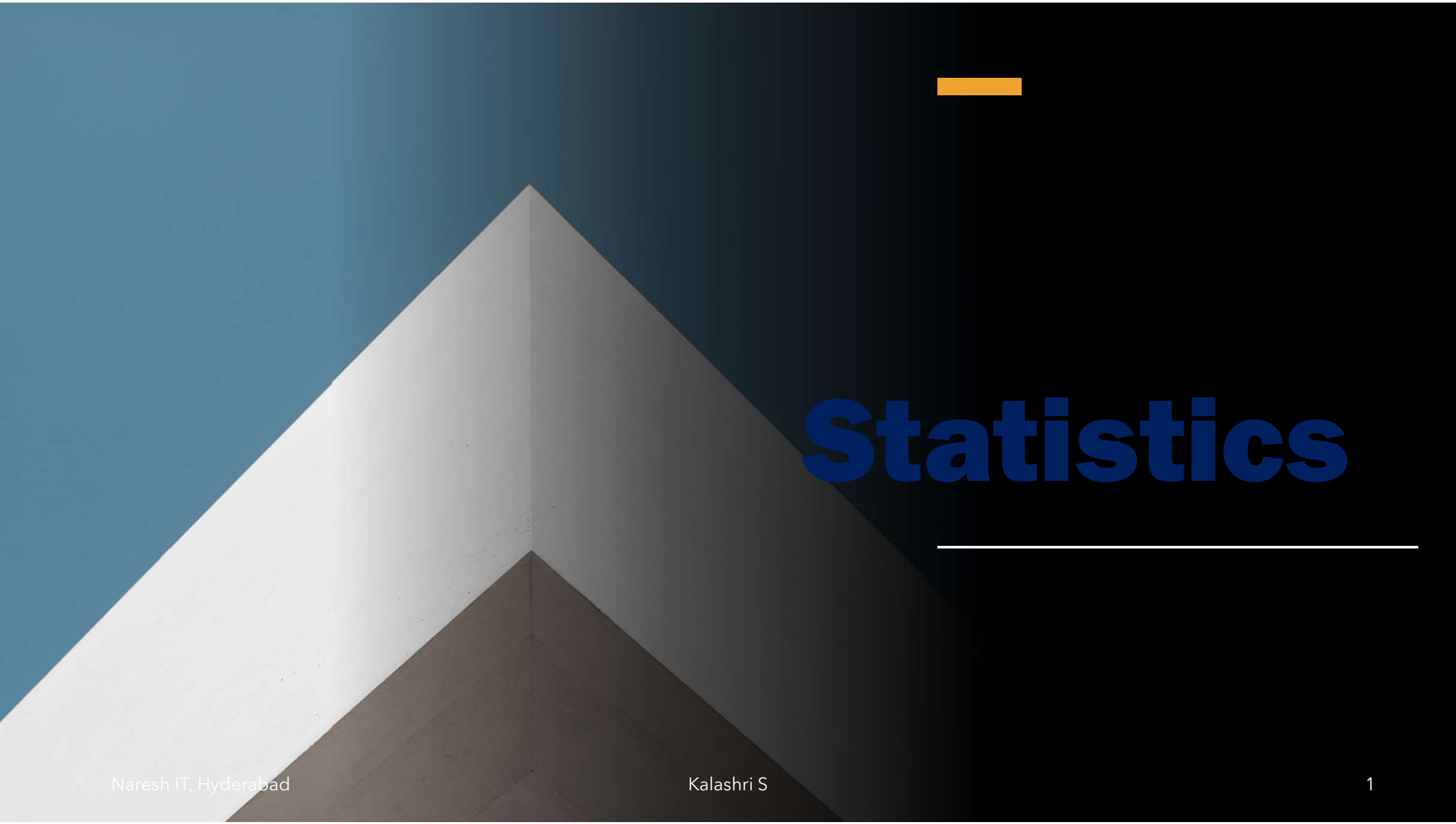


Statistics

Agenda

- Why data scientist studies statistics?
- What is statistics?

Why do we study statistics in DS?

- Numbers play an essential role in statistics. They provide the raw material of statistics. These materials must be processed to be useful, just as crude oil must be refined into petrol before it can be used an automobile engine.

- Study of statistics involves methods of refining numerical (& non-numerical) information into useful forms

Introduction to Statistics

- When numbers are collected and compiled, regardless of what they represent, they becomes statistics.

- We study statistics as a tool – and a highly valuable one – in the analysis of problems

What is statistics?

Statistics is the discipline that concerns

- **the collection,**
- **organization,**
- **analysis,**
- **interpretation,**
- **and presentation of data.**

Difference between Mathematics & Statistics

- Mathematics refers to a subject as well as to symbols, formulae, theorems & accounting.
- Statistics refers to quantitative information or to a method of dealing with quantitative information.

Numerical Measures

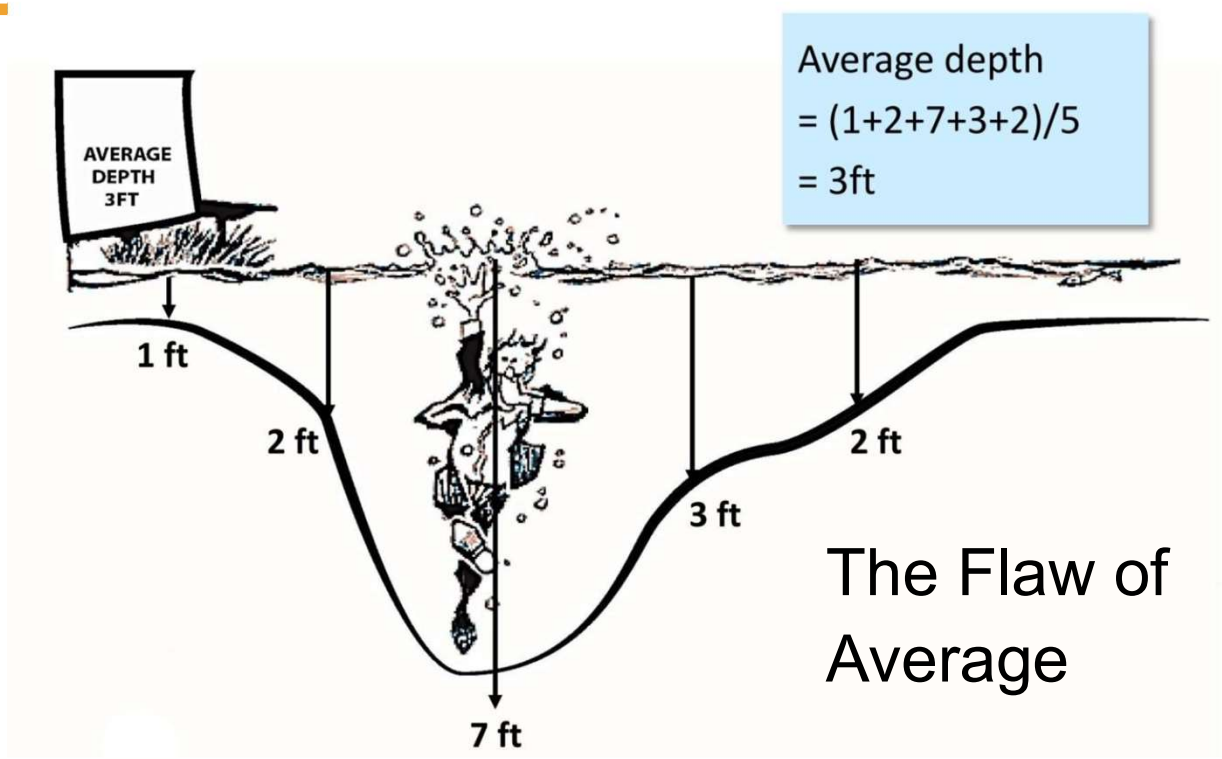
The “Average” Story

Alan went for a trek. On the way, he had to cross a stream. As Alan did not know swimming, he started exploring alternate routes to cross over.

Suddenly he saw a sign-post, which said “Average depth 3 ft”. Alan was 5’7” tall and thought he could safely cross the stream.

Alan never reached the other end and drowned in the stream.

Did Alan Drown?

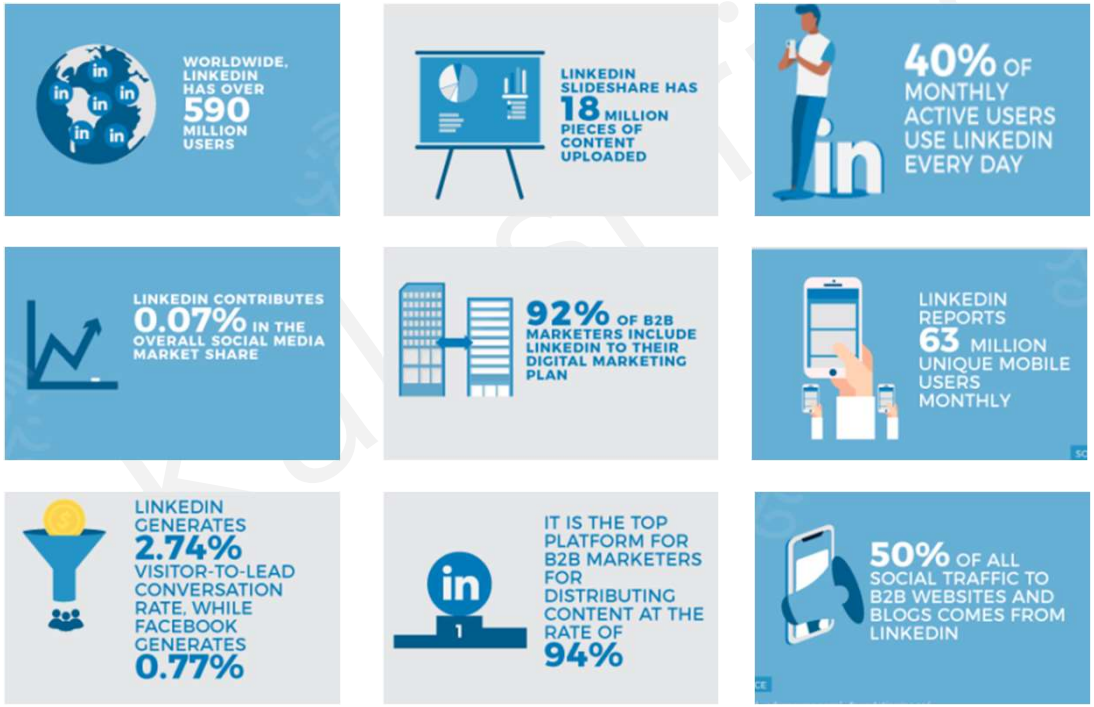


Beware of Averages!!!

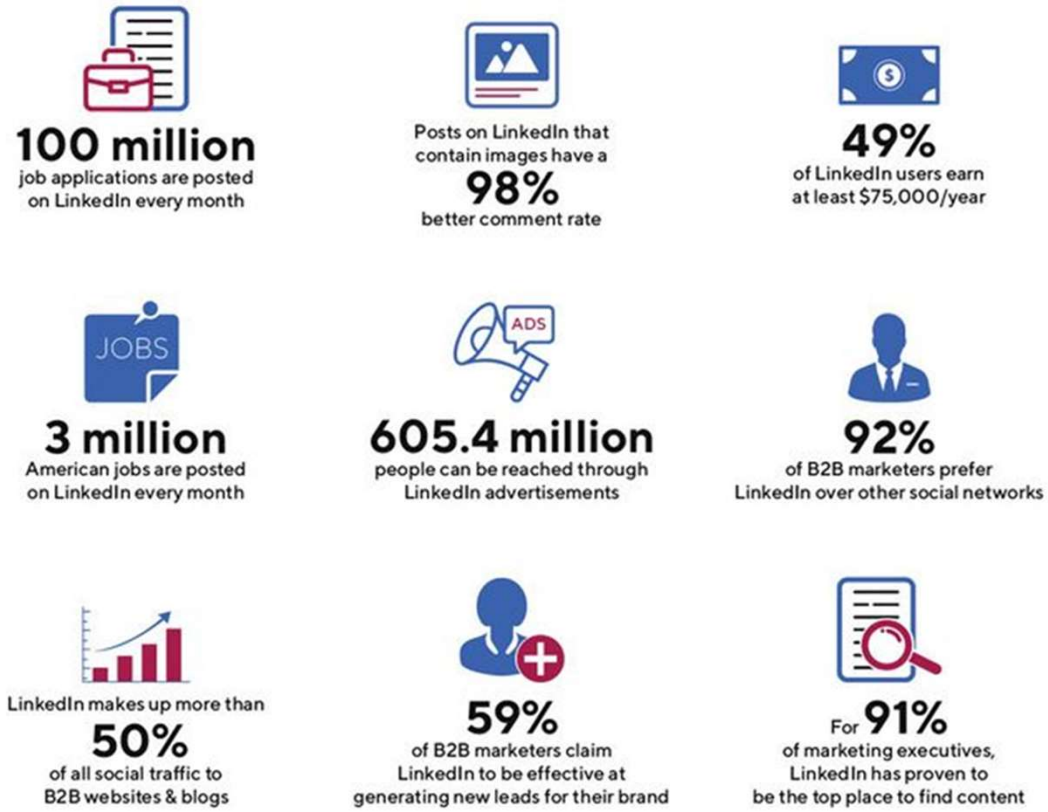
Discussion

- Can you tell a statistical story about LinkedIn?
- If you are a recent graduate, what would statistics mean to you?
- If you work in the marketing department, what decisions can you make about digital marketing strategy?
- If you want your expertise to be known to a global audience, is LinkedIn the right social media channel?
- If you are recruiter, will you post your job requirements on LinkedIn?
- If you want to invest, will you buy LinkedIn stock?

Total LinkedIn Users Statistics 2019



LinkedIn Marketing Statistics



Discussion

- Do these statistics tell you a story about LinkedIn?
- If you are a recent graduate, what would these statistics mean to you?
- If you work in the marketing department, what decisions can you make about digital marketing strategy?
- If you want your expertise to be known to a global audience, is LinkedIn the right social media channel?
- If you are recruiter, will you post your job requirements on LinkedIn?
- If you want to invest, will you buy LinkedIn stock?

Interesting Insights

- If LinkedIn users were to form a country, it would be the size of United States of America!!! Social Media's exponential growth in the past decade
- Access to internet applications is increasing by use of mobile devices
- Google reports LinkedIn company pages within the top 1-2 pages of search results; Having LinkedIn company page a no-brainer
- Job seekers and recruiters increasingly use LinkedIn with great results.
  - For example, your resume is your LinkedIn profile page.

Statistics

Statistics is **collection**, **organization**, **analysis**, and **interpretation** of data.

Statistics includes:

- Design of experiments
- Sampling
- Descriptive Statistics
- Inferential Statistics
- Probability theory

Descriptive Statistics: Key concepts

In statistical analysis, the three fundamental concepts associated with describing data are:

- Location or Central Tendency
- Dispersion or spread
- Shape or Distribution

Agenda

What is statistics?

- ✓ **Central Tendency Measures**
- ✓ Dispersion Measures
- ✓ Data Distributions

Central Tendency

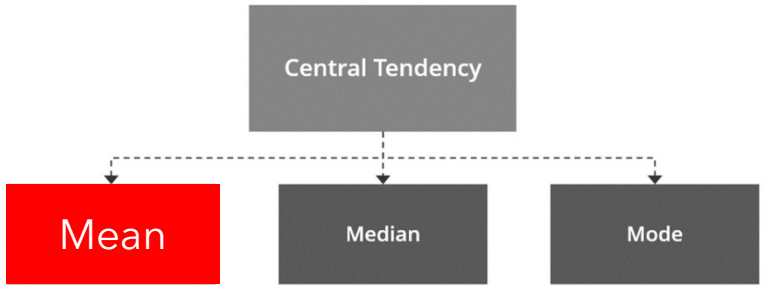
A measure of **Central Tendency** is a single value that attempts to describe a set of data by **identifying the central position** within that set of data.
In other words, the Central Tendency computes the "center" around which that data is distributed.

The three measures of Central Tendency are:

- Mean
- Median
- Mode

Measures of Central Tendency:

The Mean



- The arithmetic mean (often just called "**mean**") is the most common **measure of central tendency**

Pronounced x-bar

For a sample of size n:

The i^{th} value

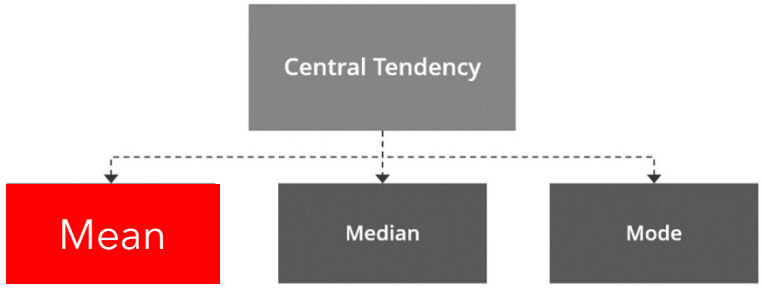
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Sample size

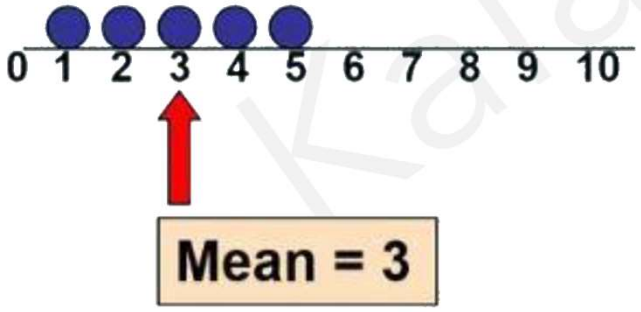
Observed values

Measures of Central Tendency:

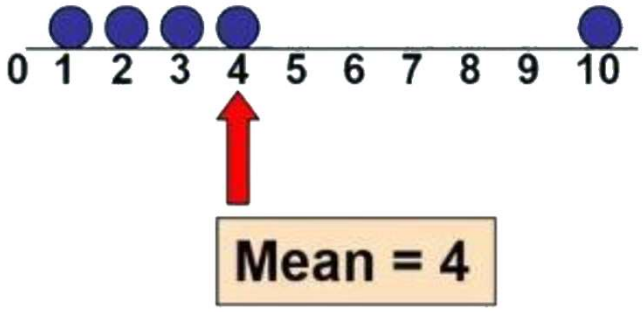
The Mean



- Mean = sum of values **divided** by the number of values
- Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The “Hotshot” Sales Executive

Kurt works as a sales manager at realityhomes.com. In the monthly sales review, Kurt reports that he will chieve his quarterly target of $1M.

Kurt claims his average deal size is $100,000 and he has 10 deals in his pipeline. Kurt’s boss Ross is very delighted with his numbers.

At the end of quarter, even after closing 8 deals Kurt fails to meet his target number and falls short by more than $500,000.

Discussion

- Why did Kurt fail to chieve his quarterly target?

- With 10 deals in pipeline and with average size of $100,00 and covering 8 of those deals how did he fail?

The Reality of "Hotshot" Salesman

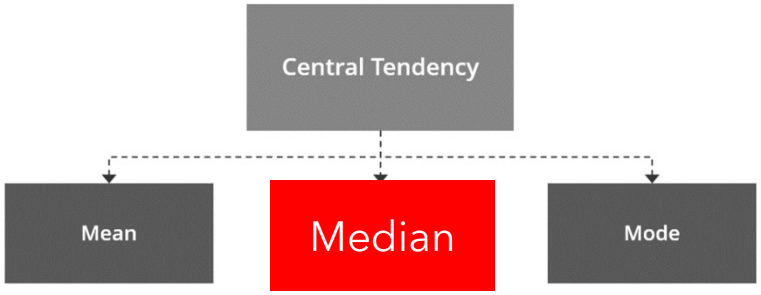
- Average deal size in pipeline =$100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

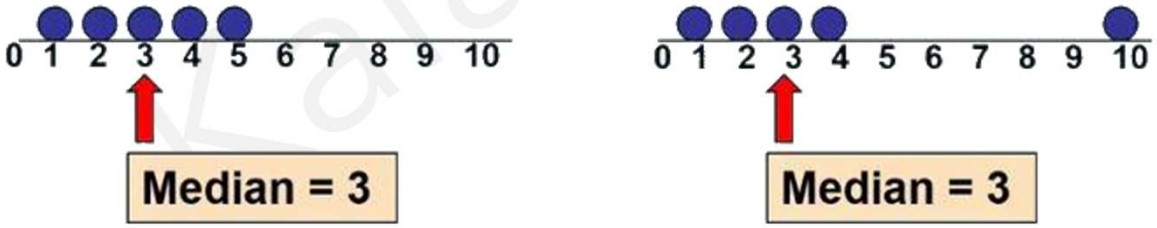
Median is less susceptible to the influence of outliers

Measures of Central Tendency:

The Median



- In an ordered array, the median is the “middle” number (50% above, 50% below)



- **Not** affected by extreme values (outliers)

Measures of Central Tendency:

Locating the Median

- First arrange the values in numerical order (smallest to largest) to find the median:

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, then the median is the middle number
- If the number of values is even, the median is the average of the two middle numbers
  - ❖ Note that (n+1)/2 is not the value of the median, only the position of the median in the ranked data

Median

The middle value

Example: 1, 2, 3, 4, 5

Median = 3

Median

Example: 1, 2, 3, 4, 5, 6

Two middle scores: 3, 4

To find the median, take the average of the two middle scores:
(3+4)/2 = 3.5

Median = 3.5

Median

Odd N: When there are an odd number of values, the median is the middle score

(1, 2, 3, 4, 5; N=5) median = 3

Even N: When there are an even number of values, the median is equal to the average of the two middle scores

(1, 2, 3, 4, 5, 6; N=6) median = 3.5

Median

Prior to calculating the median, be sure that the numbers are ordered from smallest to largest (don't pick the middle number of a set of numbers if they are not first ordered)

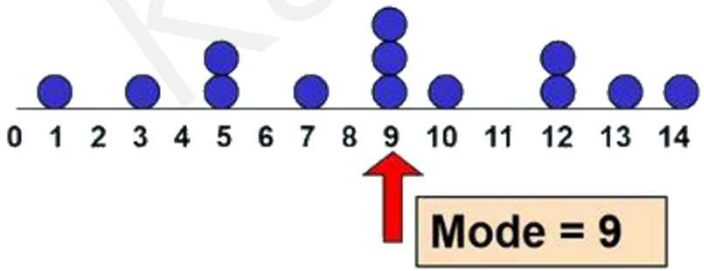
Example: 2, 3, 1, 3, 1, 6

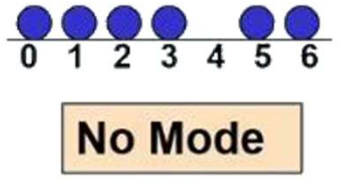
Reordered: 1, 1, 2, 3, 3, 6

Median = 2.5

Measures of Central Tendency:
the Mode

- Value that occurs **most often**
- There may be **no mode**
- There may be **several modes**





Mode

- The most frequently occurring score
- Example: 1, 2, 2, 2, 3, 3, 4
- Mode = 2

Mode

Example: 1, 2, 2, 3, 3, 4

Bimodal (two modes) = 2, 3

Mode

Example: 1, 2, 2, 3, 3, 4, 4

Multimodal (three or more modes) = 2, 3, 4

Measures of Central Tendency:

Review Example

House Prices:	
	$2,000,000
	$500,000
	$300,000
	$100,000
	$100,000
Sum	$3,000,000

- **Mean**: $3,000,000 / 5
  = **$600,000**

$$\bar{x} = \frac{\sum_{i=1}^{n} -x_i}{n}$$

- **Median**: middle value of ranked data
  (5+1)/2 = 3rd position
  = **$300,000**

$$\frac{n+1}{2}$$

- **Mode**: most frequent value
  = **$100,000**

Measures of Central Tendency:

Review Example

House Prices:	
	$2,000,000
	$500,000
	$300,000
	$150,000
	$130,000
	$120,000
Sum	$3,200,000

