

Bayes Theorem

- 1) Computer Component scores

- APPLICATION OF HYPOTHESIS TESTING

- Central Limit Theorem
- Confidence Interval
- One-tail and Two-tail test
- Null and Alternate Hypothesis
- Hypothesis Testing

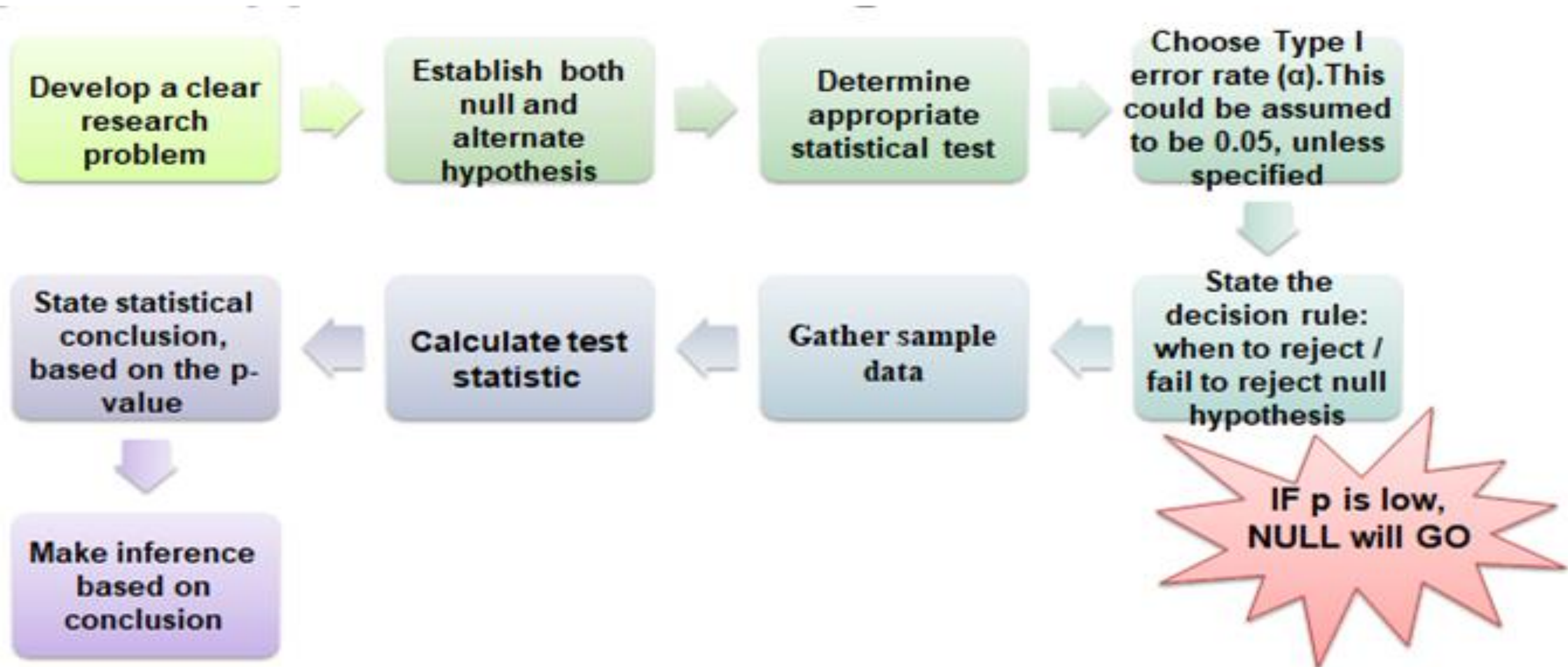
TRY ANSWERING THE FOLLOWING

- What is the formula to calculate Standard Error(SE)?
- What is the value of Level of Significance " α " for a confidence interval of 95% given a two tail test?
- Can we use Z-test for a categorical data-type?



BROAD OVERVIEW

STEPS IN HYPOTHESIS TESTING



Cosmetic Products - Market Research



1) To analyze consumer brand preference.

Hypothesis : Sales of different cosmetic brands is uniformly distributed.

Chi square test was applied. Hypothesis **rejected**. Sales of different brands is not uniformly distributed.

1) Consumer attitude towards usage of consumer products

Hypothesis : Consumers of different age groups do not differ significantly in their attitude towards usage of cosmetic products.

Z test was applied & hypothesis was **accepted**.

Can Uber double its revenue by changing a single word ?

Changing the word from “buy” to “try” more than doubled the conversions for Uber.

Uber wanted to discover how much extra its core users are willing to pay and hence increase its revenue.

<https://factordaily.com/opinions/uber-ab-test-boost-revenue-india/>



A/B Test example. Changing a single word more than doubles the conversion rate

Case Study-1 (Processing Time)

Tom is working in a credit-card processing company as a team-leader. His team is responsible to validate certain data for new credit-card applications. The time spent by his team on an application is normally distributed with average 300 minutes and standard deviation 40 minutes. Tom and his team worked on process improvement to reduce the time spent in processing new applications. After implementing the improvements, Tom checked the time spent by his team on randomly selected 25 new card applications. The average time spent is 290min. Tom is happy that, though it is a small improvement, it is a step in right direction. He shares the good news with his manager Lisa. But Lisa is not convinced about the improvement. At 95% confidence, is the process really improved?

Case Study-2 (Titan Insurance)

The Titan Insurance Company has just installed a new incentive payment scheme for its life policy sales-force. It wants to have an early view of the success or failure of the new scheme. Indications are that the sales force is selling more policies but sales always vary in an unpredictable pattern from month to month and it is not clear that the scheme has made a significant difference. Life Insurance companies typically measure the monthly output of a salesperson as the total sum assured for the policies sold by that person during the month.

Titan's new scheme is that the sales force receive low regular salaries but are paid large bonuses related to their output (i.e. to the total sum assured of policies sold by them). The scheme is expensive for the company but they are looking for sales increases to compensate for it. The scheme has now been in operation for four months. It has settled down after fluctuations in the first two months due to the changeover.

To test the effectiveness of the scheme, Titan has taken a random sample of 30 salespeople measured their output in the penultimate month prior to changeover and then measured it in the fourth month after the changeover (they have deliberately chosen months not too close to the changeover).



HAPPY LEARNING