- **Mean**: $3,200,000 / 6 = **$533,333.33**

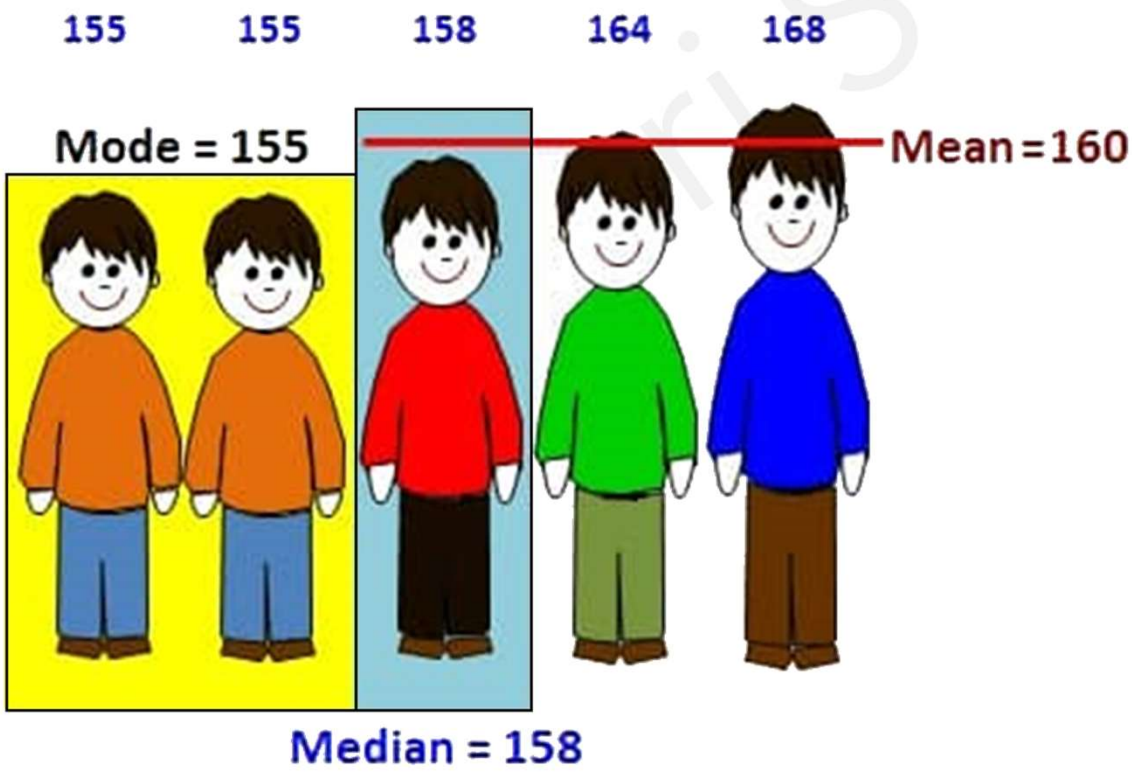
$$\bar{x} = \frac{\sum_{i=1}^{n} -x_i}{n}$$

- **Median**: middle value of ranked data
  (6+1)/2 = 3.5th position
  = **$255,000**

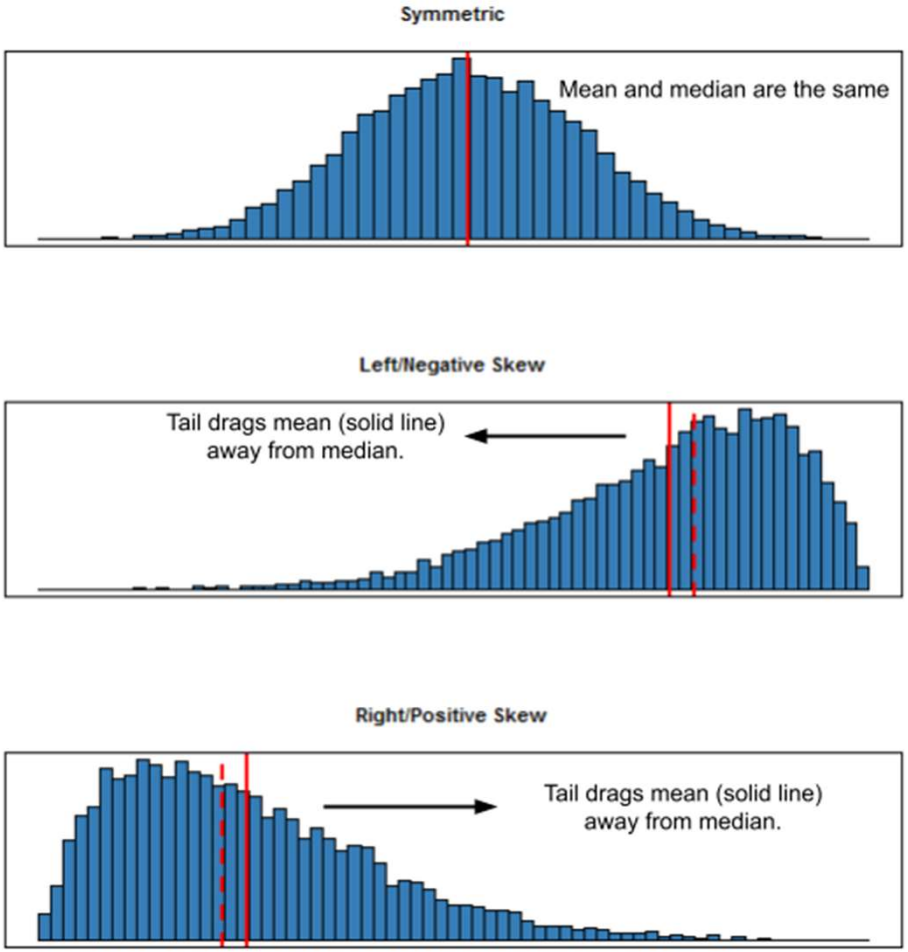
$$\frac{n+1}{2}$$

- **Mode**: most frequent value
  = **N/A**

Mean, Median & Mode Summary



Introduction to Statistics



Population Vs Sample

Population

VS

Sample

N -> Data size of Population

n -> Data size of sample

The Weighted Mean

Finding the Overall Mean for Two Separate Groups()

Weighted Mean Formula

Weighted Mean (In case of one sample) **=** $\frac{\sum WX}{\sum W}$

Weighted Mean $= \frac{(Mean\ Group\ 1)(N1)+(Mean\ Group\ 2)(N2)}{N1+N2}$

N1 = Size of first group

N2 = Size of second group

Weighted Mean Formula: Equal group sizes

N1 = 20; mean group 1 = 90

N2 = 20; mean group 2 = 110

$$\textbf{Weighted Mean} = \frac{(Mean\ Group\ 1)(N1)+(Mean\ Group\ 2)(N2)}{N1+N2}$$

$$\textbf{Weighted Mean} = \frac{(90)(20)+(110)(20)}{20+20} = \frac{1800+2200}{20+20} = \frac{4000}{40} = 100$$

(Exactly in the middle of the means: (90+110)/2 = 100)

Weighted Mean Formula: Unequal group sizes

N1 = 10; mean group 1 = 90

N2 = 30; mean group 2 = 110

$$\textbf{Weighted Mean} = \frac{(Mean\ Group\ 1)(N1)+(Mean\ Group\ 2)(N2)}{N1+N2}$$

$$\textbf{Weighted Mean} = \frac{(90)(10)+(110)(30)}{10+30} = \frac{900+3300}{10+30} = \frac{4200}{40} = 105$$

(Closer to mean of 110 than mean of 90: larger group size, greater influence)

Weighted Mean Formula: Unequal group sizes

N1 = 30; mean group 1 = 90

N2 = 10; mean group 2 = 110

$$\textbf{Weighted Mean} = \frac{(Mean\ Group\ 1)(N1)+(Mean\ Group\ 2)(N2)}{N1+N2}$$

$$\textbf{Weighted Mean} = \frac{(90)(30)+(110)(10)}{30+10} = \frac{2700+1100}{30+10} = \frac{3800}{40} = 95$$

(Closer to mean of 90 than mean of 110: larger group size, greater influence)

Weighted Mean

Summary Results:

Equal group sizes: N1 = 20; mean group1 = 90; N2 = 20; mean group 2 = 110

Weighted mean = 100 (**Exactly in the middle with equal N**)

Unequal group sizes: N1 = 10; mean group1 = 90; N2 = 30; mean group 2 = 110

Weighted mean = 105 (**Closer to mean with larger N**)

Unequal group sizes: N1 = 30; mean group1 = 90; N2 = 10; mean group 2 = 110

Weighted mean = 95 (**Closer to mean with larger N**)

Weighted Mean

The weighted mean takes into account the **size of each group**. When group sizes are equal, the weighted mean is exactly in-between the two group means. When group sizes are unequal, the mean with the larger N has greater influence, pulling the overall (weighted) mean of the two groups closer to it.

Agenda

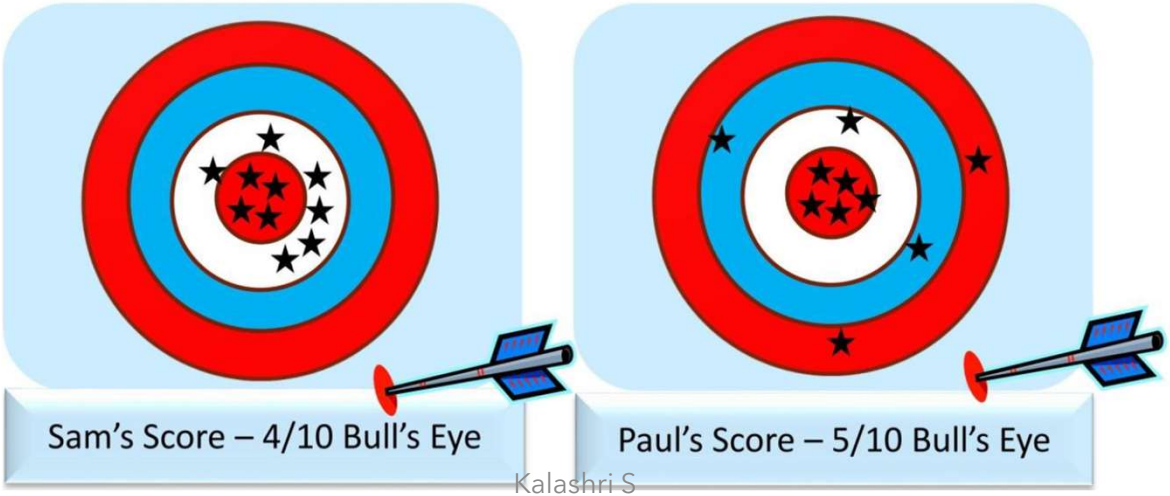
What is statistics?

- ✓ Central Tendency Measures
- ✓ **Dispersion Measures**
- ✓ Data Distributions

Dispersion Measures

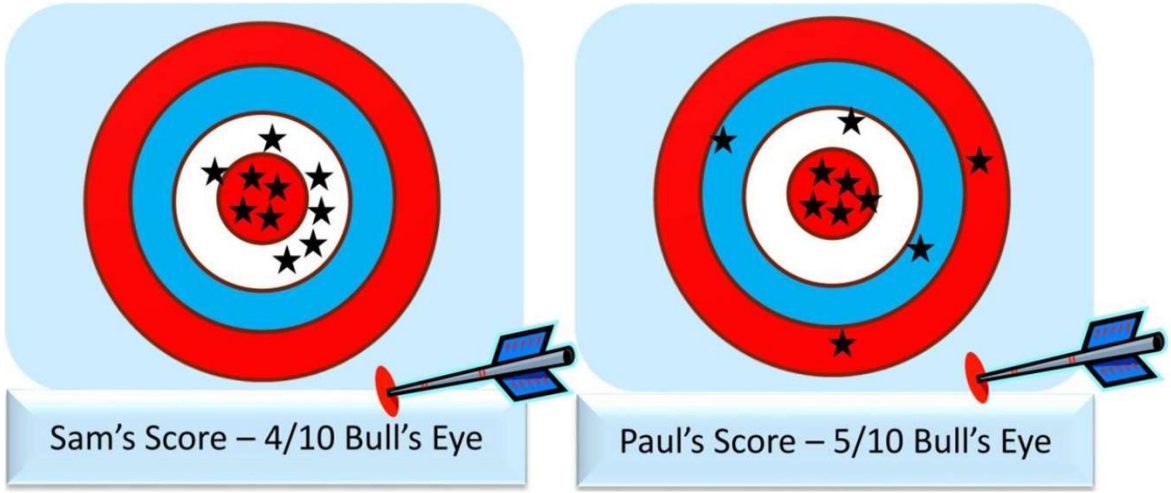
Bull's Eye

Sam and Paul are throwing darts at the local sports bar. A few of their friends start a betting pool. Both Sam and Paul shoot 10 practice shots each so that their friends can decide their bets.



Discussion

Who is better bet: Sam or Paul?

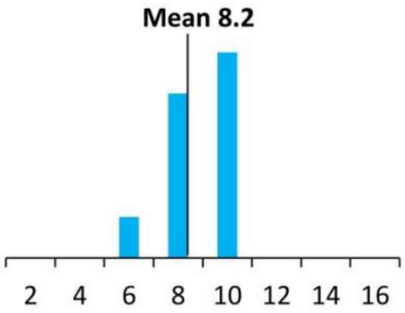


Who is the Better Agent?

Sheila and Sarah work as customer service agents in vsellhomes.com. During the annual performance review, the manager reviews their call handle times data. Bothe Sheila and Sarah have a average handle time of 8.2 minutes, which is as per the team's of < 10 minutes.

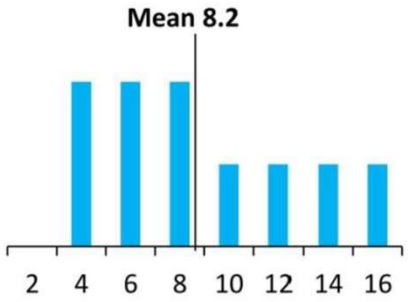
Sheila	
Sample Call #	Handle Time in Minutes
1	8
2	7
3	9
4	10
5	9
6	7
7	6
8	9
9	10
10	7

Average handle time = 8.2 minutes



Sarah	
Sample Call #	Handle Time in Minutes
1	15
2	5
3	8
4	9
5	4
6	12
7	7
8	13
9	6
10	3

Average handle time = 8.2 minutes



Discussion

Who is the better agent: Sheila or Sarah?

Should both of them be rated at par for meeting the team's target performance?

Dispersion Measures

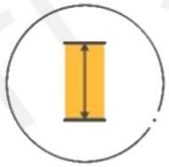
Measures of Dispersion describe the data spread or how far the measurements are from the centre.

The Measures of Dispersion are:

- Range
- Variance / Standard Deviation
- Mean Absolute Deviation
- Interquartile Range

Dispersion Measures: Examples

Range



Variance

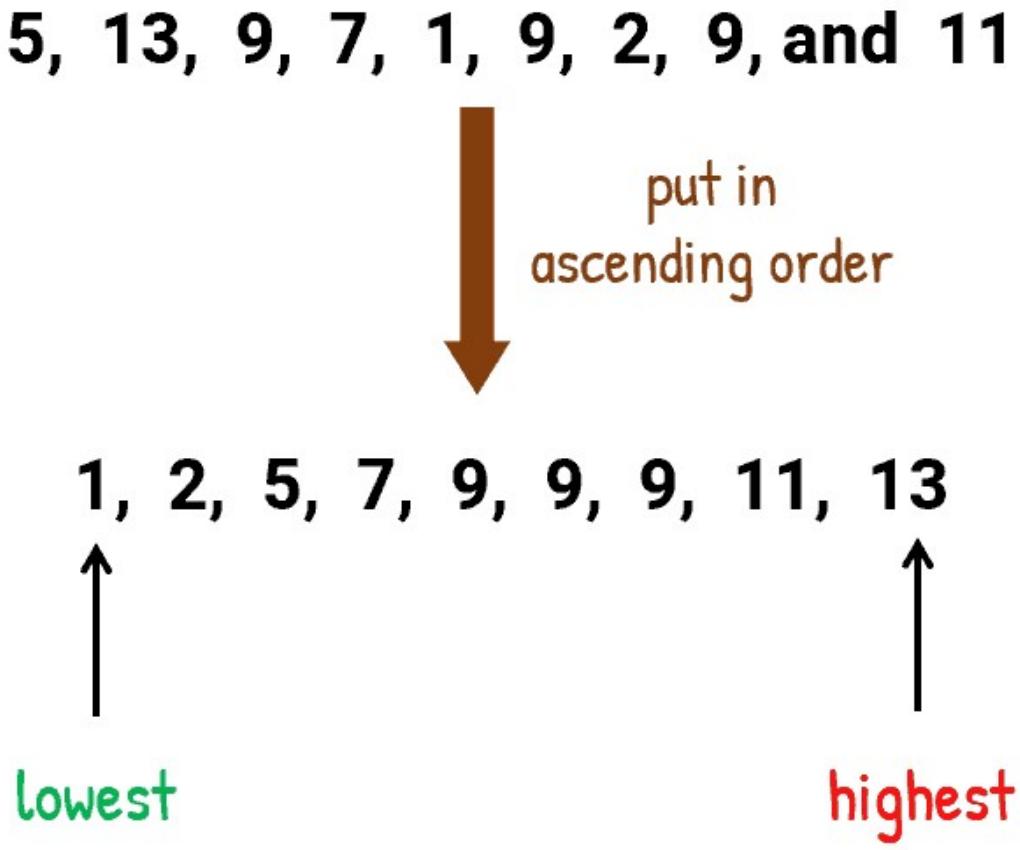
Interquartile Range

Standard Deviation

Range

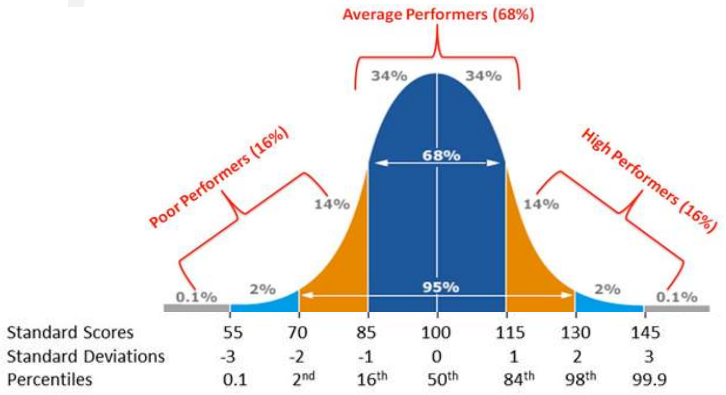
- Range is the difference between the highest and lowest data point in a distribution.
- Formula:

  Range (R) = Maximum – Minimum

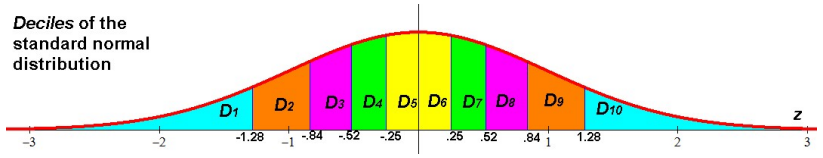


Dispersion Measures: Percentiles and Deciles

- **Percentiles**: Of a distribution are the 99 values that split the data into a hundred equal parts

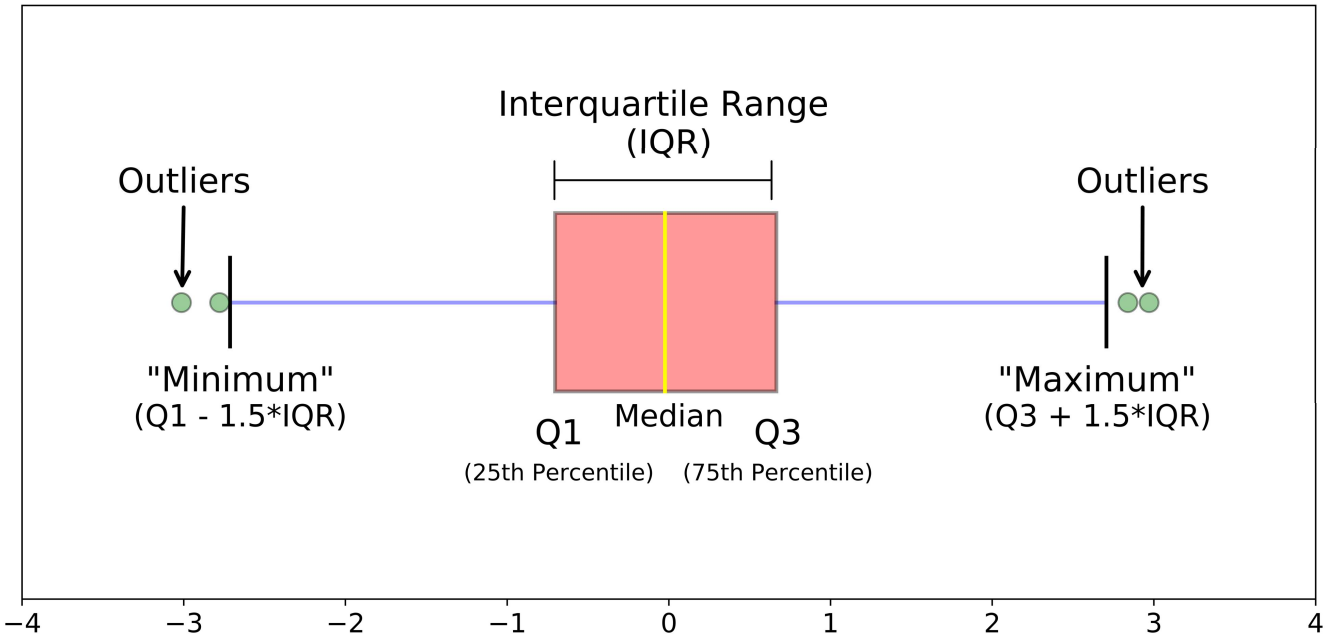


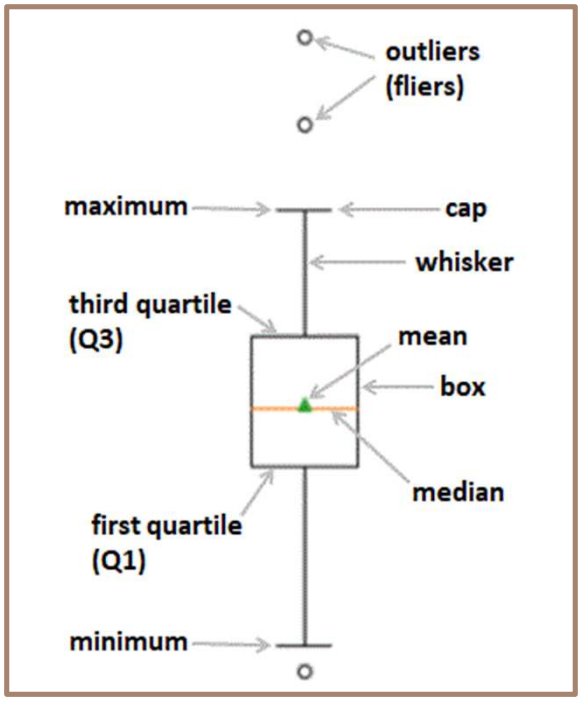
- **Deciles**: of a distribution are the 9 values that split the data into 10 equal parts



Quartile Measures

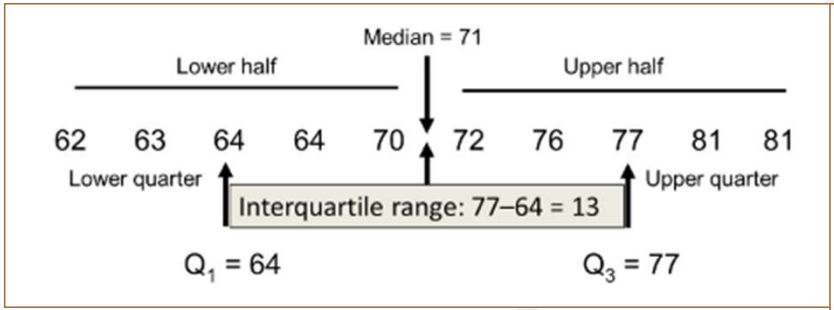
- Quartiles split the ranked data into 4 segments with equal number of values per segment

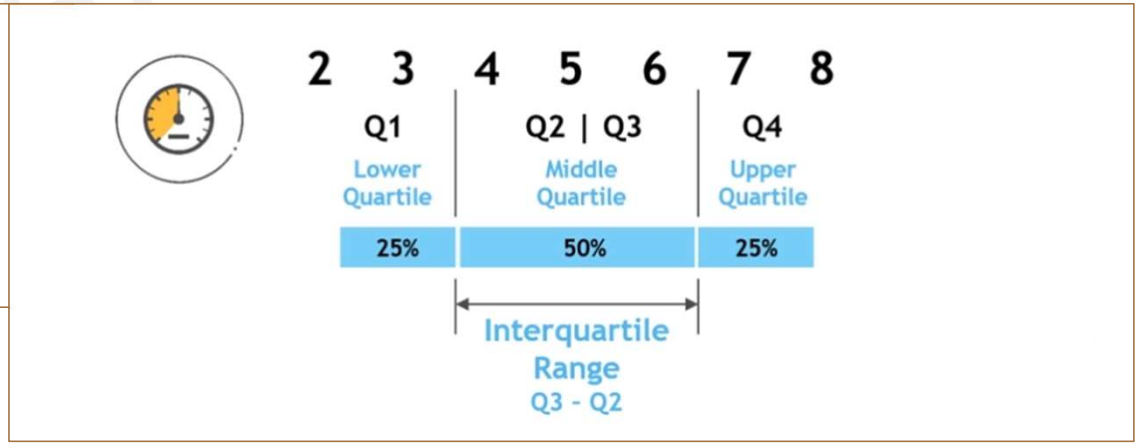




Interquartile Range (IQR)

- The IQR describes the middle 50% of values when ordered from lowest to highest.
- Majority of values (50%) closest to the center, which gives better prospective of data, or we can say unbiased data.

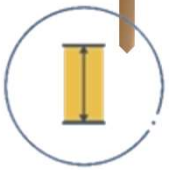




The Five Number Summary

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)
- $X_{largest}$

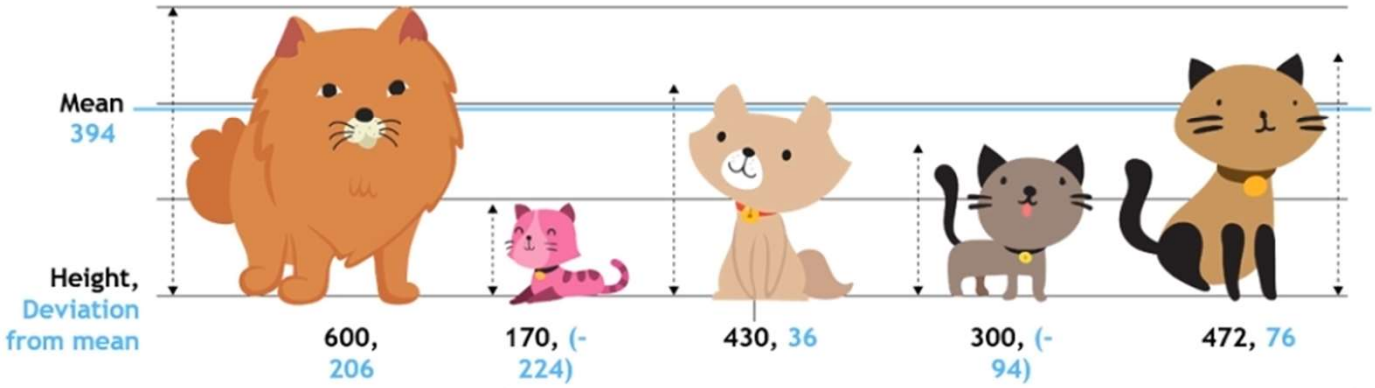


Variance

- Variance describe how much an element differ from mean.

$$var = s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

Variance is the average of the squared distances from the mean



$$Variance = \frac{\sum 206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = 21704\ mm^2$$

Standard Deviation

- The standard deviation in statistics that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
- Its symbol is σ (the Greek letter sigma).

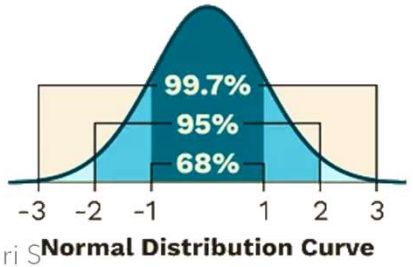
Calculating Standard Deviation

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

n = The number of data points

x_i = Each of the values of the data

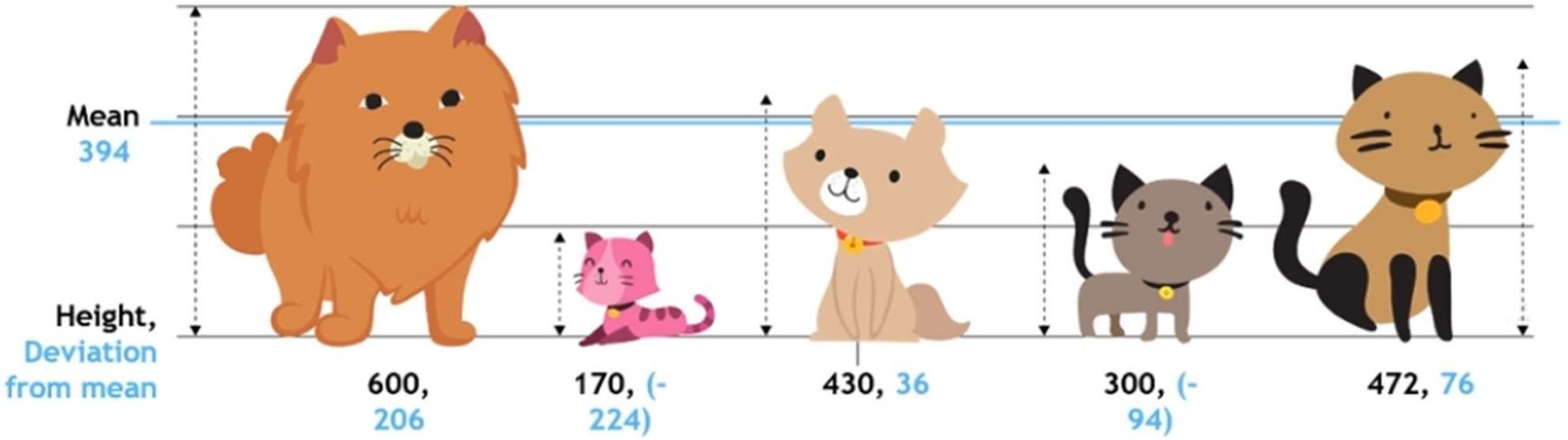
$\bar{x}$ = The mean of x_i



Normal Distribution Curve

Standard Deviation

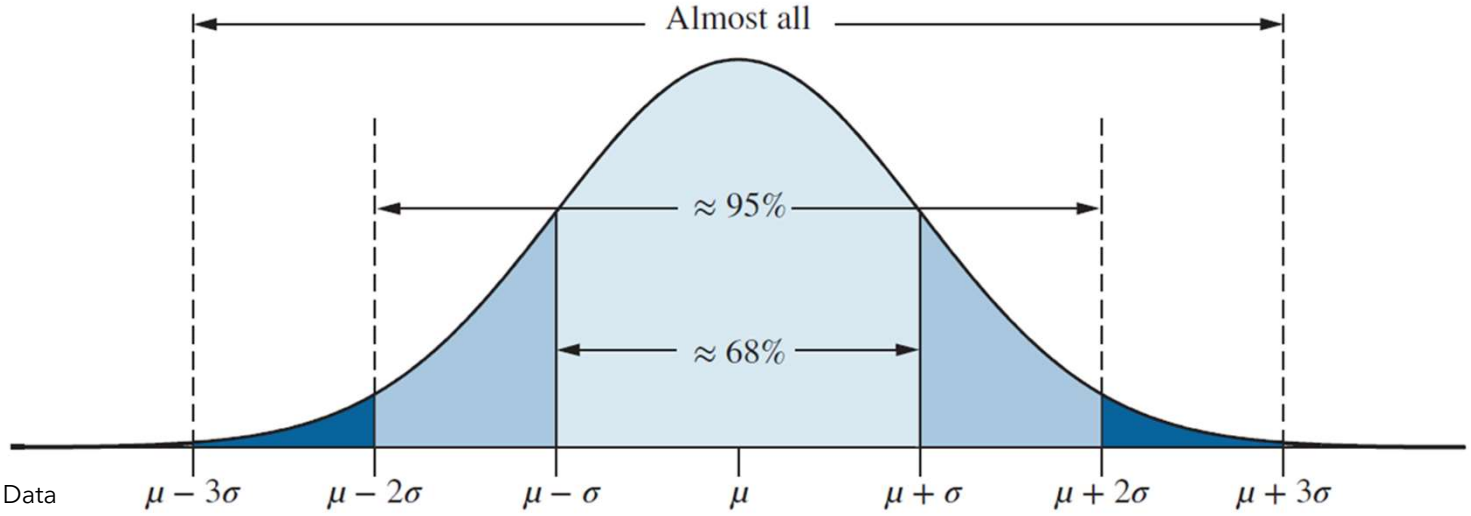
Standard Deviation is the square root of variance $\sqrt{var} = \sqrt{s^2} = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}$



$$SD = \sqrt{21704} = 147.3\ mm$$

The Empirical Rules

- The empirical rule approximates the variation of data (Entire population) in the bell-shaped distribution
- Approximately 68% of data in a bell shaped distribution is within 1 standard deviation of the mean or **$\mu \pm 1\sigma$**
- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean **$\mu \pm 2\sigma$**
- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations od the mean, **$\mu \pm 3\sigma$**



Note: Implemented on Continuous Data

Sample Population Variance & Standard Deviation

Sample/Population Variance & Standard Deviation

Sample Variance:

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

Sample Standard Deviation:

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

Population Variance:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

Population Standard Deviation:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

Measures of Position:

Locating Extreme Outliers: Z-Score

- To compute the Z-**score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

Measures of Position:

Locating Extreme Outliers: Z-Score

$$Z = \frac{X - \overline{X}}{S} \quad \text{OR} \quad Z = \frac{X - \mu}{\sigma}$$

where X represents the data value

$\overline{X}$ is the sample mean

S is the sample standard deviation

Measures of Position:

Locating Extreme Outliers: Z-Score

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

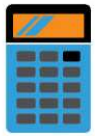
$$Z = \frac{X - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

Measures of Variation:

The Coefficient of Variation = CV (Covariance)

- The CV is a dimensionless number
- Always in **percentage (%)**
- It can be used to **compare the variation among two or more sets** of data
- The CV is particularly useful when comparing dispersion in datasets with : markedly different means or different units of measurement
- This is also called as **Covariance**



$$\text{Coefficient of Variation Formula} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

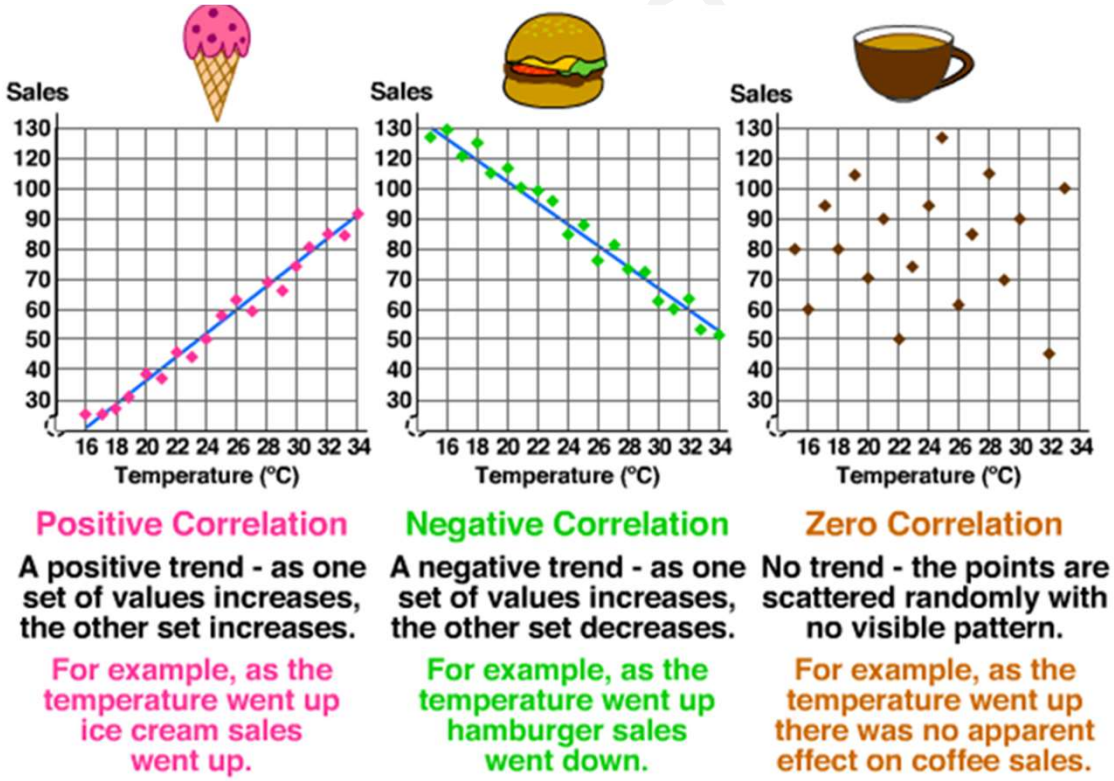
CV ~ 100 means two variables are 100% differ

CV ~ 0 means two variables are 0% differ

Note: If we multiply any formula with number that number becomes coefficient

Correlation coefficient

- The three scatter plots below show a positive correlation, a negative correlation and no correlation.
- To manage ordering supplies more effectively, three scatter plots were made to see if there was any correlation between daily temperatures and sales od ice cream, burgers and coffee.



Correlation coefficient

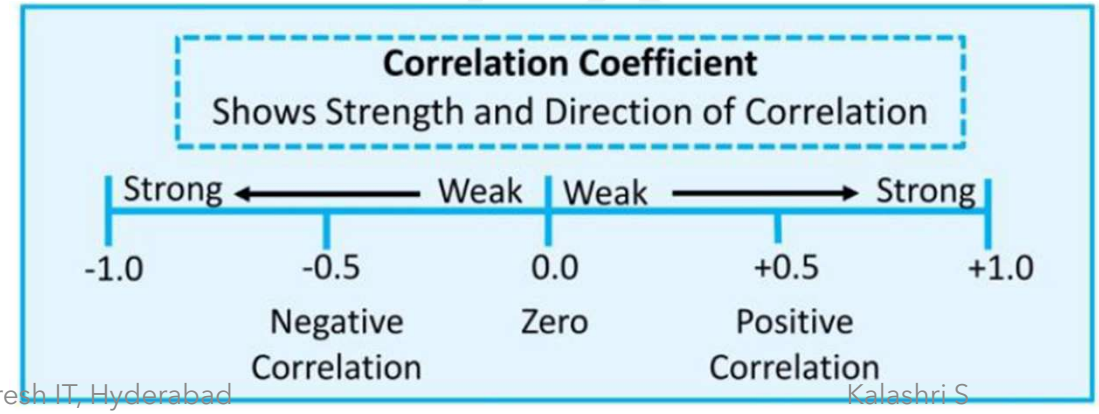
- The **correlation coefficient** of two variables in a data set equals to their covariance divided by the product of their individual standard deviations.
- It is a normalized measurement of how the two variables are linearly related.
- Correlation Coefficient (also called Pearson Correlation Coefficient)

$$r \text{ or } R = \frac{Covariance}{Product\ of\ Standard\ Deviation\ of\ the\ variable}$$

Correlation coefficient

The Correlation Coefficient ranges from -1 to 1.

- +1 indicates perfect collinearity, which means, if one value increases, the other also increases in the same proportion
- -1 indicates perfect negative collinearity, which means, if one value decreases, the other increases in the same proportion
- Zero indicates no relationship between the variables



Coefficient of Variation ~ 100%
Correlation Coefficient ~ 0 (Very less correlated)

Coefficient of Variation ~ 0%
Correlation Coefficient ~ 1 (Highly correlated)

Shape of a Distribution

Shape of a Distribution

- Describes how data are distributed
- Three useful **shapes** related to statistics are:
  - ❑ **Symmetric** 

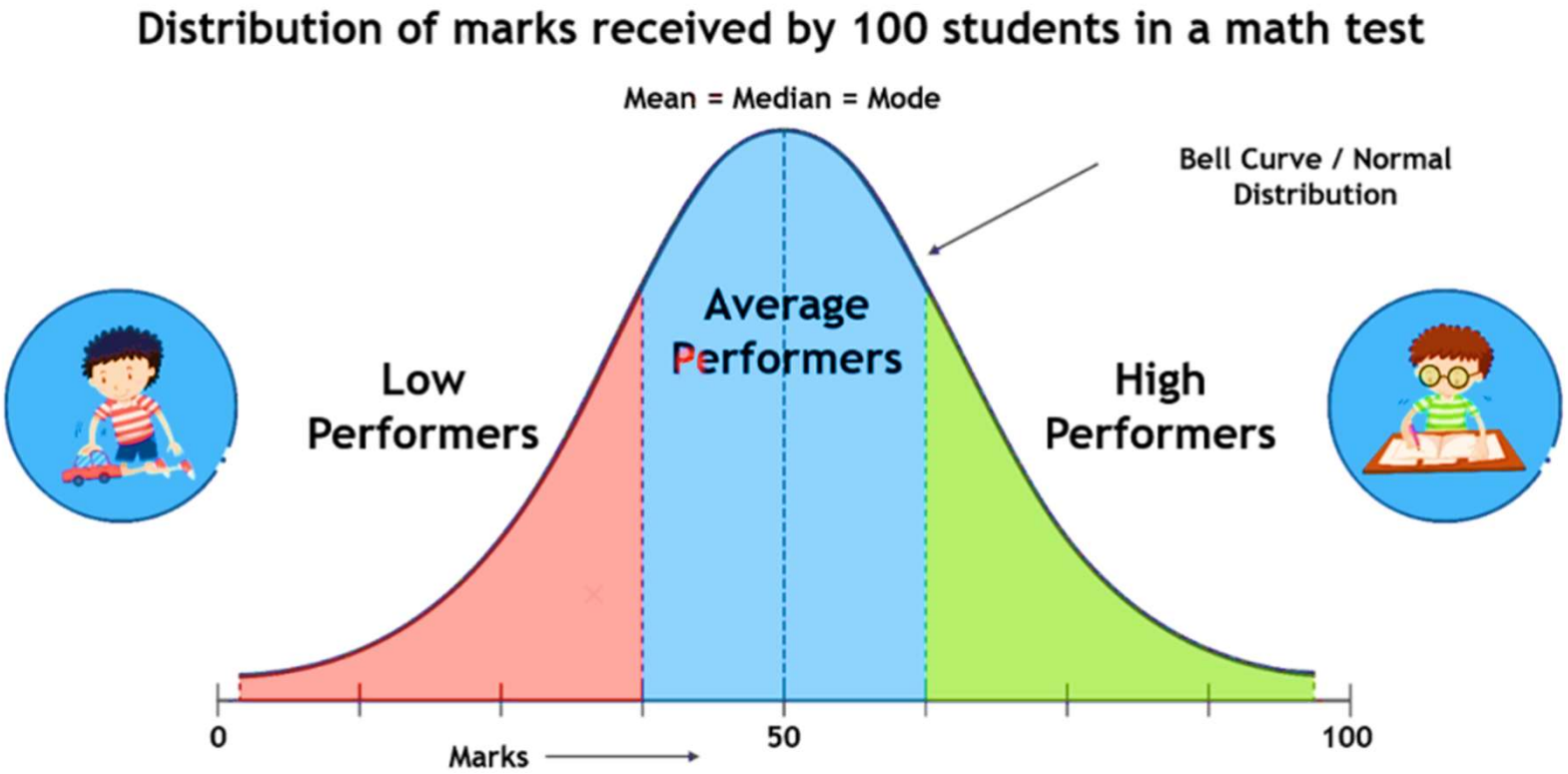
    Measures the amount of **asymmetry** in a distribution
  - ❑ **Skewness** 

    Measures the amount of **asymmetry** in a distribution
  - ❑ **Kurtosis** 

    Measures the **relative concentration** of values in the **center** of a distribution as compared with the **tails**

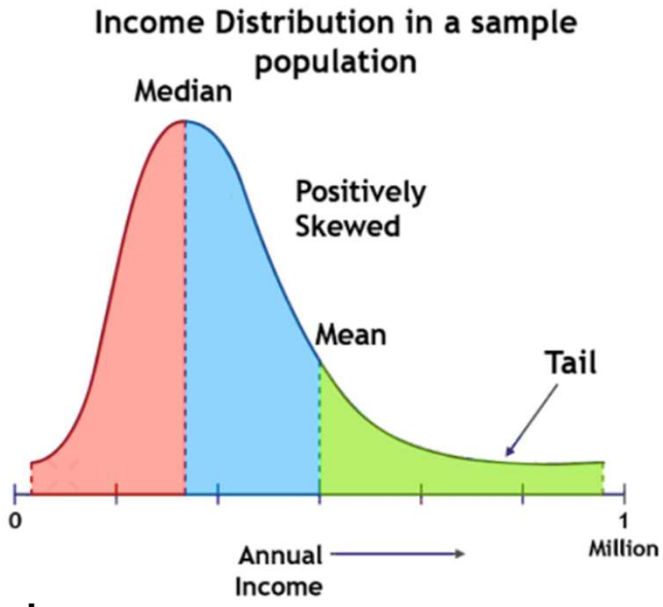
Symmetric

Symmetric means if the distributions having the same shape on both sides of the center. Moreover, those with only one peak are known as a normal distribution.



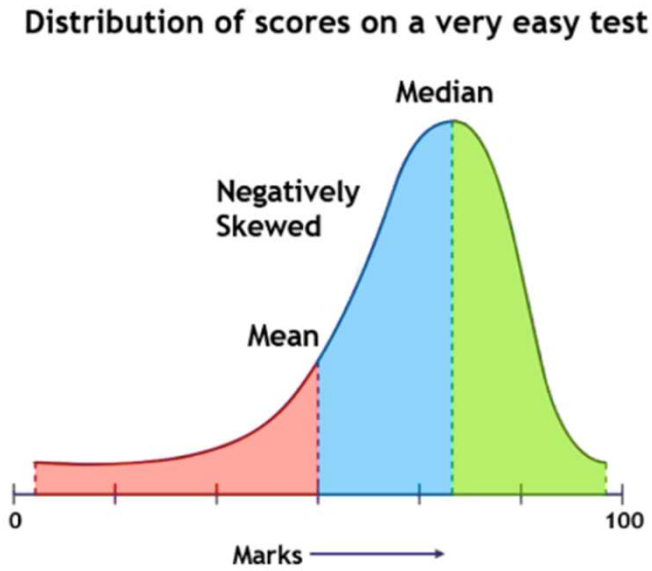
Skewness

Skewness refers to the degree of asymmetry in a distribution. And, asymmetry reflects extreme scores in a distribution. Moreover, it includes positive and negative skewness.



1. Positively skew

- The mean is greater than the median and
- The mean is sensitive to each score in the distribution
- Moreover, it is subject to large shifts when the sample is small and contains extreme scores



2. Negatively skew

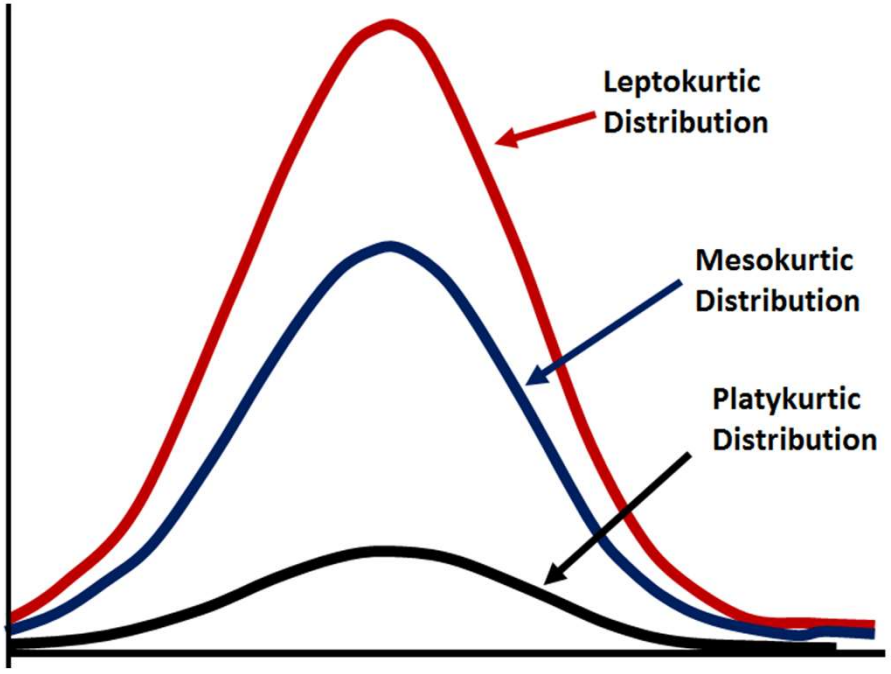
- Reflects bunching of numbers in the upper part of the distribution
- Moreover, it has fewer scores at the lower end of the measurement scale

Kurtosis

Kurtosis explain what is the concentration of data at center point.

Like skewness, Kurtosis is a descriptor of shape and it describes the shape of the of the distribution in terms of height or flatness.

There are three types of kurtosis: Mesokurtic, Leptokurtic, and Platykurtic.



Mesokurtic: Distributions that are moderate in breadth and curves with a medium peaked height.

Leptokurtic: More values in the distribution tails and more values close to the mean (i.e. sharply peaked with heavy tails)

Platykurtic: Fewer values in the tails and fewer values close to the mean (i.e. the curve has a flat peak and has more dispersed scores with lighter tails).

Probability Distribution

Why Probability is Important in Statistics?

- Statisticians use the basic ideas of probability to draw conclusions about populations by studying samples drawn from them.
- **Sampling** an individual from a population is a probability experiment.
- The collection of all the possible outcomes of a probability experiment is called a **sample space**.
- The population is the sample space, and the members of the population are equally likely outcomes. For this reason, the ideas of **probability are fundamental to statistics**.

Sampling

Agenda

Sampling

- Methods
- Estimation of Sample Size

Discussion

- Why Sampling?
- Why not analyze population data?

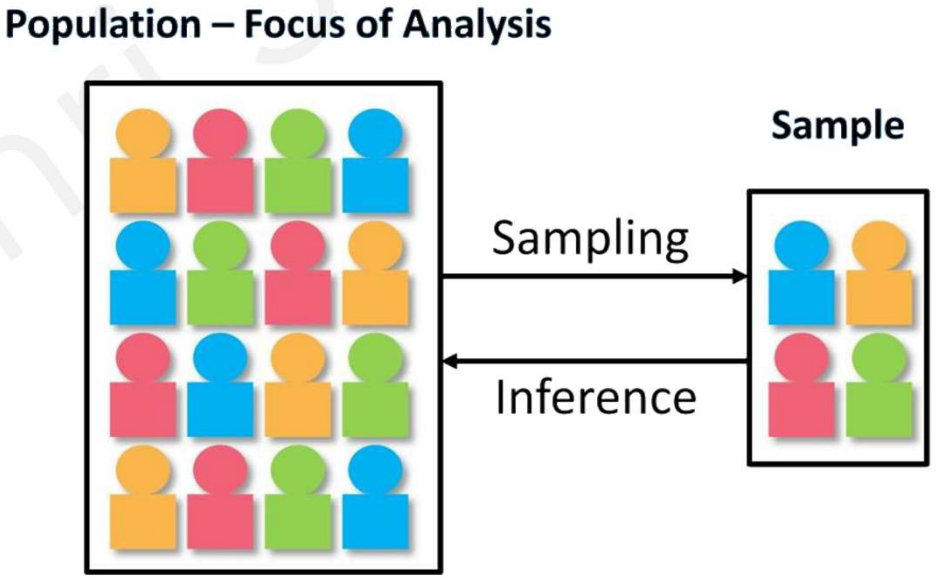
Population and Sample

- Population:

  A complete set of items that have common properties which, are the subject of statistical analysis

- Sample:

  A subset of the population of a manageable size selected through a defined procedure

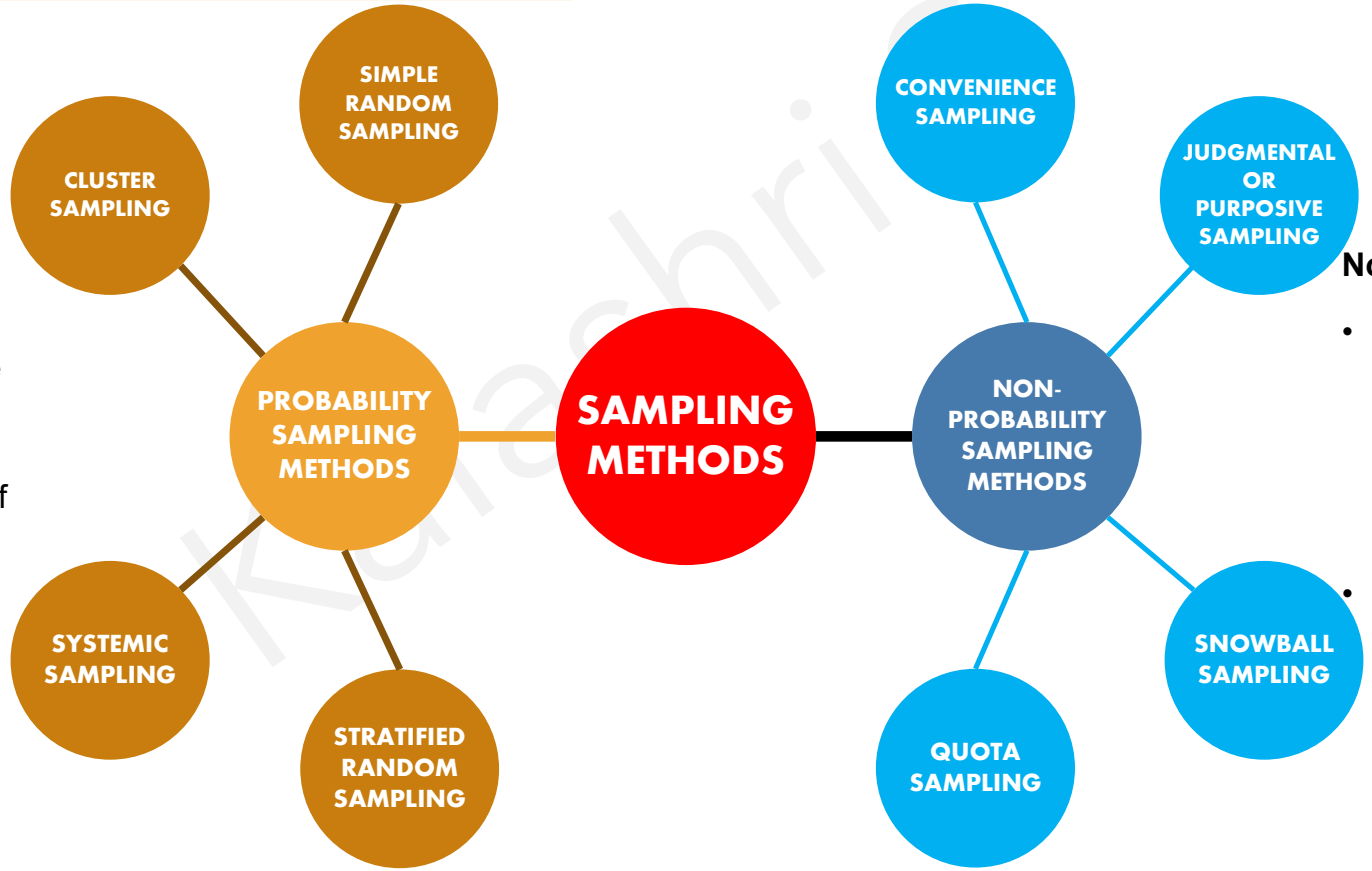


Why Sampling?

Saves Cost	Less expensive to study the sample than the population
Saves Time	Less time needed to study the sample than the population
Accuracy	Sampling process is designed and conducted in a systematic manner by skilled personnel so that the expected results are accurate
Complete	No missing units and no duplication

Probability Sampling:

- Is the Sampling method where each member if the population has a known probability (non-zero) of being selected as a part of the sample
- Sample is not biased

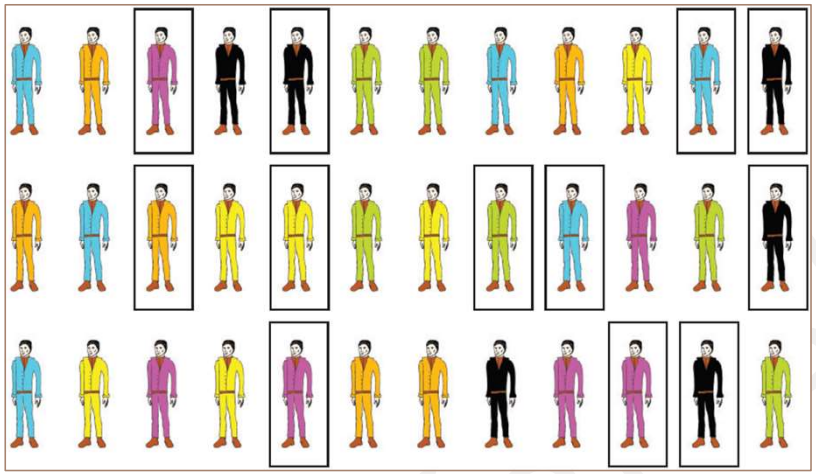


Non - Probability Sampling:

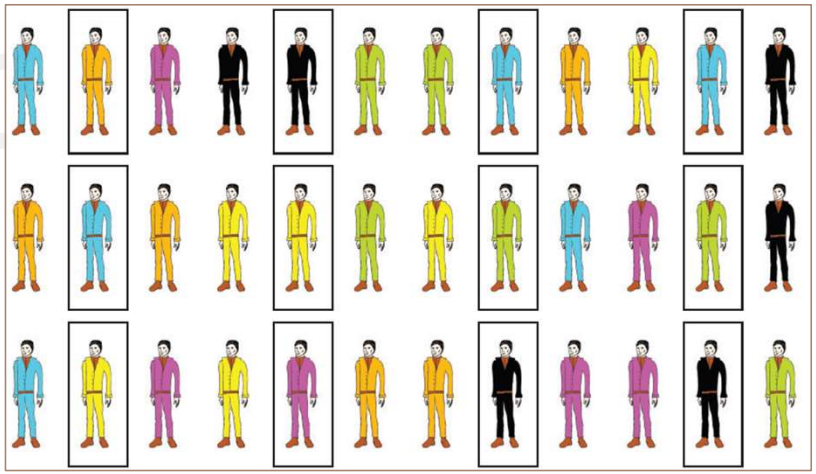
- Is the Sampling method where some members of the population have no chance of being selected
- Exclusive of samples can lead to sampling bias

Probability Sampling

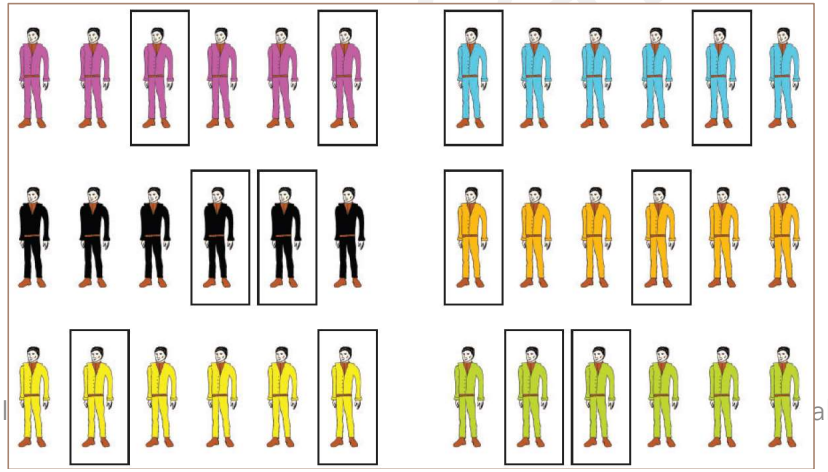
Simple Random Sampling (SRS)



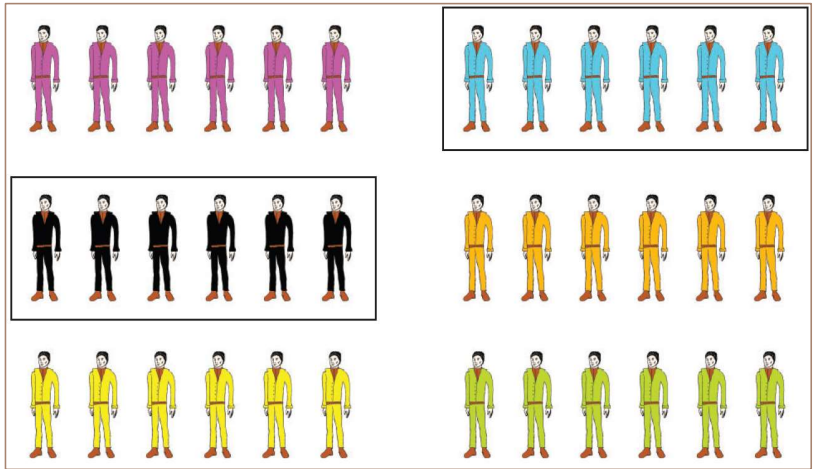
Systematic Sampling



Stratified Random Sampling



Cluster Sampling



Sampling

There are 300 employees in a certain company. The Human Resources department wants to draw a simple random sample of 20 employees to fill out a questionnaire about their attitudes toward their jobs. Describe how can you select 20 employees.

Solution: Simple random sampling

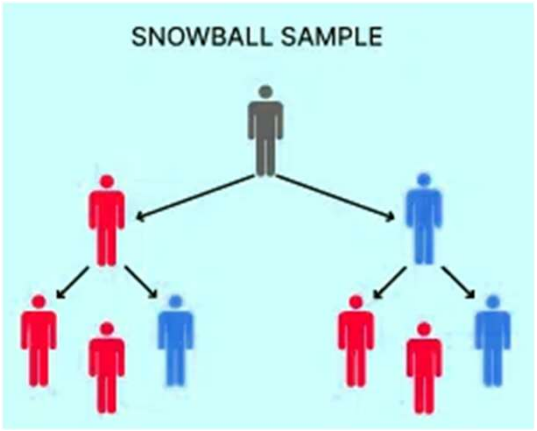
Sampling?

- The professor wants to draw a sample of 50 students to fill out a questionnaire about which sports they play. The professor's 10:00 A.M. class has 50 students.
- She uses the first 20 minutes of class to have the students fill out the questionnaire. Is this a simple random sample?

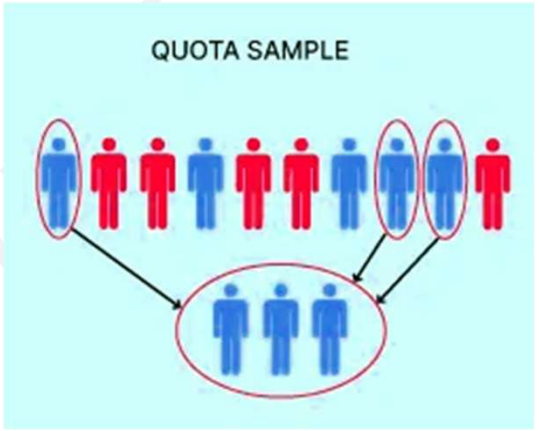
Solution:

- No. A simple random sample is like a lottery, in which each student in the population has
- an equal chance to be part of the sample. In this case, only the students in a particular class
- had a chance to be in the sample.

Non - Probability Sampling

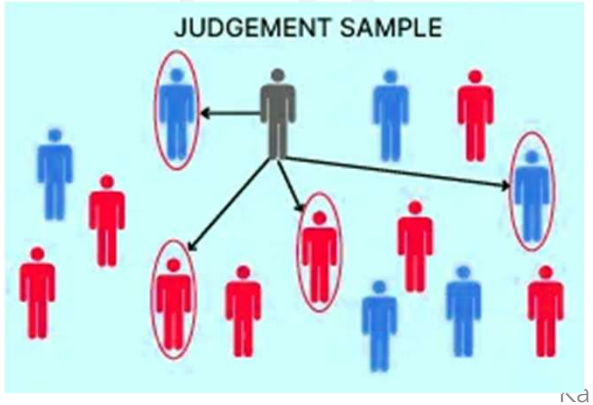


Snowball Sampling
Selection happens by referral

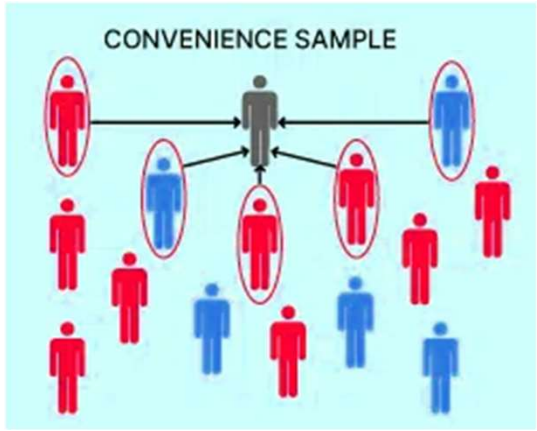


Quota Sampling
Population is divided into subgroups before expert judgment for sample selection from each subgroup

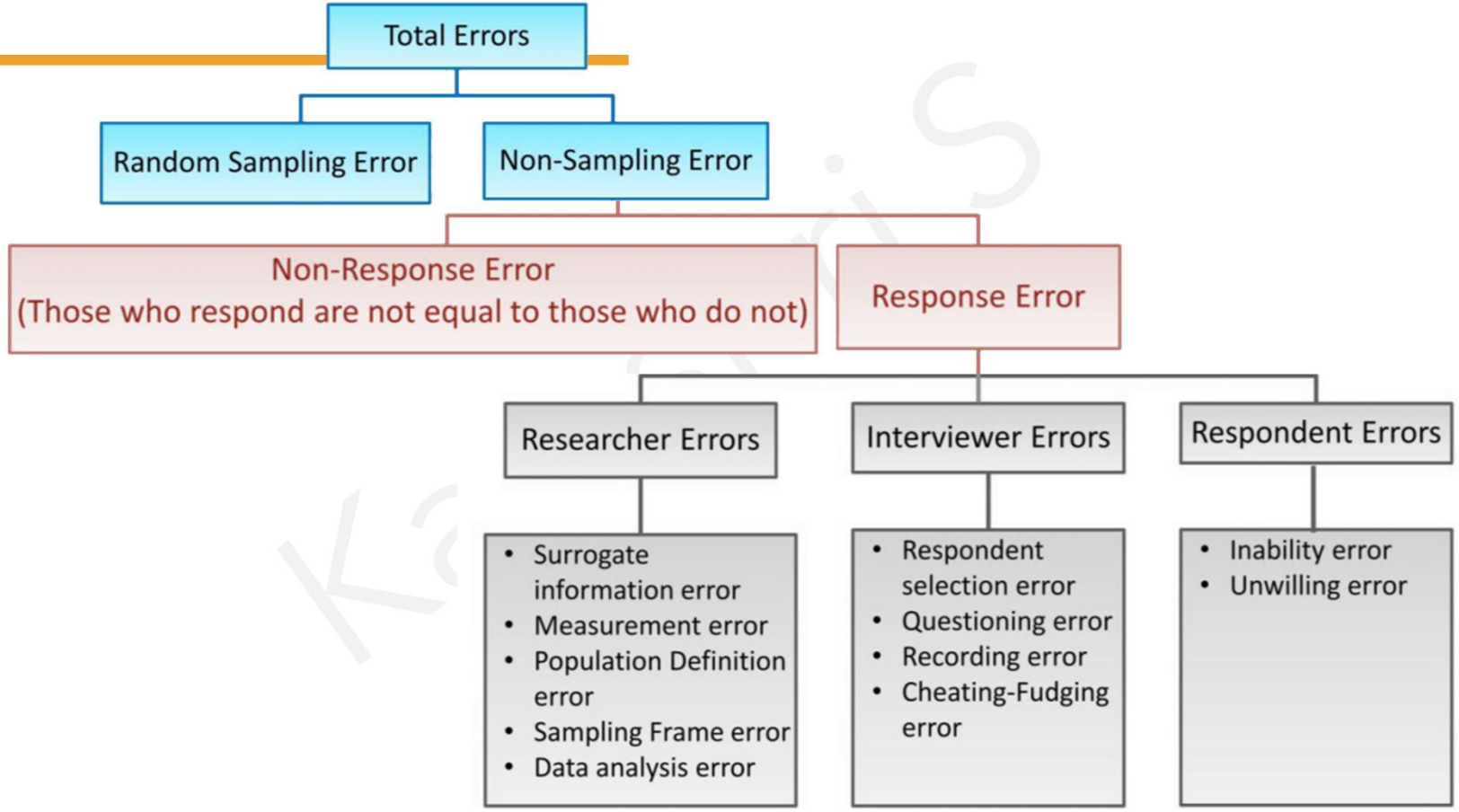
Judgmental Sampling
Researcher uses expert judgement for sample selection



Convenience Sampling
Sample is selected, based on their availability

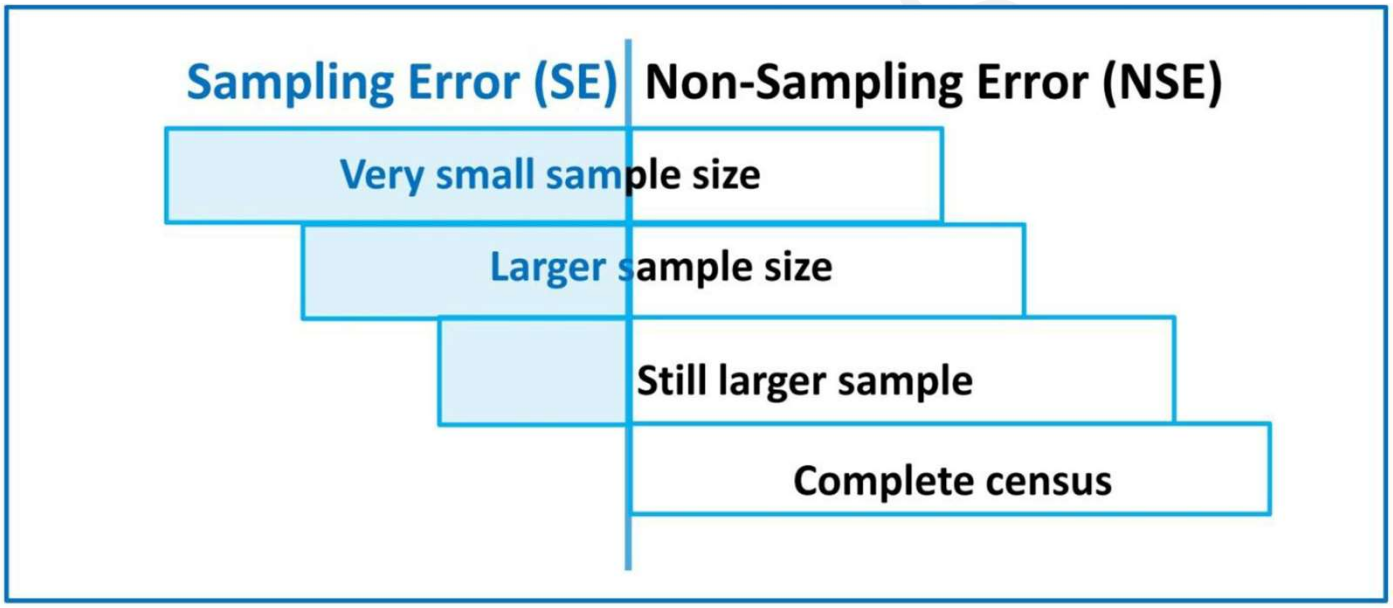


Errors in Data Collection



Non-Sampling errors are the major contributors to "Total Errors".

Sampling vs. Non-Sampling Errors



Increasing the sample size can reduce sampling error. However, the reduction in non-sampling error can happen only by ensuring a rigorous sampling process

Introduction to Statistics

Distribution: Example

- What is the probability that the dice will land on 6 in each trial –

Ans: 1/6 or 16.66%

Slide 92

k1 KS, 24-04-2022

Discussion

Which of the following random variables are discrete and which are continuous?

a. The number that comes up on the roll of a die

b. The height of a randomly chosen college student

c. The number of siblings a randomly chosen person has

d. Amount of electricity used to light a randomly chosen classroom

Discrete probability distribution: Example

A fair coin is tossed twice. Let *X* be the number of heads that come up. Find the probability distribution of *X*.

First Toss	Second Toss	X = Number of Heads
H	H	2
H	T	1
T	H	1
T	T	0

Solution

There are four equally likely outcomes to this probability experiment, listed below. For each outcome, we count the number of heads, which is the value of the random variable *X*.

Discrete probability distribution: Example

There are three possible values for the number of heads: 0, 1, and 2. One of the four outcomes has the value ‘‘0,’’ two of the outcomes have the value ‘‘1,’’ and one outcome has the value ‘‘2.’’ Therefore, the probabilities are

$$P(0) = \frac{1}{4} = 0.25 \qquad P(1) = \frac{2}{4} = 0.50 \qquad P(2) = \frac{1}{4} = 0.25$$

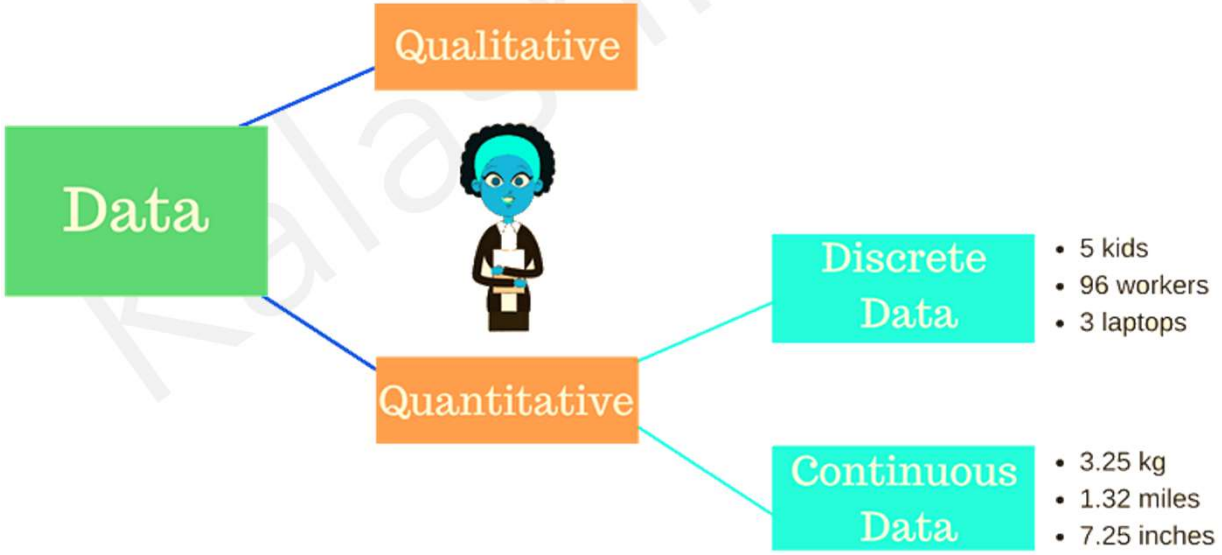
Probability Distribution of X

x	0	1	2
$P(x)$	0.25	0.50	0.25

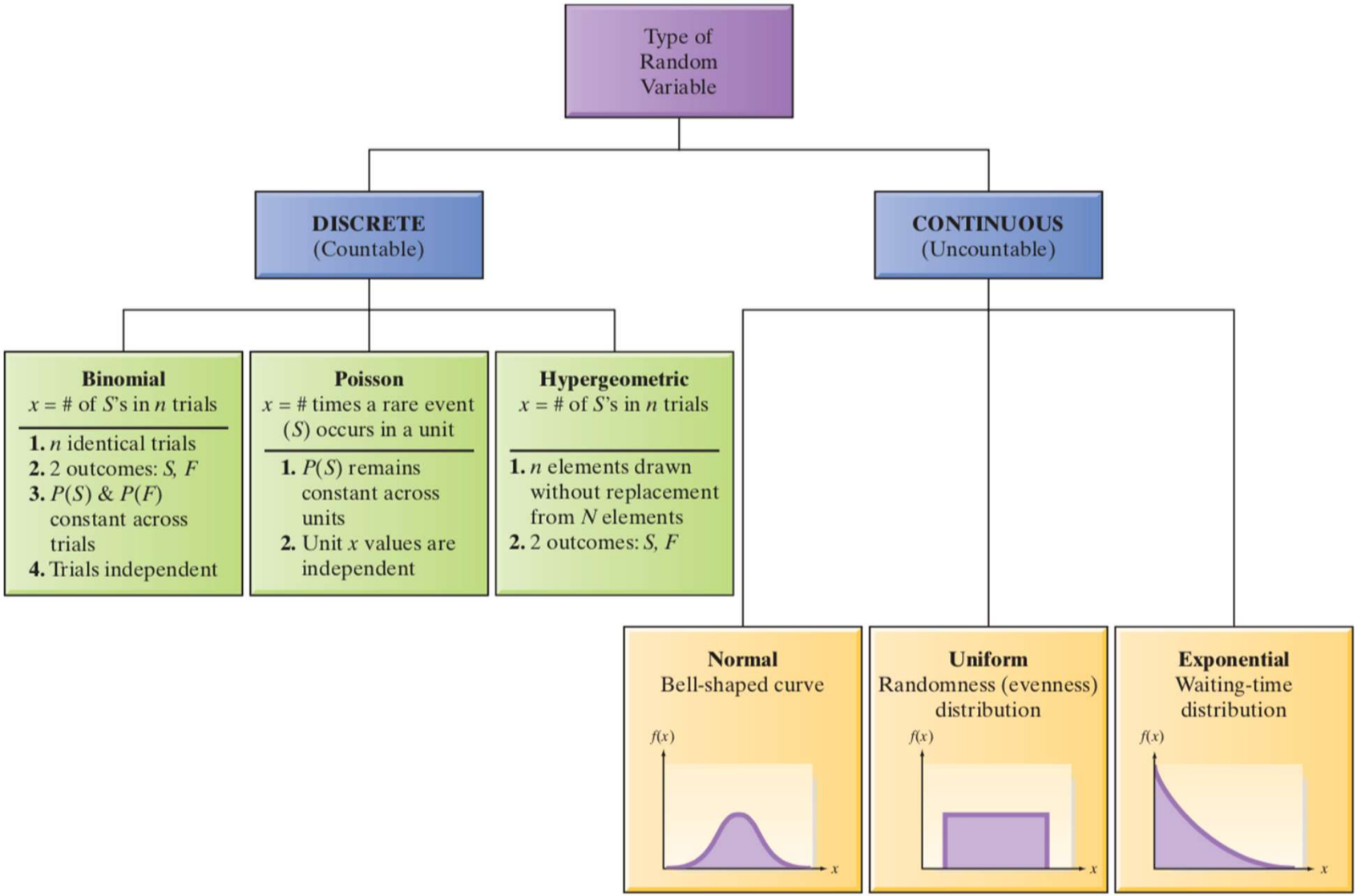
Discrete probability distributions satisfy two properties. First, since the values $P(x)$ are probabilities, they must all be **between 0 and 1**. Second, since the random variable always takes on one of the values in the list, the sum of the **probabilities must equal 1.**

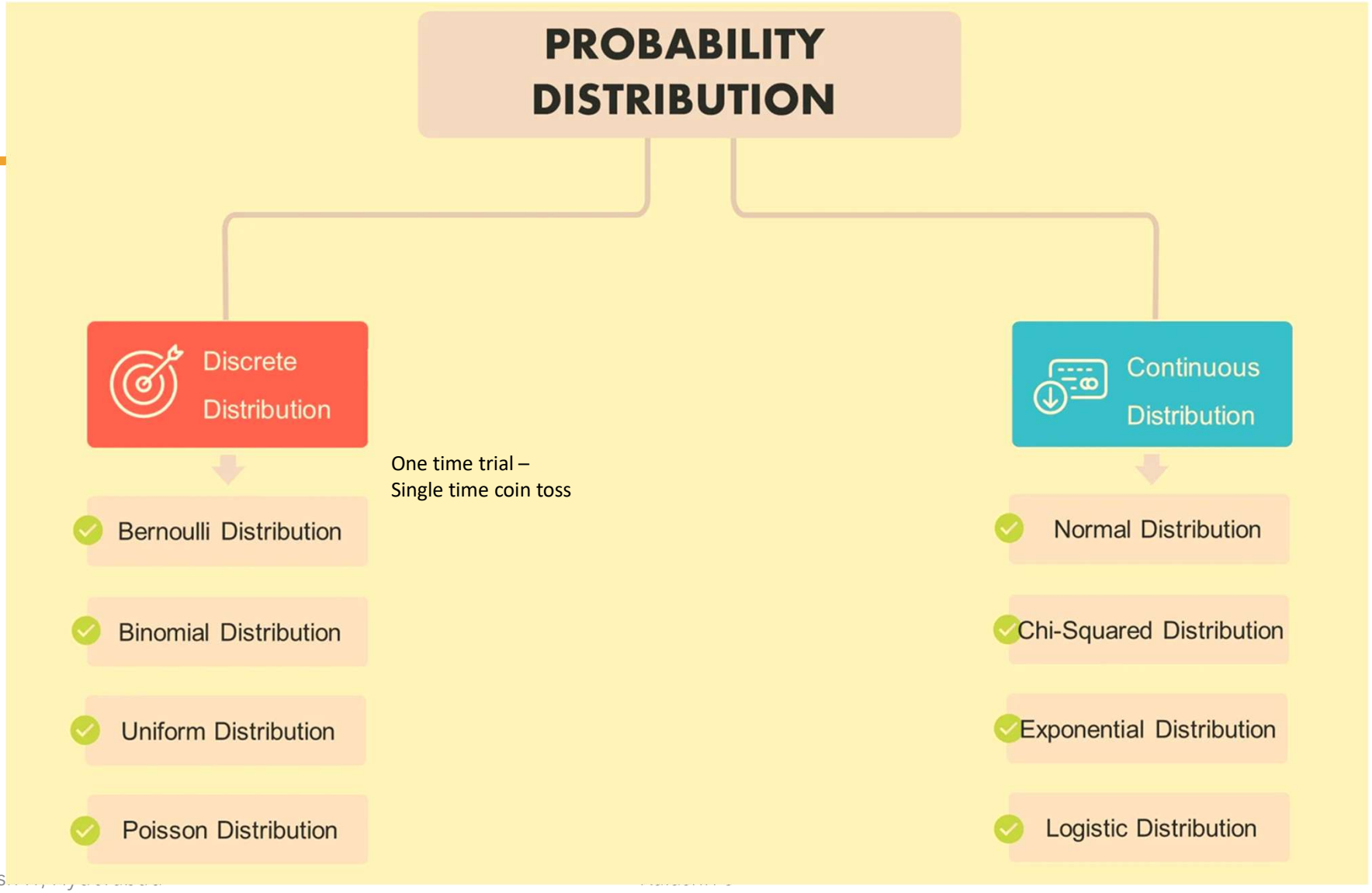
Random Variable

- A **random variable** is a numerical outcome of a probability experiment



Guide to Selecting a Probability Distribution



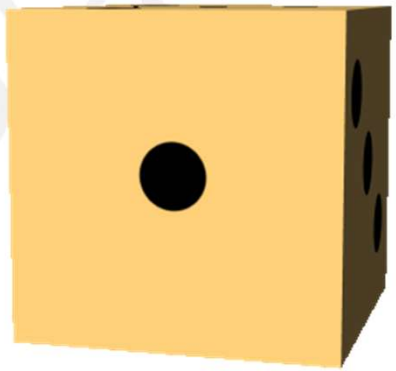

PROBABILITY DISTRIBUTION
Discrete Distribution
Continuous Distribution
One time trial –
Single time coin toss
Bernoulli Distribution
Binomial Distribution
Uniform Distribution
Poisson Distribution
Normal Distribution
Chi-Squared Distribution
Exponential Distribution
Logistic Distribution

Binomial Distribution

- Binomial distribution to model binary data, such as coin tosses

Uniform Distribution

- Uniform distribution to model multiple events with the same probability, such as rolling a die.



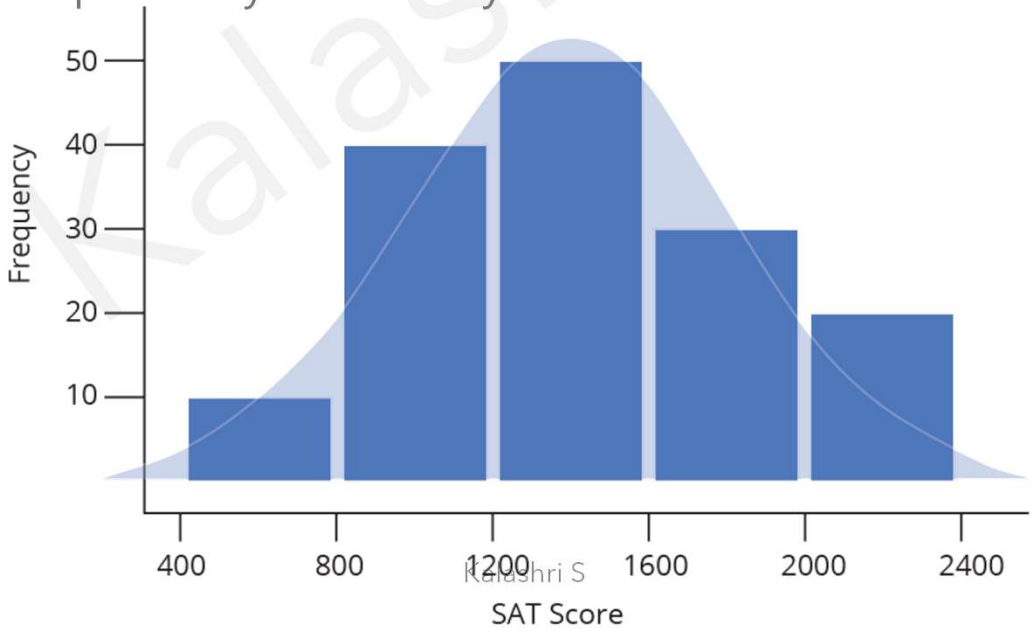
Poisson Distribution

- Poisson distribution to model count data, such as the count of library book checkouts per hour.

Poisson Distribution	
Number of cars passing in a drive-thru in a hour	
Number of phone calls received in a call center in an hour	
Number of machines breakdown in a month	

Normal Disctribution

- Normal/Gaussian Distribution is a bell-shaped graph which encompasses two basic terms- mean and standard deviation.
- It is a symmetrical arrangement of a data set in which most values cluster in the mean and the rest taper off symmetrically towards either extreme.

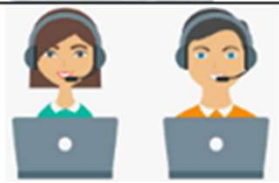


Chi-squared Distribution

- We will discuss later

Exponential Distribution

- The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.

Poisson Distribution		Exponential Distribution
Number of cars passing in a drive-thru in a hour		Number of minutes between car arrivals in drive-thru
Number of phone calls received in a call center in an hour		The time it takes for a call center executive respond to a caller
Number of machines breakdown in a month		Time between machine breakdown

Logistic Distribution

Central Limit Theorem

Example

There are 10,000 families in a certain town. They are categorized by their type of housing as follows.

Own a house	4753
Own a condo	1478
Rent a house	912
Rent an apartment	2857

A pollster samples a single family at random from this population.

a. What is the probability that the sampled family owns a house?

b. What is the probability that the sampled family rents?

Solution

a. The sample space consists of the 10,000 households. Of these, 4753 own a house, so the probability that the sampled family owns a house is

$$P(\text{Owns a house}) = \frac{4753}{10{,}000} = 0.4753$$

b. The number of families who rent is $912 + 2857 = 3769$. Therefore, the probability that the sampled family rents is

$$P(\text{Rents}) = \frac{3769}{10{,}000} = 0.3769$$

- We have seen Shape of Distribution

Introduction to Statistics

Introduction to Statistics

Agenda

- Overview to Hypothesis Testing
- Chi-Square Test
- Test for continuous Data
- Non-Normal Data
- Correlation and Regression

Hypothetical Testing

Agenda

- Introduction
- Objectives
- Classification of Statistical Tests
- Parametric Tests
- Non-parametric Tests

Discussion

- Earth is the center of universe
- Earth is flat
- Continents do not move
- Stress causes Ulcers
- 10000 hours of appropriately guided practice is “the number of greatness” (Malcolm Gladwell)
- Men are better drivers than women

What are these statements? How were they proved or disapproved?

Hypothesis: Business Examples

- No difference in performance of the sales team across geographies and product lines
- Change in gas price will have no impact on losses in automotive finance
- Change in CEO will have no impact on the stock price
- Real estate yields are the same in all metros
- Compensation changes will not impact attrition

Hypothesis Testing

- Is a method an inference about a population parameter based on sample data
- Is statistical analysis used to determine if the difference observed in samples is not a random occurrence but a true difference

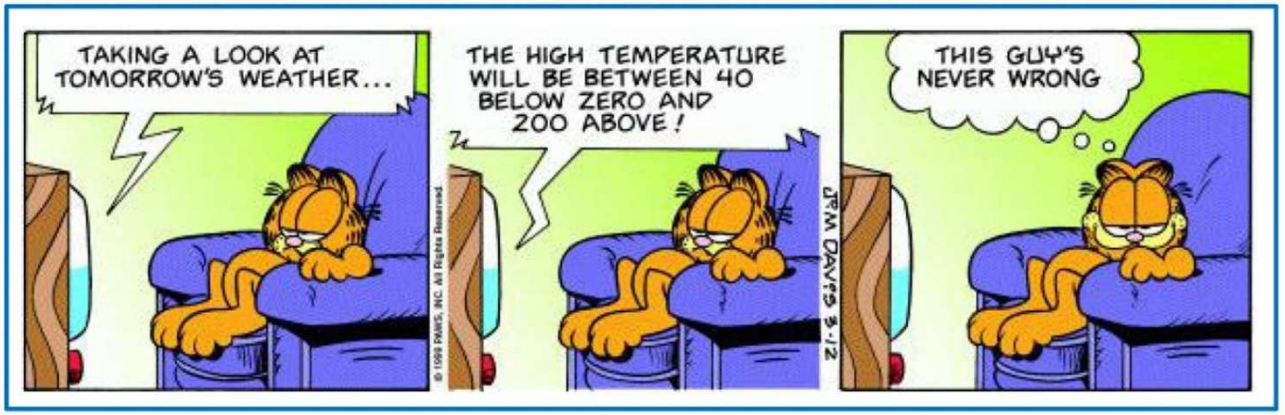
Key Terms

- Three key terms that you need to understand in Hypothesis Testing are:

Confidence Interval	Measure for reliability of an estimate; sample is used for estimating a population parameter so we need to know the reliability of that estimate
Degrees of Freedom	Number of values that are free to vary in a study
P-value	Probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the Null Hypothesis is true

Confidence Interval

- Confidence Interval
  - Describe the reliability of an estimate
  - Range of values (lower and upper boundary) within which the population parameter is included
  - Width of the interval indicates the uncertainty associated with the estimate
- Confidence Level
  - Probability associated with the confidence interval



100% Confidence Interval

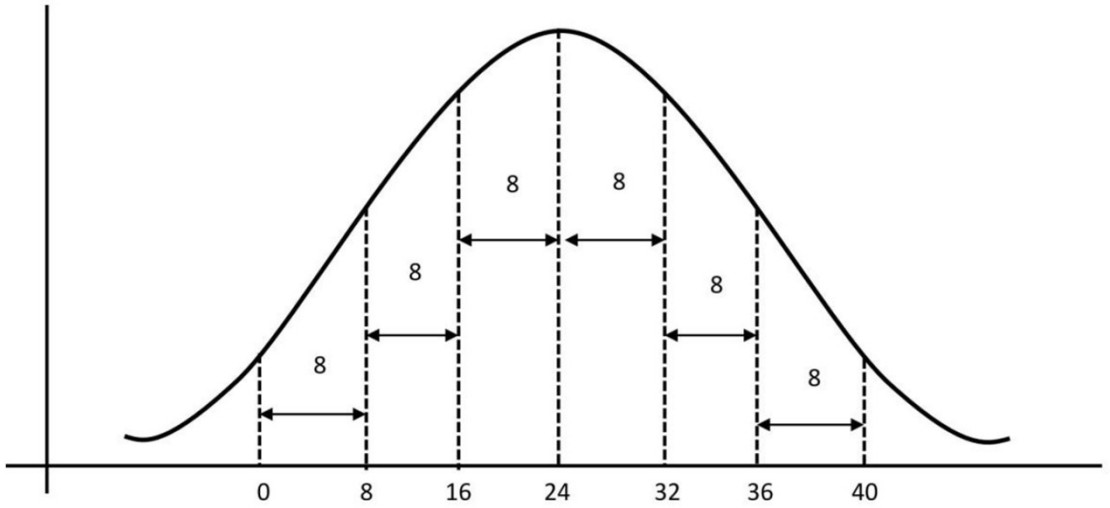
Discussion

“Average Miles Per Gallon (MPG) for automobiles running in metropolitan cities is 24 MPG with a Standard Deviation of 6 MPG”

What does it mean?

Discussion (Cont'd)

- 68.2% cars have MPG between 16-32
- 99% cars have MPG between 6-42



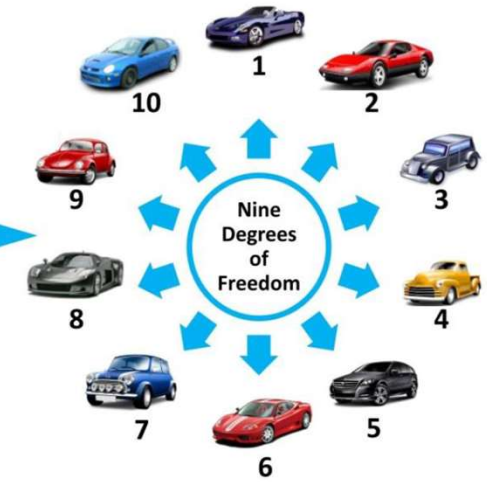
Change the scale

Degrees of Freedom

Degrees of Freedom is the measures of number of values in a study that are free to vary.

Example: You go to a car rental a car. You observe that there are 10 cars in the lot to choose from. While you are still thinking about your choice, other customers start renting out cars. As the cars are driven out of the lot by other customers, your choice decreases.

Initially, there are 10 cars to rent out. When other customers start to rent out and take them out of the lot, your choice decreases. If 9 cars are rented out, there is no choice – only one car is left for you to rent.

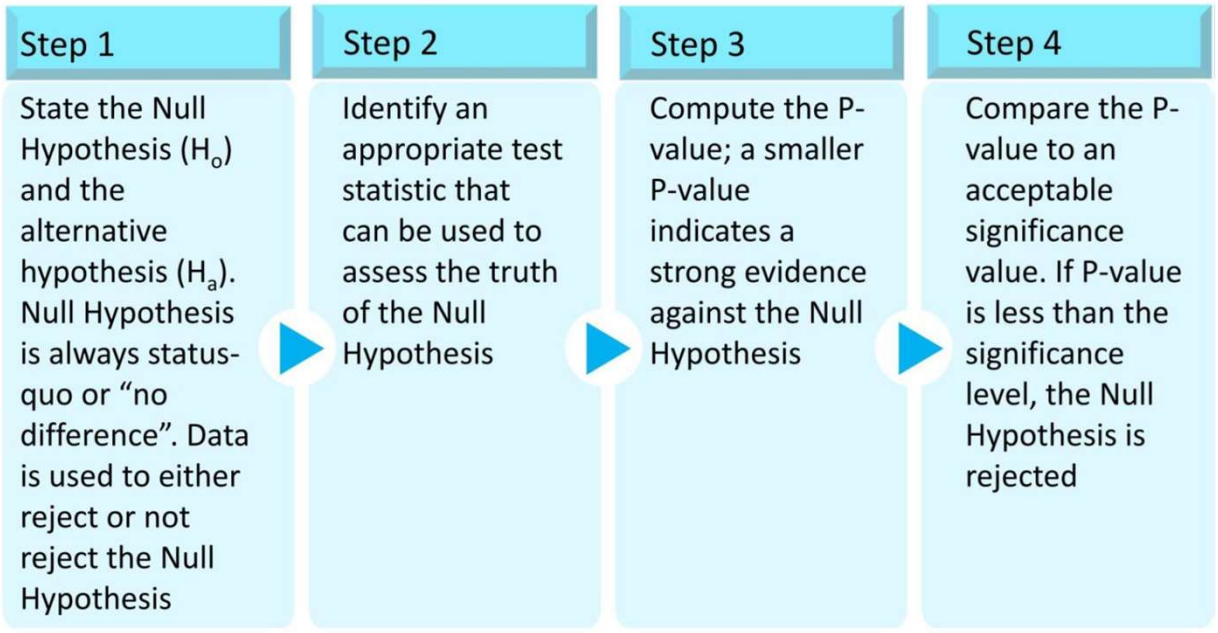


P-value

- P-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the Null Hypothesis Is True.
- When P-value is less than a certain significance level (often 0.05), you “reject the null hypothesis”. This result indicates that the observed result is not due to a random occurrence but a true difference.

Process of Hypothesis Testing

The process of Hypothesis Testing consist of four steps:

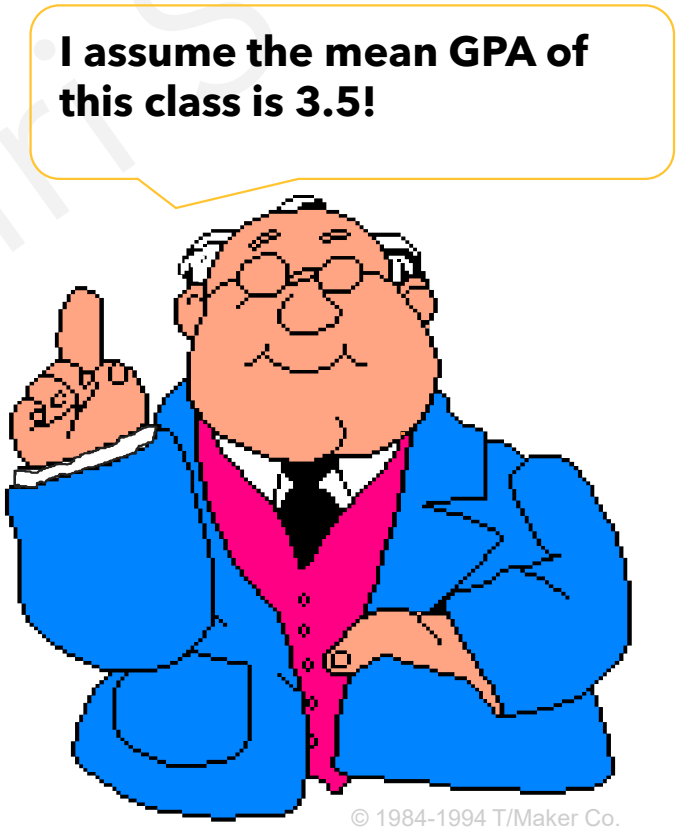


Agenda

- Hypothesis Testing Methodology
- Z Test for the Mean (σ Known)
- p-Value Approach to Hypothesis Testing
- Connection to Confidence Interval Estimation
- One Tail Test
- t Test of Hypothesis for the Mean
- Z Test of Hypothesis for the Proportion

What is a Hypothesis?

- A hypothesis is an assumption about the population parameter.
  - A parameter is a Population mean or proportion
  - The parameter must be identified before analysis.



The Null Hypothesis, H_0

- States the Assumption (numerical) to be tested
- e.g. The average # TV sets in US homes is at least 3 (H_0: $\mu \geq 3$)
- Begin with the assumption that the null hypothesis is TRUE.

(Similar to the notion of innocent until proven guilty)

- Refers to the Status Quo
- Always contains the ' = ' sign

- The Null Hypothesis may or may not be rejected.

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis

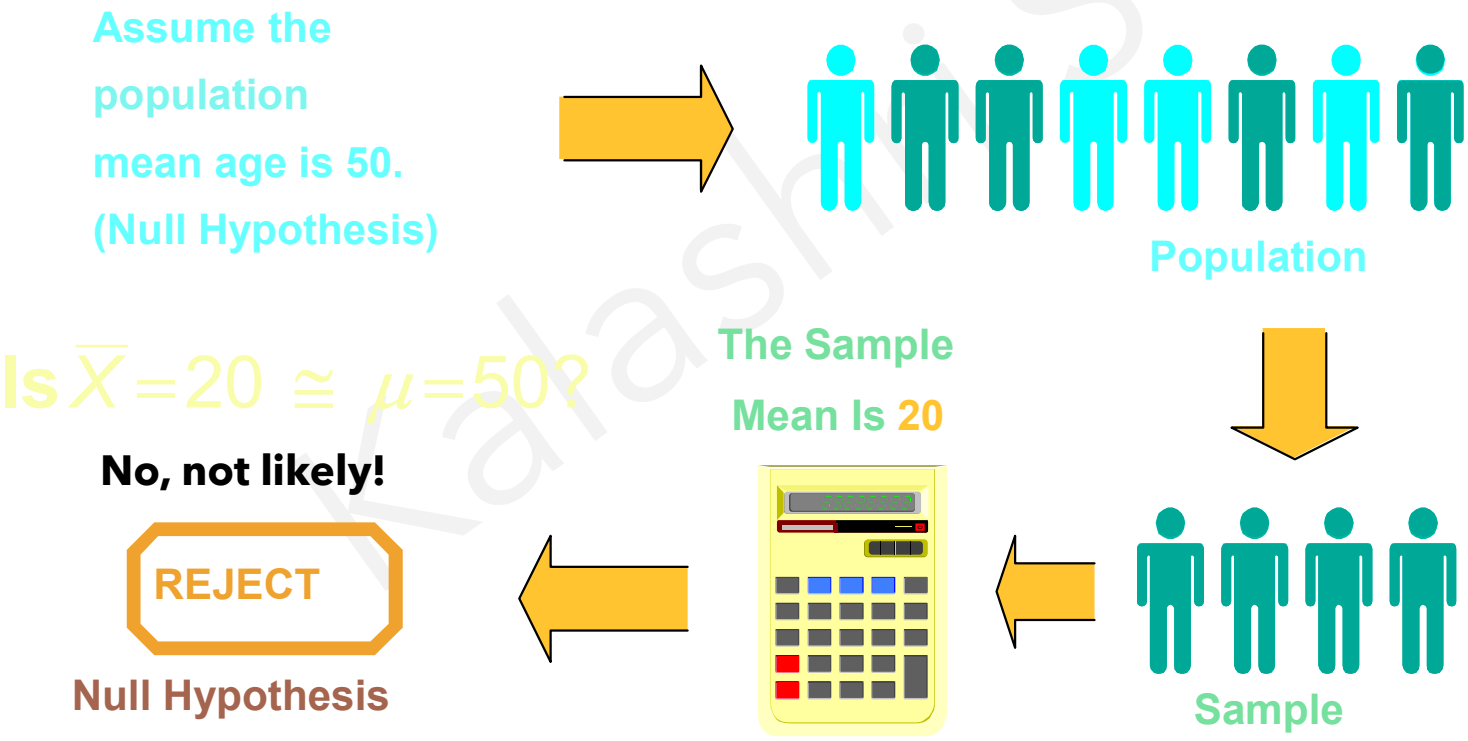
  e.g. The average # TV sets in US homes is less than 3 (H_1: $\mu < 3$)

- Challenges the Status Quo
- Never contains the '=' sign
- The Alternative Hypothesis may or may not be accepted

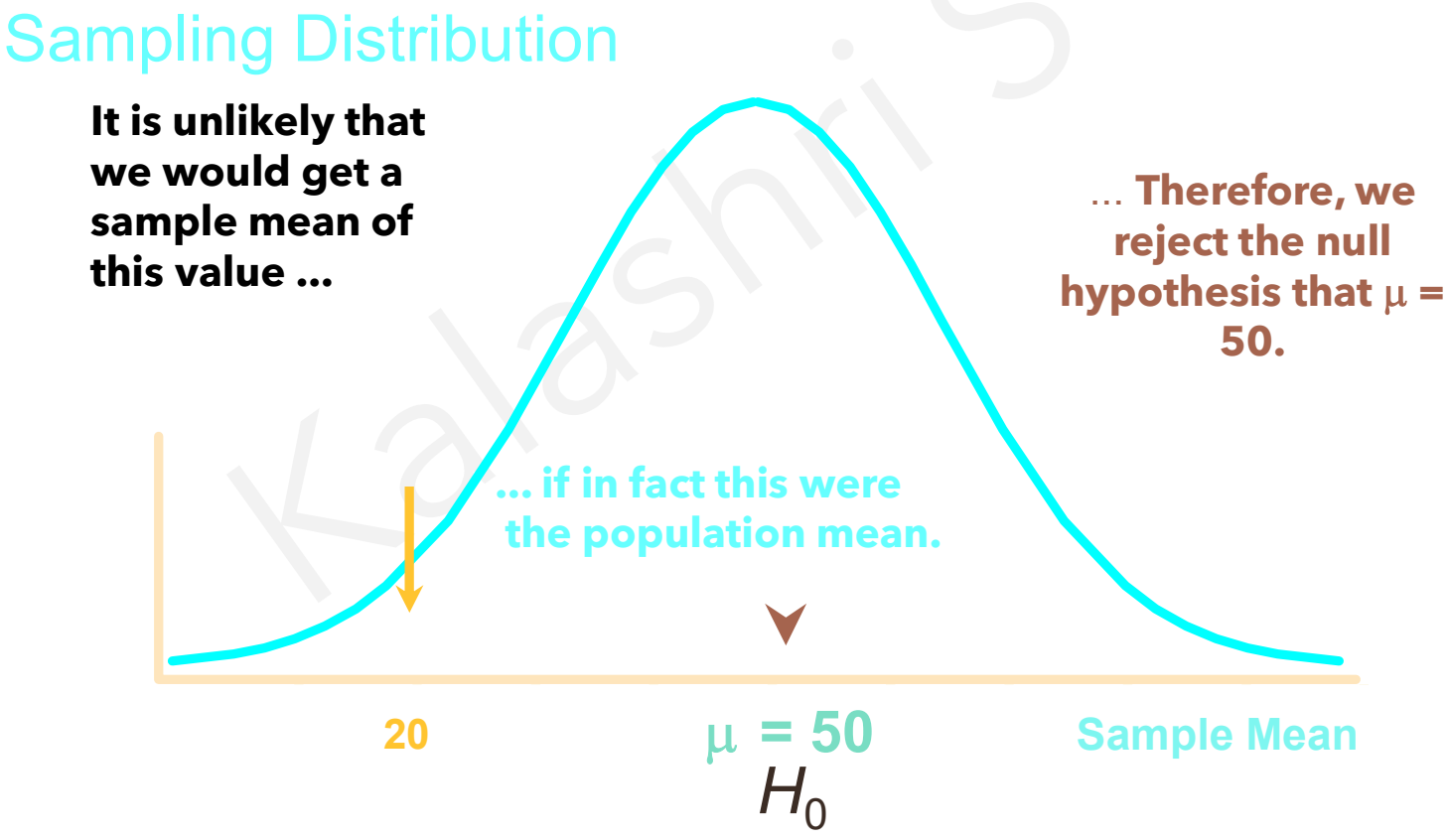
Identify the Problem

- Steps:
  - State the Null Hypothesis (H_0: $\mu \geq 3$)
  - State its opposite, the Alternative Hypothesis (H_1: $\mu < 3$)
    - Hypotheses are **mutually exclusive & exhaustive**
    - Sometimes it is easier to form the alternative hypothesis first.

Hypothesis Testing Process



Reason for Rejecting H_0



Level of Significance, α

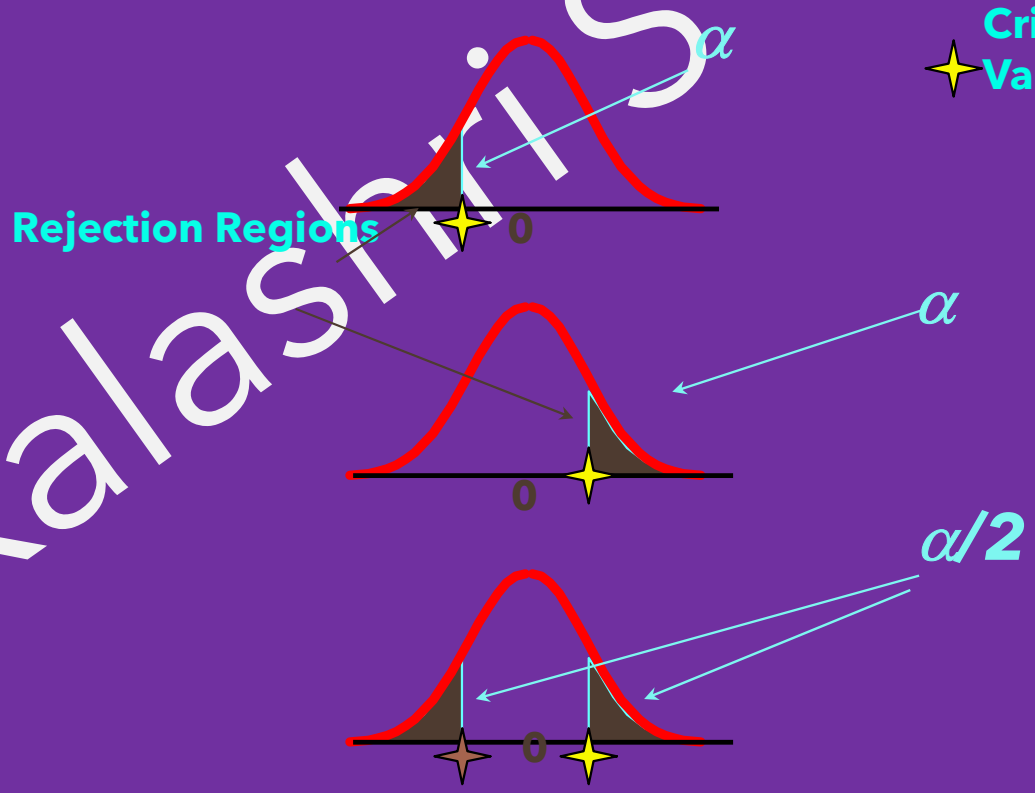
- Defines Unlikely Values of Sample Statistic if Null Hypothesis Is True
  - Called Rejection Region of Sampling Distribution
- Designated α (alpha)
  - Typical values are 0.01, 0.05, 0.10
- Selected by the Researcher at the Start
- Provides the Critical Value(s) of the Test

Level of Significance, α and the Rejection Region

$H_0: \mu \geq 3$
$H_1: \mu < 3$

$H_0: \mu \leq 3$
$H_1: \mu > 3$

$H_0: \mu = 3$
$H_1: \mu \neq 3$



Errors in Making Decisions

- Type I Error
  - Reject True Null Hypothesis
  - Has Serious Consequences
  - Probability of Type I Error Is α
    - Called Level of Significance
- Type II Error
  - Do Not Reject False Null Hypothesis
  - Probability of Type II Error Is β (Beta)

Result Possibilities

H_0: Innocent

	Jury Trial: Actual Situation		Hypothesis Test: Actual Situation		
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Do Not Reject H_0	1 - α	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power (1 - β)

Result Possibilities

H_0: Innocent

Jury Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Do Not Reject H_0	1 - α	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power (1 - β)

α & β Have an Inverse Relationship

Z-Test Statistics (σ Known)

- A **Z-test** is a type of hypothesis test - a way to figure out if results from a test are valid or repeatable.
- **Convert Sample Statistic (e.g., $\bar{X}$) to Standardized Z Variable**

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ *Test Statistic*

- **Compare to Critical Z Value(s)**
  - **If Z test Statistic falls in Critical Region, Reject H_0; Otherwise Do Not Reject H_0**

When you can run a Z Test

Several different types of tests are used in statistics (i.e. f test, chi square test, t test).

You would use a Z test if:

- **Sample size is greater than 30**. Otherwise, use a t test.
- **Data points should be independent** from each other. In other words, one data point isn't related or doesn't affect another data point.
- **Data should be normally distributed**. However, for large sample sizes (over 30) this doesn't always matter.
- Data should be **randomly selected** from a population, where each item has an equal chance of being selected.
- **Sample sizes should be equal** if at all possible.

p Value Test

- Probability of Obtaining a Test Statistic More Extreme ($\leq$ or $\geq$) than Actual Sample Value Given H_0 Is True
- Called Observed Level of Significance
  - Smallest Value of a H_0 Can Be Rejected
- Used to Make Rejection Decision
  - If *p* value $\geq \alpha$, Do Not Reject H_0
  - If *p* value $< \alpha$, Reject H_0

Hypothesis Testing: Steps

Test the Assumption that the true mean # of TV sets in US homes is at least 3.

1.	State H_0	$H_0: \mu \geq 3$
2.	State H_1	$H_1: \mu < 3$
3.	Choose α	$\alpha = .05$
4.	Choose n	$n = 100$
5.	Choose Test:	*Z Test (or p Value)*

Hypothesis Testing: Steps
(continued)

Test the Assumption that the average # of TV sets in US homes is at least 3.

6.	Set Up Critical Value(s)	*Z = -1.645*
7.	Collect Data	*100 households surveyed*
8.	Compute Test Statistic	*Computed Test Stat.= -2*
9.	Make Statistical Decision	*Reject Null Hypothesis*
10.	Express Decision	*The true mean # of TV set is less than 3 in the US households.*

Test Statistics

The **test statistic** is a number calculated from a **statistical test** of a hypothesis.

It shows how closely your observed data match the distribution expected under the null hypothesis of that **statistical test**.

Common test statistics

Z-test :

One-sample &
Two-sample,
One-proportion z-test,
Two-proportion z-test

t-test:

One-sample,
Paired *t*-test,
Two-sample pooled *t*-test,
equal variances Two-sample unpooled *t*-test,
unequal variances (Welch's *t*-test)

Chi-squared test:

for variance,
for goodness of fit

One - sample z-test

- Assumptions:
  - Normal Population or n large
  - σ known

$$z = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$$

- ***Note:***
  - *z* is the distance from the mean in relation to the standard deviation of the mean).
  - For non-normal distributions it is possible to calculate a minimum proportion of a population that falls within *k* standard deviations for any *k*

Two - sample z-test

- Assumption:
  - Normal population **and** independent observations **and**
  - σ_1 and σ_2 are known

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where
equation and equation are the means of the two samples,
d0 is the hypothesized difference between the population means (0 if testing for equal means),
σ 1 and σ 2 are the standard deviations of the two populations
n1 and n2 are the sizes of the two samples.

Introduction to Statistics

Upper-tailed, Lower-tailed, Two-tailed Tests

Step 1. Set up hypotheses and select the level of significance α.

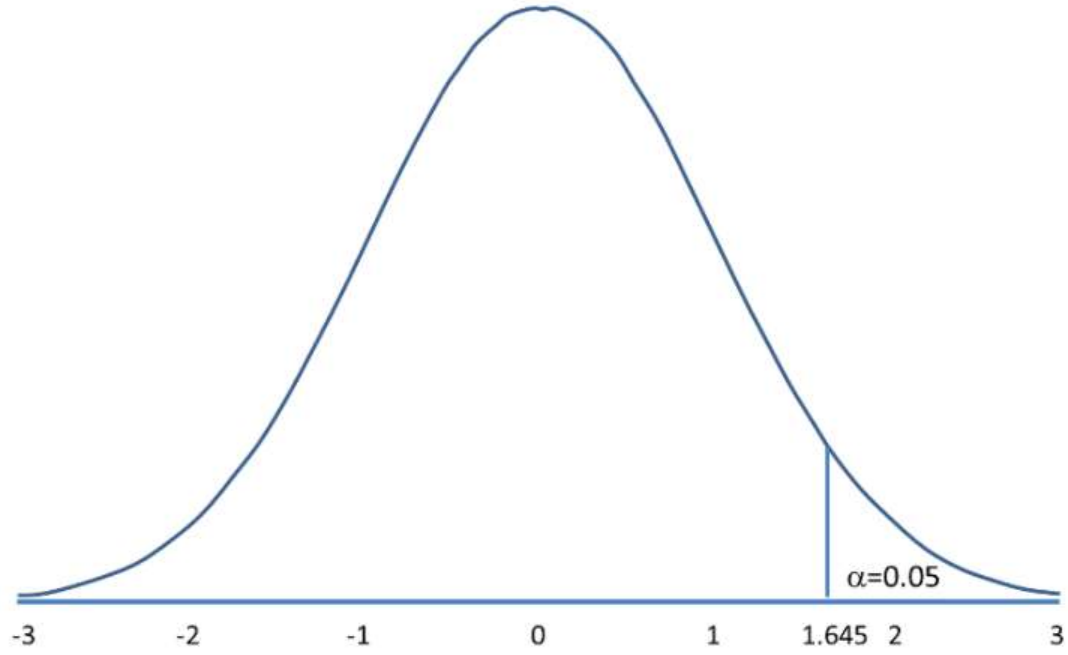
- H_0: Null hypothesis (no change, no difference);
- H_1: Research hypothesis (investigator's belief); α =0.05

The research or alternative hypothesis can take one of three forms. An investigator might believe that the parameter has increased, decreased or changed. For example, an investigator might hypothesize:

1. H_1: $\mu > \mu_0$, where μ_0 is the comparator or null value (e.g., μ_0 =191 in our example about weight in men in 2006) and an increase is hypothesized - this type of test is called an **upper-tailed test**;
2. H_1: $\mu < \mu_0$, where a decrease is hypothesized and this is called a **lower-tailed test**; or
3. H_1: $\mu \neq \mu_0$, where a difference is hypothesized and this is called a **two-tailed test**.

- The exact form of the research hypothesis depends on the investigator's belief about the parameter of interest and whether it has possibly increased, decreased or is different from the null value. The research hypothesis is set up by the investigator before any data are collected.

Upper Tailed Z Test

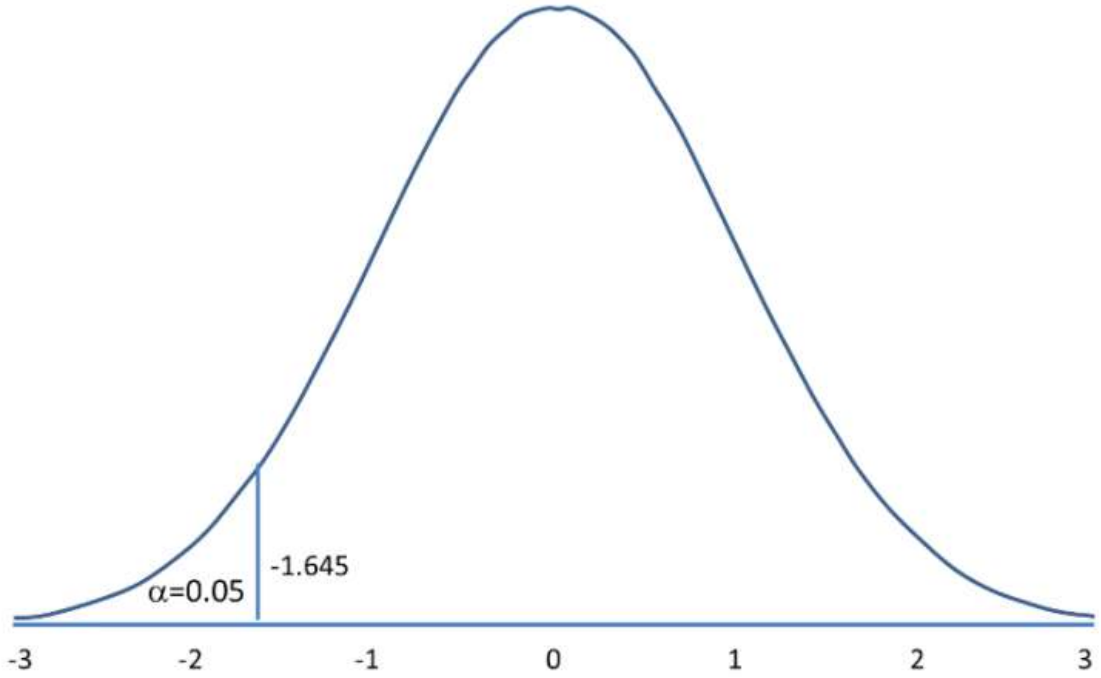


Rejection Region for Upper-Tailed Z Test (H_1: $\mu > \mu_0$) with α=0.05

The decision rule is: Reject H_0 if $Z \geq 1.645$.

Upper-Tailed Test	
α	Z
0.10	1.282
0.05	1.645
0.025	1.960
0.010	2.326
0.005	2.576
0.001	3.090
0.0001	3.719

Lower Tailed Z Test

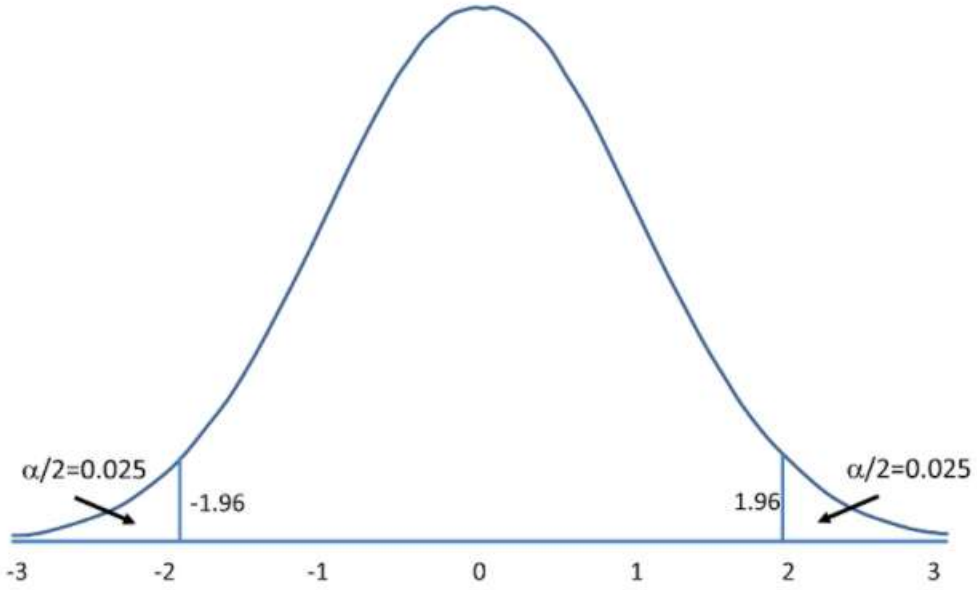


Rejection Region for Lower-Tailed Z Test (H_1: $\mu < \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \leq 1.645$.

Lower-Tailed Test	
a	Z
0.10	-1.282
0.05	-1.645
0.025	-1.960
0.010	-2.326
0.005	-2.576
0.001	-3.090
0.0001	-3.719

Two Tailed Z Test



Rejection Region for Two-Tailed Z Test (H_1: $\mu \neq \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Two-Tailed Test	
α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

One – proportion Z-test

The test statistic is a z-score (z) defined by the following equation. $z=(p-P) / \sigma$ where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and σ is the standard deviation of the sampling distribution

One – proportion Z-test

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

Where −
z = Test statistics
n = Sample size
p_o = Null hypothesized value
p^ = Observed proportion

Example

Problem Statement:

A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is obtained. Of these 100 doctors, 82 indicate that they recommend aspirin. Is this claim accurate? Use alpha = 0.05.

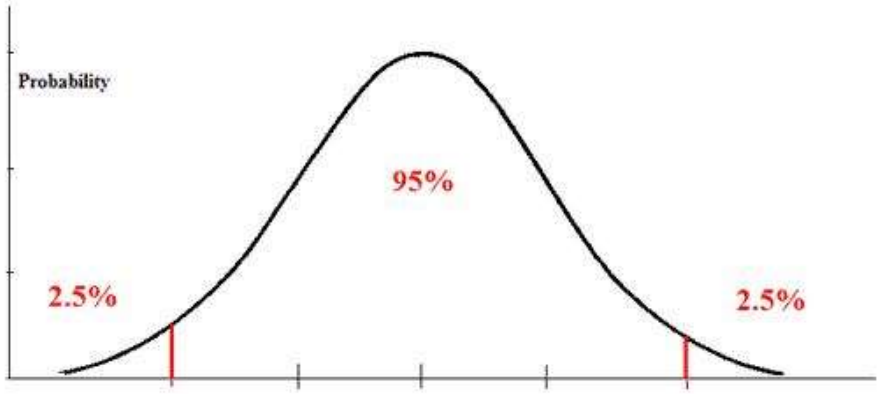
Solution:

Define Null and Alternative Hypotheses

H_0;p=0.90

H_0;p≠0.90

Here Alpha = 0.05. Using an alpha of 0.05 with a two-tailed test, we would expect our distribution to look something like this:



Here we have 0.025 in each tail. Looking up 1 - 0.025 in our z-table, we find a critical value of 1.96. Thus, our decision rule for this two-tailed test is: If Z is less than -1.96, or greater than 1.96, reject the null hypothesis.

Calculate Test Statistic:

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

$$\hat{p} = .82$$

$$p_o = .90$$

$$n = 100$$

$$z_o = \frac{.82 - .90}{\sqrt{\frac{.90(1-.90)}{100}}}$$

$$= \frac{-.08}{0.03}$$

$$= -2.667$$

As z = -2.667
Thus as result we should reject the null hypothesis and as conclusion, The claim that 9 out of 10 doctors recommend aspirin for their patients is not accurate,
z = -2.667, p < 0.05.

Two proportion Z test

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- $n_1\, p_1 > 5$ **and** $n_1(1 - p_1) > 5$ **and**
- $n_2\, p_2 > 5$ **and** $n_2(1 - p_2) > 5$ **and** independent observations

Example

Let's say you're testing two flu drugs A and B. Drug A works on 41 people out of a sample of 195. Drug B works on 351 people in a sample of 605. Are the two drugs comparable? Use a 5% alpha level.

Step 1: Find the **two proportions**:

P_1 = 41/195 = 0.21 (that's 21%)

P_2 = 351/605 = 0.58 (that's 58%).

Set these numbers aside for a moment.

Step 2: Find the **overall sample proportion**. The numerator will be the total number of "positive" results for the two samples and the denominator is the total number of people in the two samples.

p = (41 + 351) / (195 + 605) = 0.49.

Set this number aside for a moment.

Step 3: Insert the numbers from Step 1 and Step 2 into the test statistic formula:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{(.58 - .21) - 0}{\sqrt{.49\ (1 - .49)\left(\frac{1}{195} + \frac{1}{605}\right)}}$$

Solving the formula, we get:
Z = 8.99
We need to find out if the z-score falls into the "rejection region."

Step 4: Find the z-score associated with α/2. I'll use the following table of known values:

Confidence Level	Alpha	Alpha/2	z alpha/2
90%	10%	5.0%	1.645
95%	5%	2.5%	1.96
98%	2%	1.0%	2.326
99%	1%	0.5%	2.576

The z-score associated with a 5% alpha level / 2 is 1.96.

Step 5: Compare the calculated z-score from Step 3 with the table z-score from Step 4. If the calculated z-score is larger, you can reject the null hypothesis.

8.99 > 1.96, so we can reject the null hypothesis.

One – sample t test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

When can I use the test?

You can use the test for continuous data. Your data should be a random sample from a normal population.

What do we need?

One-sample *t*-test assumptions:

For a valid test, we need data values that are:

- Independent (values are not related to one another).
- Continuous.
- Obtained via a simple random sample from the population.
- Also, the population is assumed to be normally distributed.
- Example:
  - A hospital has a random sample of cholesterol measurements for men. These patients were seen for issues other than cholesterol. They were not taking any medications for high cholesterol. The hospital wants to know if the unknown mean cholesterol for patients is different from a goal level of 200 mg.
  - We measure the grams of protein for a sample of energy bars. The label claims that the bars have 20 grams of protein. We want to know if the labels are correct or not.

Example

Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.

Step 1: Write your null hypothesis. The accepted hypothesis is that there is no difference in sales, so: H0: μ = $100.

Step 2: Write your alternate hypothesis. This is the one you're testing. You think that there is a difference (that the mean sales increased), so: H1: μ > $100.

Step 3: The sample mean($\bar{x}$) = $130.

The population mean(μ) = $100 (from past data).

The sample standard deviation(s) = $15.

Number of observations(n) = 25.

Step 4: Insert the items from above into the t score formula. $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

t = (130 – 100) / ((15 / √(25))
t = (30 / 3) = 10

Degree of Freedom

- Degrees of Freedom - maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.

Df=$n-1$

where:

- Df=degrees of freedom
- n=sample size

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Step 5: Find the t-table value. You need two values to find this:

1. The alpha level: given as 5% in the question.
2. The degrees of freedom, which is the number of items in the sample (n) minus 1: 25 – 1 = 24.

- Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1.711.This is your one-tailed critical t-value.
- What this critical value means is that we would expect most values to fall under 1.711. If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.

Step 5: Compare Step 4 to Step 5. The value from Step 4 does not fall into the range calculated in Step 5, so we can reject the null hypothesis. The value of 10 falls into the rejection region (the left tail).

Introduction to Statistics

Paired t-test

Also called a

- **correlated pairs t-test**,
- a **paired samples t test** or
- **dependent samples t test**
- is where you run a t test on dependent samples.

• Dependent samples are essentially connected – they are tests on the same person or thing.

For example:

- Knee MRI costs at two different hospitals,
- Two tests on the same person before and after training,
- Two blood pressure measurements on the same person using different equipment.

When to Choose a Paired T Test

- Choose the paired t-test if we have two measurements on the same item, person or thing.
- We should also choose this test if we have two items that are being measured with a unique condition.
- For example, you might be measuring car safety performance in vehicle research and testing and subject the cars to a series of crash tests.
- Although the manufacturers are different, you might be subjecting them to the same conditions.

- With a “regular” two sample t test, we’re comparing the means for two different samples.
- For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills.
- If we take a random sample each group separately and they have different conditions, our samples are independent and we should run an independent samples t test (also called between-samples and unpaired-samples).

- The null hypothesis for the independent samples t-test is

  $\mu_1 = \mu_2$.

- In other words, it assumes the means are equal.
- With the paired t test, the null hypothesis is that the *pairwise difference* between the two tests is equal (H_0: $\mu_d = 0$).
- The difference between the two tests is very subtle; which one we choose is based on our data collection method.

Example

Calculate a paired t test by hand for the following data

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

Step 1: Subtract each Y score from each X score.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

Step 2: Add up all of the values from Step 1. Set this number aside for a moment.

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2
		SUM:	**-73**

Step 3: Square the differences from Step 1.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	

Step 4: Add up all of the squared differences from Step 3.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Step 5: Use the following formula to calculate the t-score:

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

- ΣD: Sum of the differences (Sum of X-Y from Step 2)
- ΣD^2: Sum of the squared differences (from Step 4)
- $(\Sigma D)^2$: Sum of the differences (from Step 2), squared.

$$t = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}} \qquad t = -2.74$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this example problem, with df = 10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to our calculated t-value (-2.74).
The calculated t-value is greater than the table value at an alpha level of .05.
The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.
Note: We can ignore the minus sign when comparing the two t-values, as ± indicates the direction; the p-value remains the same for both directions.

Introduction to Statistics

Two Sample t-test

This function gives an unpaired two sample t test with a confidence interval for the difference between the means.

The unpaired t method tests the null hypothesis that the population means related to two independent, random samples from an approximately normal distribution are equal.

Assuming equal variances, the test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Two Sample t-test

- where x bar 1 and x bar 2 are the sample means, s^2 is the pooled sample variance, n1 and n2 are the sample sizes and t is a Student t quantile with n1 + n2 - 2 degrees of freedom.

Power is calculated as the power achieved with the given sample sizes and variances for detecting the observed difference between means with a two-sided type I error probability of (100-CI%)%.

The unpaired t test should not be used if there is a significant difference between the variances of the two samples.

- For the situation of unequal variances, we can calculate Satterthwaite's approximate t test; a method in the Behrens-Welch family.
- Assuming unequal variances, the test statistic is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}{n_1 - 1}$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2} (x_j - \bar{x}_2)^2}{n_2 - 1}$$

- where x bar 1 and x bar 2 are the sample means,
- s^2 is the sample variance,
- n1 and n2 are the sample sizes,
- d is the Behrens-Welch test statistic evaluated as a t quantile with df freedom using Satterthwaite's approximation.

- Note that is often more robust to use the nonparametric Mann-Whitney test as an alternative method in the presence of unequal variances.

F Statistic

- In analysis of variance, the aim is to test the null hypothesis that the means of two or more population are equal. In other words, our null and the alternate hypothesis are:

  H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$

  H_1: at least one mean is different

- The above hypothesis is tested by the F statistic with (c-1) and (N-c) degrees of freedom in the numerator and denominator respectively. The F statistic is given by the following formula:

$$F = \frac{\frac{SS_x}{(c-1)}}{\frac{SS_{Error}}{(N-c)}}$$

- The rule Is when the calculated value of F is greater than critical value reject the null hypothesis.

Example

Consider the gain in weight of 19 female rats between 28 and 84 days after birth. 12 were fed on a high protein diet and 7 on a low protein diet.

High protein	Low protein
134	70
146	118
104	101
119	85
124	107
161	132
107	94
83	
113	
129	
97	
123	

- Unpaired t test
- Mean of High Protein = 120 (n = 12)
- Mean of Low Protein = 101 (n = 7)
- Assuming equal variances
- Combined standard error = 10.045276
- df = 17
- t = 1.891436
- One sided P = 0.0379
- Two sided P = 0.0757
- 95% confidence interval for difference between means = -2.193679 to 40.193679
- Power (for 5% significance) = 82.25%
- Note: Two sided F test is not significant

- Assuming unequal variances
- Combined standard error = 9.943999
- df = 13.081702
- t(d) = 1.9107
- One sided P = 0.0391
- Two sided P = 0.0782
-
- 95% confidence interval for difference between means = -1.980004 to 39.980004
-
- Power (for 5% significance) = 40.39%

- Comparison of variances
- Two sided F test is not significant
- No need to assume unequal variances
-
- Thus we have a difference that is not quite significant at the 5% level. The most important information is, however, conveyed by the confidence interval. The 95% CI includes zero therefore we can not be confident (at the 95% level) that these data show any difference in weight gain. As most of the interval is toward weight gain and as the test result is in the grey "suggestive" 5%-10% zone we have good evidence for repeating this experiment with larger numbers. Bigger samples will probably shrink the range of uncertainty so that the confidence interval contracts to a narrower band that excludes zero.
-
- N.B. We did not consider a one sided P value here because we could not be absolutely certain that the rats would all benefit from a high protein diet in comparison with those on a low protein diet.

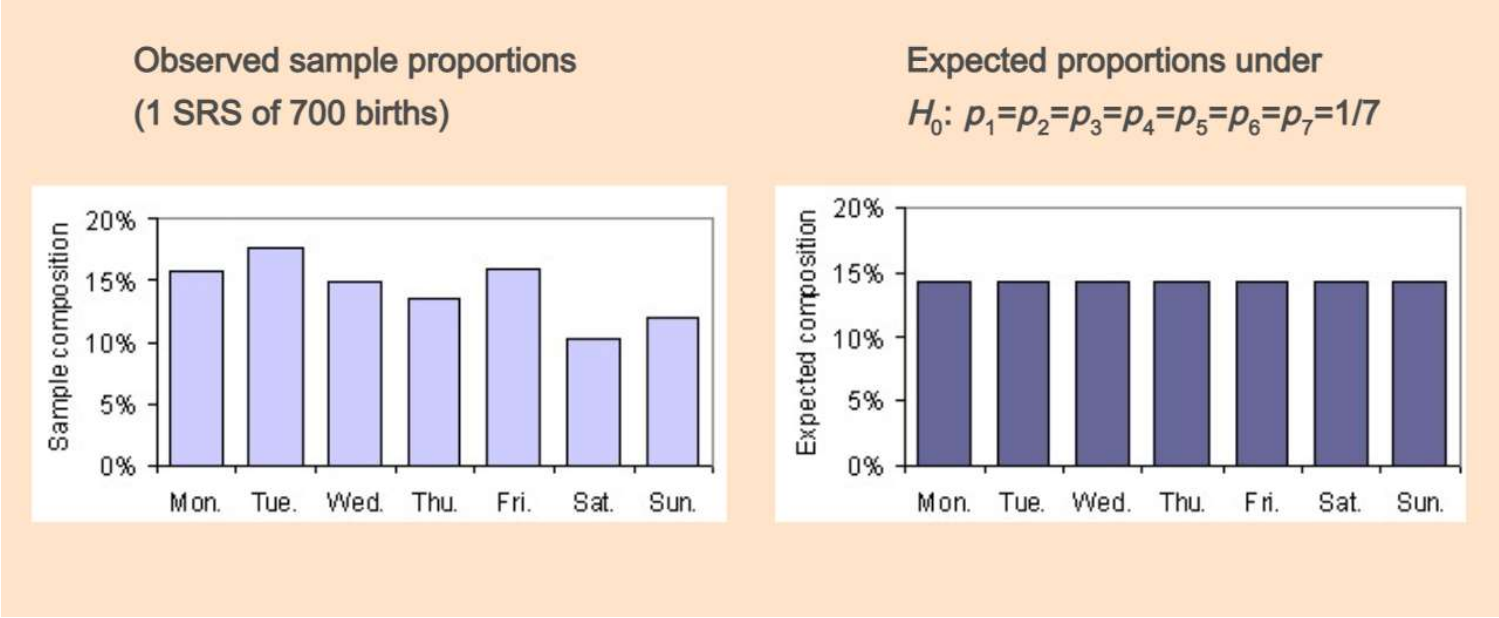
Introduction to Statistics

Chi Squared Test

- Idea of the chi-square test
- The chi-square distributions
- Goodness of fit hypotheses
- Conditions for the chi-square goodness of fit test
- Chi-square test for goodness of fit

Idea of the chi-square test

- The chi-square ($\chi 2$) test is used when the data are categorical. It detects differences between the observed data and when we would expect if H_0 was true.



The chi-square statistic

The chi-square ($\chi 2$) statistic compares observed and expected counts.

- **Observed counts** are the actual number of observations of each type.
- **Expected counts** are the number of observations that we would expect to see of each type if the null hypothesis was true.

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

(calculated for each category separately and then summed)

The chi-square distributions

The χ^2 distributions are a family of distributions that take only positive values, are skewed to the right, and are described by a specific degrees of freedom.

Published tables & software give the upper-tail area for critical values of many χ^2 distributions.

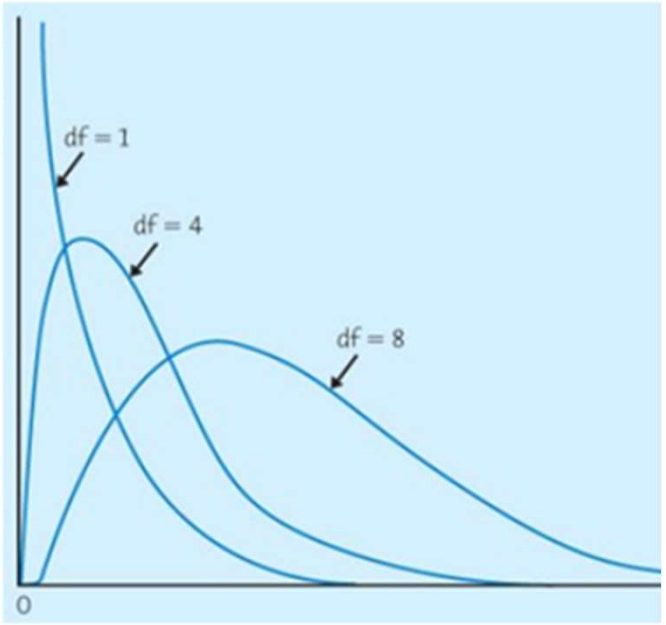


Table A

Press Esc to exit full screen

Ex: df = 6

	p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.70
80	88.13	90.41	93.11	96.58	101.90	106.60	108.10	112.30	116.30	120.10	124.80	128.30
100	109.10	111.70	114.70	118.50	124.30	129.60	131.10	135.80	140.20	144.30	149.40	153.20

Goodness of fit hypotheses

The chi-square test can be used to for a categorical variable (1 SRS) with **any number *k* of levels**.
The null hypothesis can be that all population proportions are equal (uniform hypothesis)

Are hospital births uniformly distributed in the week?

H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7$

or that they are equal to some specific values, as long as the sum of all the population proportions in H0 equals 1.

When crossing homozygote parents expressing two co-dominant phenotypes A and B, we would expect in F2

H_0: $p_A = ¼$, $p_{AB} = ½$, $p_B = ¼$ where AB is an intermediate phenotype.

For 1 SRS of size *n* with *k* levels of a categorical variable

When testing

H_0: $p_1 = p_2 = \ldots = p_k$ (a uniform distribution)

The **expected counts are all = *n* / *k***

When testing

H_0: $p_1 = p_{1Ho}$ and $p_2 = p_{2Ho}$ … and $p_k = p_{kHo}$

The expected counts in each level *i* are

expected count$_i$ = $n\ p_{iHo}$

Conditions for the goodness of fit test

The chi-square test for goodness of fit is used when we have a single SRS from a population, and the data are categorical, with *k* mutually exclusive levels.

The sampling distribution of the X^2 statistic will be approximately chi-square distributed when:

- **all expected counts** are 1 or more (≥1)
- no more than 20% of **expected counts** are less than 5

Recall:Chi-square test for goodness of fit

The **chi-square statistic for goodness of fit with *k* proportions** measures how much observed counts differ from expected counts. It follows the chi-square distribution **with *k* − 1 degrees of freedom** and has the formula:

$$X^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}}$$

The *P*-value is the tail area under the X^2 distribution with df = k – 1.

Example: River ecology

Three species of large fish (A, B, C) that are native to a certain river have been observed to co-exist in equal proportions.

A recent random sample of 300 large fish found 89 of species A, 120 of species B, and 91 of species C. Do the data provide evidence that the river's ecosystem has been upset?

H_0: $p_A = p_B = p_C = 1/3$ H_a: H_0 is not true

Number of proportions compared: $k = 3$

All the expected counts are : $n / k = 300 / 3 = 100$

Degrees of freedom: $(k - 1) = 3 - 1 = 2$

X^2 calculations:

$$\chi^2 = \frac{(89-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(91-100)^2}{100}$$
$$= 1.21 + 4.0 + 0.81 = 6.02$$

If H_0 was true, how likely would it be to find by chance a discrepancy between observed and expected frequencies yielding a X^2 value of 6.02 or greater?

TABLE E Chi-square distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20

From Table E, we find $5.99 < X^2 < 7.38$, so $0.05 > P > 0.025$

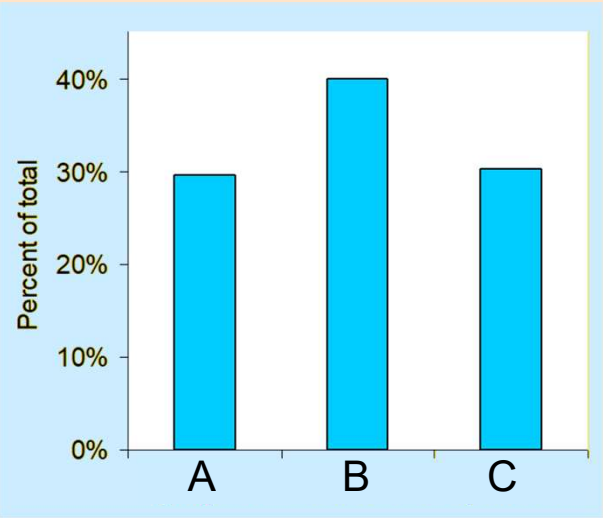
Software gives P-value = 0.049

Using a typical significance level of 5%, we conclude that the results are significant. We have found evidence that the 3 fish populations are not currently equally represented in this ecosystem ($P < 0.05$).

Interpreting the x^2 output

The individual values summed in the X^2 statistic are the X^2 **components.**

- **When the test is statistically significant**, the **largest components** indicate which condition(s) are most different from the expected H_0.
- You can also compare the actual proportions qualitatively in a graph.



$$\chi^2 = \frac{(89-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(91-100)^2}{100}$$
$$= 1.21 + 4.0 + 0.81 = 6.02$$

The largest X^2 component, 4.0, is for species B. The increase in species B contributes the most to significance.

Lack of significance: Avoid a logical fallacy

A **non-significant *P*-value** is not conclusive: H_0 could be true, or not.

This is particularly relevant in the **X^2 goodness of fit test** where we are often **interested in H_0** that the data fit a particular model.

- A significant *P*-value suggests that the data do not follow that model (but by how much?).
- But finding **a non-significant *P*-value is NOT a validation of the null hypothesis** and does NOT suggest that the data do follow the hypothesized model. It only shows that the data are not inconsistent with the model.

Goodness of fit for a genetic model

Under a genetic model of dominant epistasis, a cross of white and yellow summer squash will yield white, yellow, and green squash with probabilities 12/16, 3/16 and 1/16 respectively (expected ratios 12:3:1).

Suppose we observe the following data:

Color	Number of Offspring
white	155
yellow	40
green	10

Are they consistent with the genetic model?

H_0: $p_{white} = 12/16$; $p_{yellow} = 3/16$; $p_{green} = 1/16$

H_a: H_0 is not true

We use H_0 to compute the expected counts for each squash type.

Color	Observed	Expected
white	155	$205 \times 12/16 = 153.75$
yellow	40	$205 \times 3/16 = 38.4375$
green	10	$205 \times 1/16 = 12.8125$
Total	205	205

We then compute the chi-square statistic:

$$\chi^2 = \frac{(155-153.75)^2}{153.75} + \frac{(40-38.4375)^2}{38.4375} + \frac{(10-12.8125)^2}{12.8125} = 0.069106$$

Color	Observed	Expected	χ^2
white	155	153.75	0.01016
yellow	40	38.4375	0.06352
green	10	12.8125	0.61738
Total	205	205	0.69106

Degrees of freedom = $k - 1 = 2$, and $X^2 = 0.691$.

Using Table D we find $P > 0.25$. Software gives $P = 0.708$.

This is not significant and we fail to reject H_0. The observed data are **consistent with** a dominant epistatic genetic model (12:3:1). The small observed deviations from the model could simply have arisen from the random sampling process alone.

Introduction to Statistics

Agenda

- Overview to Hypothesis Testing
- Chi-Square Test
- Test for continuous Data
- Non-Normal Data
- Correlation and Regression

Diet Crackers and Bloating Stomachs

- A diet cracker manufacturer wants to launch a new type of diet cracker, which has high content of a certain kind edible fibre. Before the launch, the manufacturer wants to analyse the result of eating crackers.
- A study was conducted on few overweight subjects who ate crackers with different types of fibre (bran, gum, both, and no fibre). After this, they were allowed to eat as much as they wished from a prepared menu. The amount od food they consumed and their weight was monitored. Additionally, the side effects were also reported.
- Unfortunately, some subjects developed uncomfortable bloating and gastric upset after eating the crackers.
- Is there a relation between eating diet crackers and bloating?

Discussion

- How can the manufacturer establish that the crackers did not have any side effects and the bloating is not due to eating crackers?

Chi-Square Test

- Tests a Null Hypothesis that he frequency distribution of certain events observed in a sample is consistent with a particular distribution. Events considered must be mutually exclusive and have a total probability of 1
- Used for:
  - Goodness of Fit: Observed frequency distribution differs from a theoretical distribution
  - Test of Independence: Paired observations on two variables, expressed in a contingency table are independent of each other.

Goodness of Fit

- When you toss a coin, you have an equal probability of getting a head or a tail. So, if you toss the coin 100 times, the expected distribution is:

Head	Tail
50	50

- Take a scenario where you get this result:

Head	Tail
57	43

- How do you know if the coin is biased or if you toss another 100 times, you will get the expected distribution?

Goodness of Fit test helps to establish if the observed distribution fits the expected distribution.

Test of Independence

- An ice cream vender conducts a survey to capture the relation between ice cream flavour preference and gender

	Vanilla	Strawberry	Chocolate
Men	30	45	35
Women	20	40	50

- Based on the above data, how can the vendor establish the relation between gender and ice cream flavour preference?

Test of Independence helps to establish association or relation between two categorical variables.

Introduction to Statistics

Chi-Square Test Steps

1. Calculate the Chi-squared test statistic: Chi-Square statistic is the normalized sum of squared deviations between observed and theoretical frequencies.

$$X^2 = \sum \frac{(o-e)^2}{e}$$

o – observed or actual frequency
e – expected or theoretical frequency

2. Determine the Degrees of Freedom of that statistic: Number of frequencies reduced by the number of parameters of the fitted distribution
3. 3. Compare the Chi-square to the critical value from the Chi-Square distribution.

Chi-Square Test Example

A university wants to analyse students decision to enroll in part-time courses. It is assumed that students who have children enrolled for the part-time courses. Sarah collected the data generated the contingency table.

			Do you have children?		
			Yes	No	Total
Are you a full-time or part-time student?	Full-time	Count	31	169	200
		Expected Count	40.2	159.8	200
	Part-time	Count	15	14	29
		Expected Count	5.8	23.2	29.o
Total		Count	46	183	229
		Expected Count	46	183	229

The Chi-Square statistic is calculated as:

$$\frac{(31-40.2)^2 + (169-159.8)^2 + (15-5.8)^2 + (14-23.2)^2}{229}$$

$= 1.47$

$DF = (2-1)*(2-1) = 1$

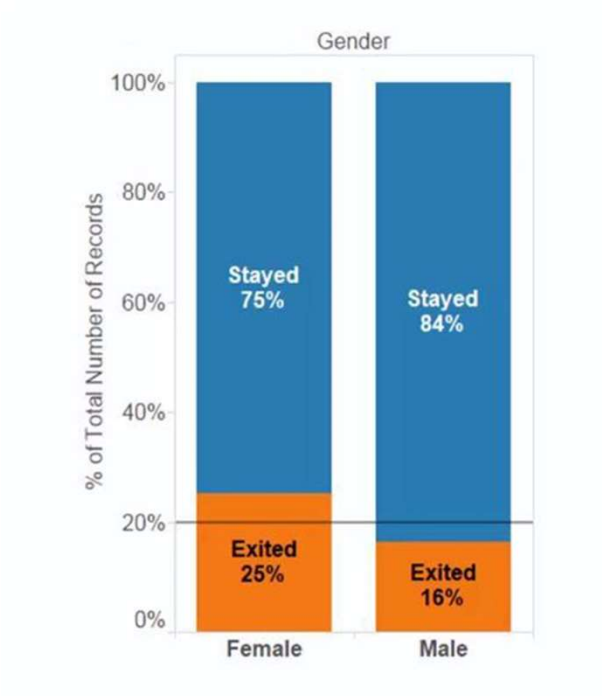
At significance level of 0.05, the critical Chi- Square is 3.841.

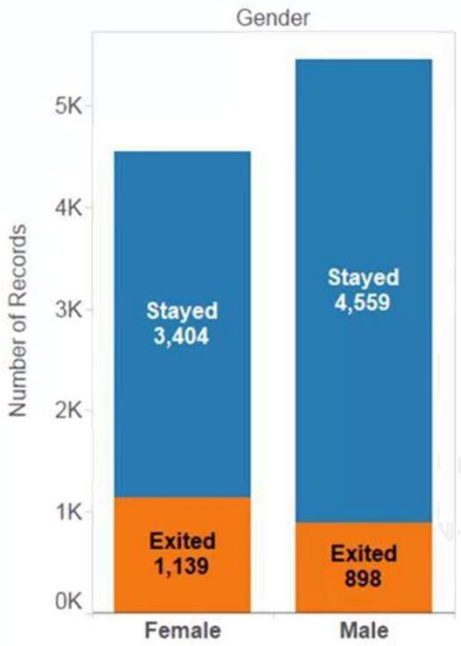
Chi-Square the test is less than the critical Chi-Square and P-value is less than the significance level. Null Hypothesis is rejected .

Sarah can conclude that having children had an association with student enrolment into part-time courses

Chi-Squared

Chi-Squared





Observed:

	Stayed	Exited
Male	4,559	898
Female	3,404	1,139

Expected:

	Stayed	Exited
Male	4,366	1,091
Female	3,634	909

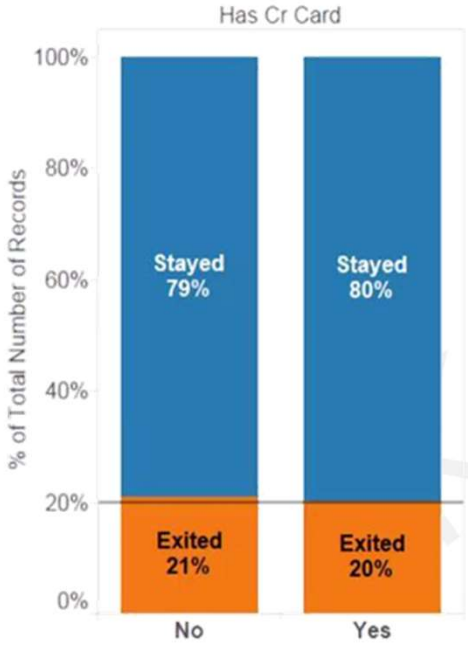
20% x Total Males

20% x Total Females

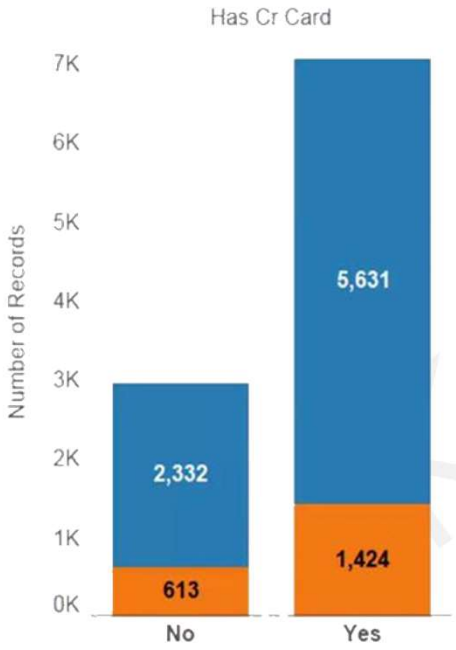
Chi-Squared

Chi-Squared is a test designed to test the **probability of independence**

Chi-Squared



Chi-Squared



Observed:

	Stayed	Exited
Yes	5,631	1,424
No	2,332	613

Expected:

	Stayed	Exited
Yes	5,644	1,411
No	2,356	589

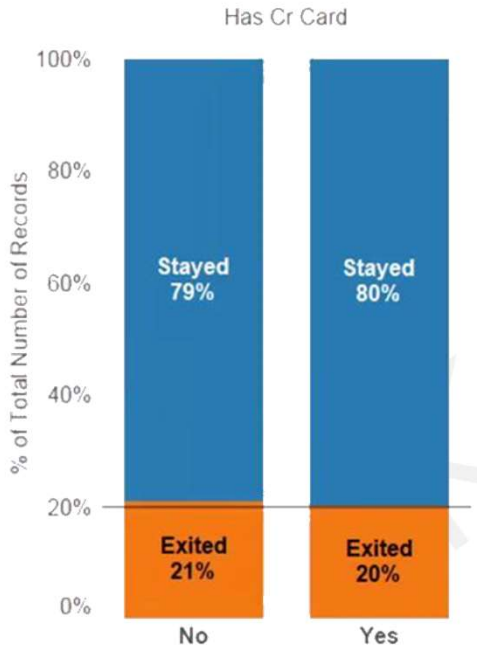
20% x Total Yes

20% x Total No

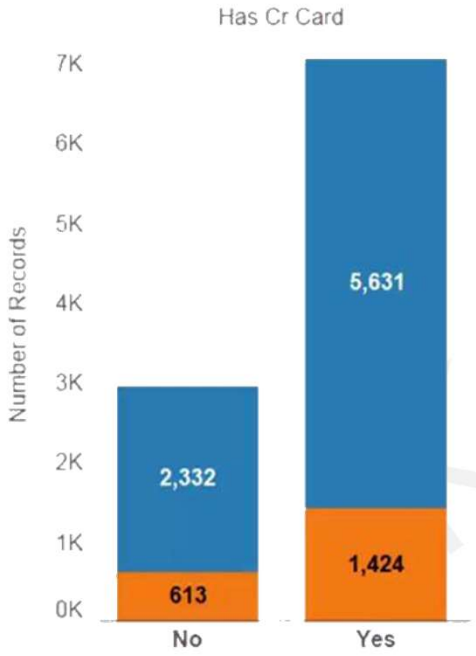
Excel

2	Chi-Squared Test			
3				
4	**Observed**			
5		Stayed	Exited	
6	Yes	5631	1424	7055
7	No	2332	613	2945
8		7963	2037	10000
9				
10	**Expected**			
11		Stayed	Exited	
12	Yes	5617.8965	1437.1	7055
13	No	2345.1035	599.897	2945
14		7963	2037	10000
15				
16	**P-Value**	0.4753654		
17	**Sign Level**	0.05		
18		Independent		

Chi-Squared



Chi-Squared



Observed:

	Stayed	Exited
Yes	5,631	1,424
No	2,332	613

Expected:

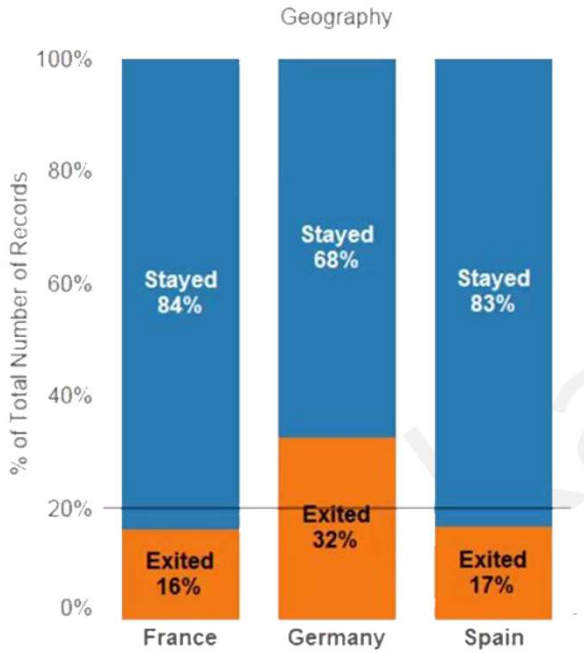
	Stayed	Exited
Yes	5,644	1,411
No	2,356	589

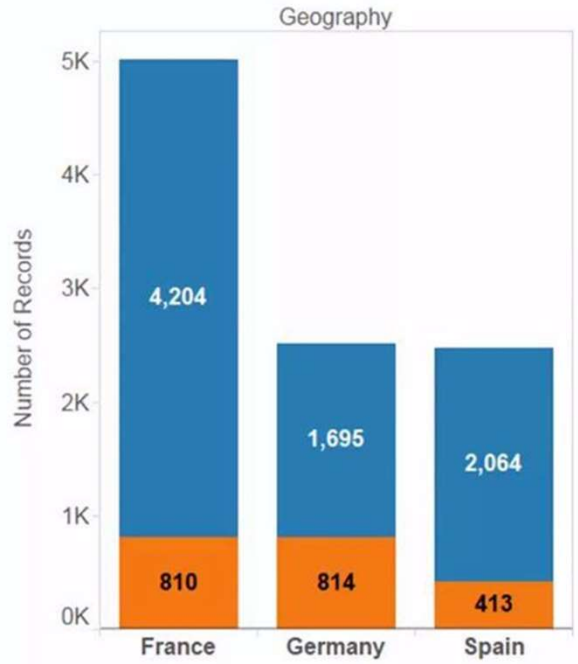
Chi-Squared

Rules

1. Probability of Independence
2. NOT the relationship between variables
3. Cannot use %, need absolute values
4. Categories must be Mutually Exclusive
5. Never exclusive one of the outcomes
6. Minimum 5 observations in each cell

Introduction to Statistics





Observed:

	Stayed	Exited
France	4,204	810
Germany	1,695	814
Spain	2,064	413

Expected:

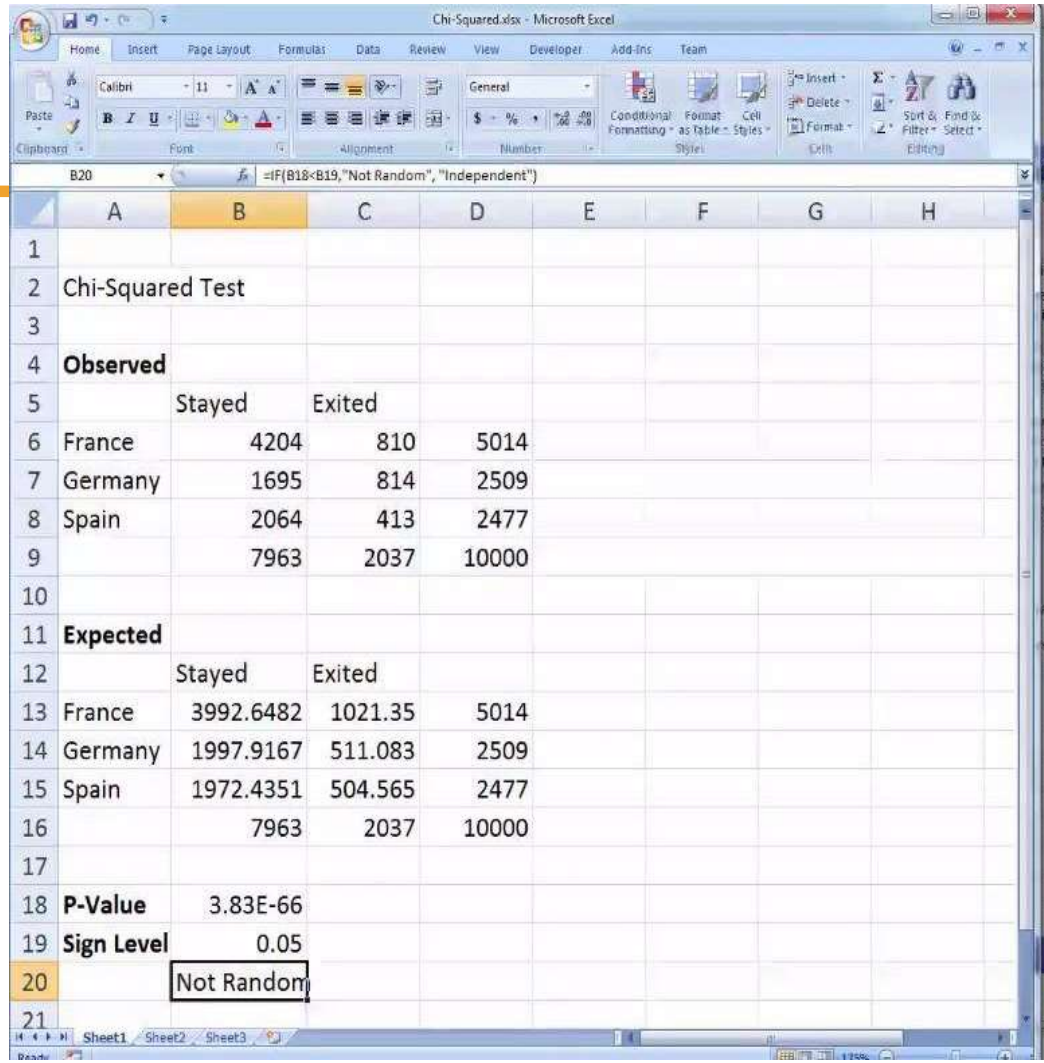
	Stayed	Exited
France	4,011	1,003
Germany	2,007	502
Spain	1,982	495

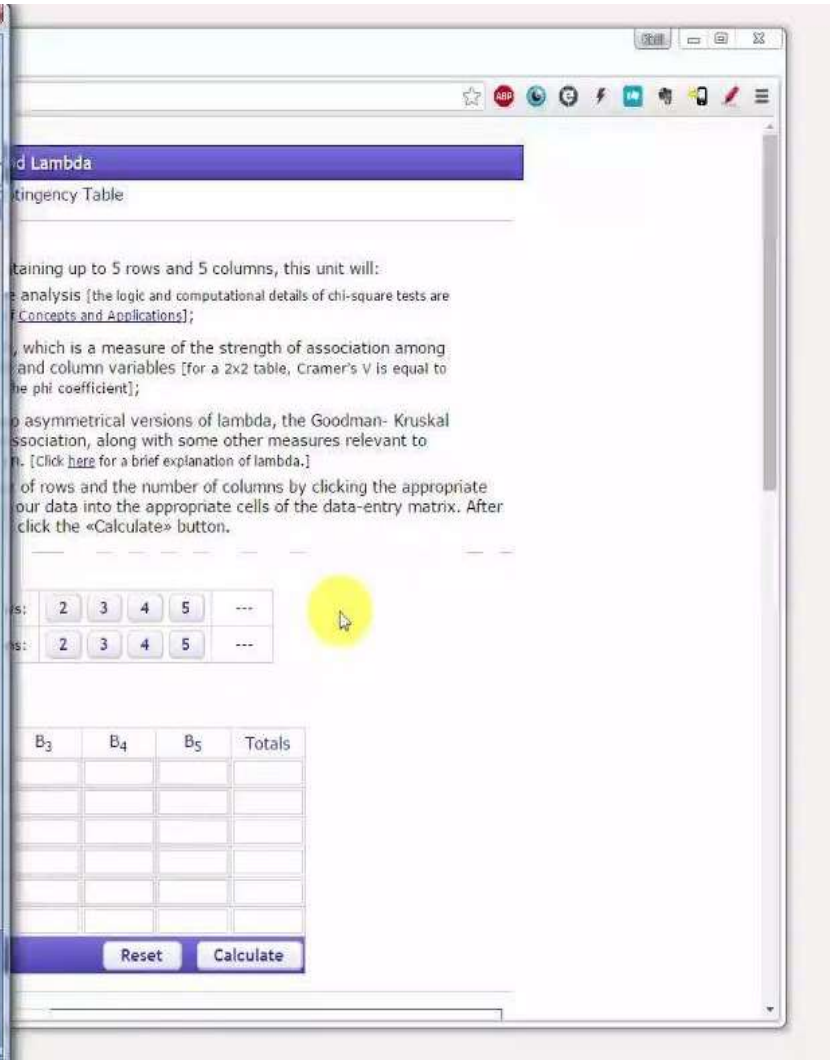
20% x Total France

20% x Total Germany

20% x Total Spain

Introduction to Statistics



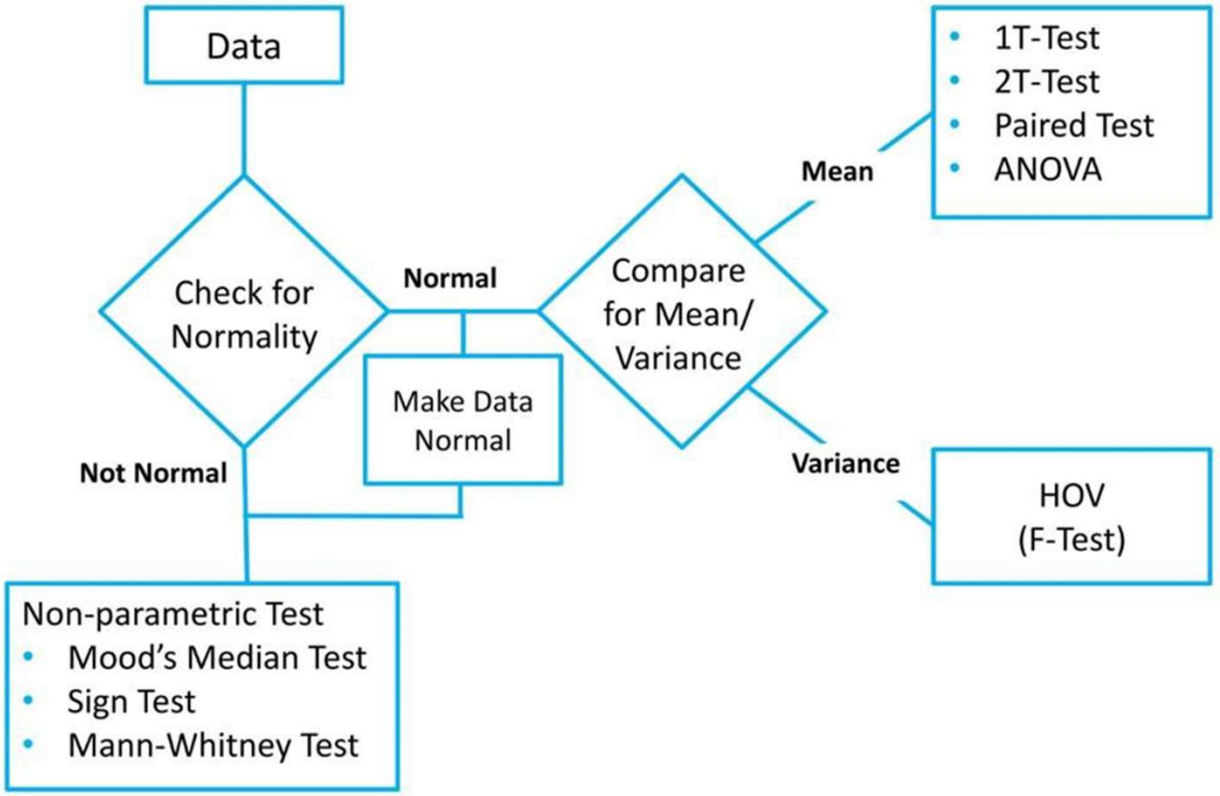


Introduction to Statistics

Agenda

- Overview to Hypothesis Testing
- Chi-Square Test
- Test for continuous Data
- Non-Normal Data
- Correlation and Regression

Selection of the Hypothesis Test



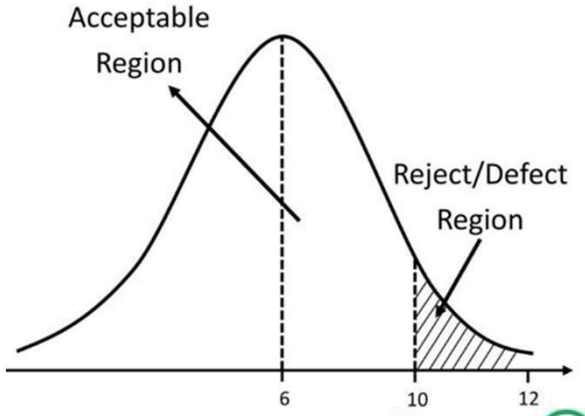
Normality Tests

- Normality test establishes if the data approximately follows a normal distribution
- Tests for hypothesis are different for normal and non-normal data hence the need to first check the distribution type
- Common tests for normality are:

  - Normal probability plot: Graphical method
  - Shapiro-Wilk's W test
  - Anderson Darling test
  - Kolmogorov-Smirnov (KS) test

One-Tailed Tests

- Test where the region of rejection is on only one side of the sampling distribution

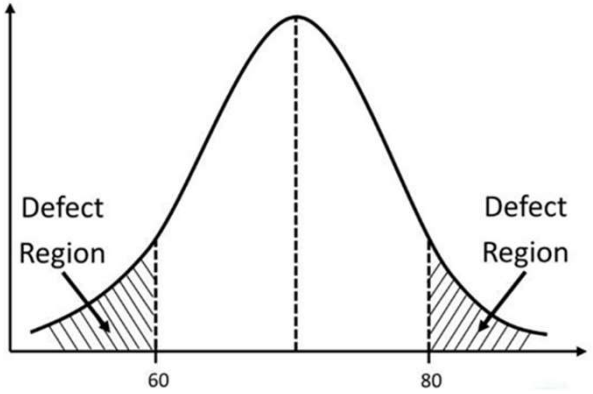


Examples:

- Null Hypothesis: Response time to customer query <=10 minutes
- Alternative Hypothesis: response time > 10 minutes
- Region of rejection would be the numbers greater than 10 (there is no bound on the lesser time interval)

Two – Tailed Tests

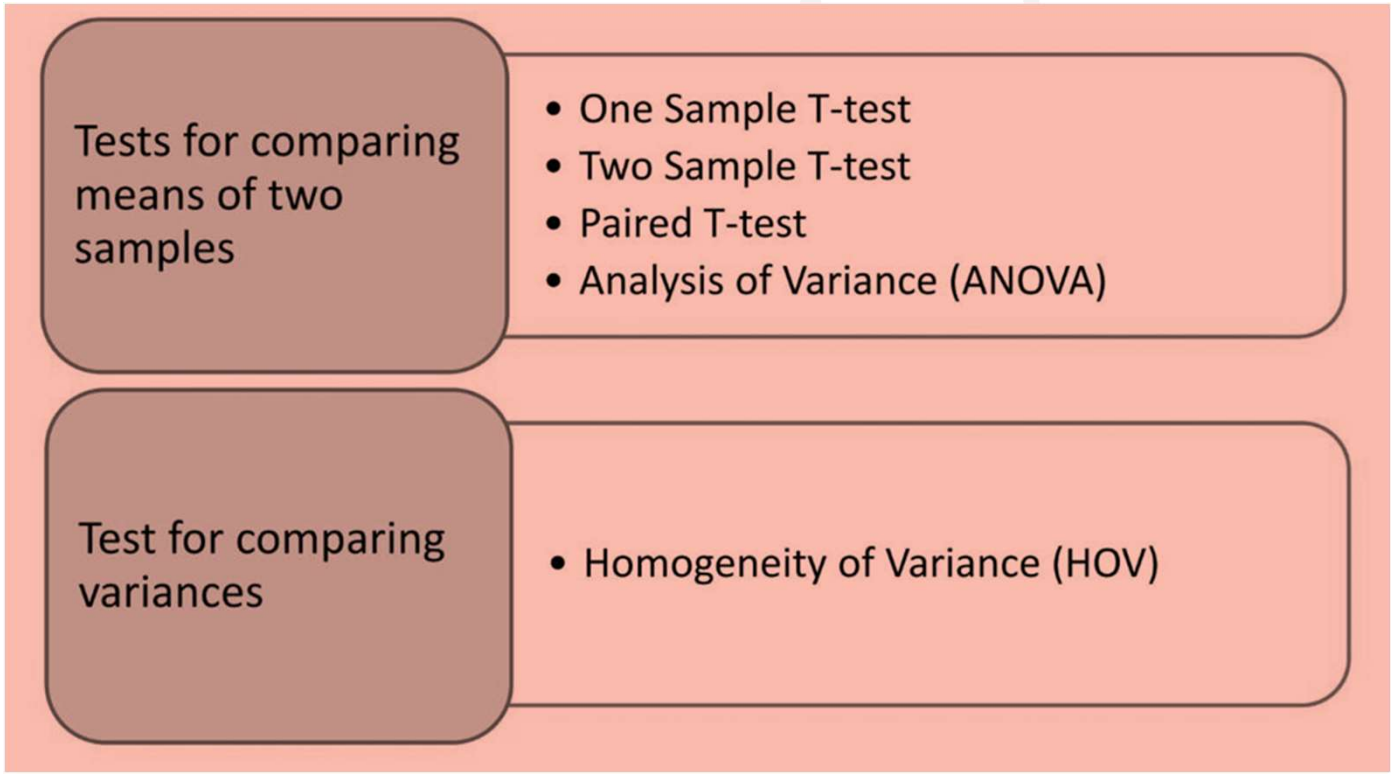
- Test where the region of rejection is on both sides of the sampling distribution



Examples:

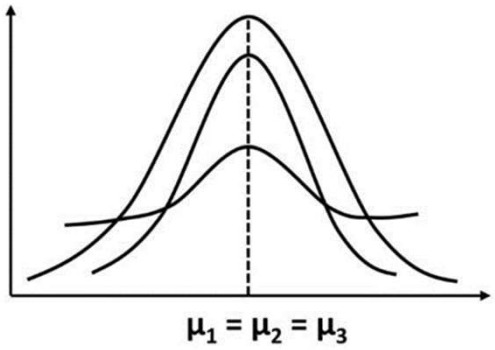
- Speed limit in a freeway 60 – 80 mph (acceptable range of values)
- Region of rejection would be numbers from both sides of the distribution, that is, both <60 and >80 are defects

Hypothesis Tests for Normal Data



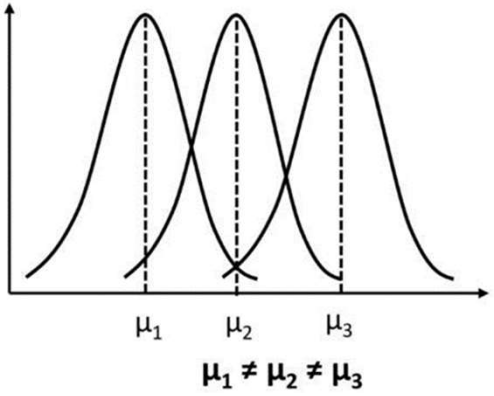
ANOVA – Analysis of Variance

- ANOVA is used for comparing means of more than 2 samples
- Null Hypothesis = Means of all the samples are equal
- Alternative hypothesis = Mean of atleast one of the samples is different
- Variance of all samples is assumed to be similar



Examples:

- Null Hypothesis: Query response time same across all five query categories
- Null Hypothesis: No difference in student performance across the 6 modules of the Analytics course



ANOVA – Analysis of Variance

Introduction

- Analysis of variance is an important statistical technique used to test the hypothesis that the means of two or more populations are equal.
- After discussing how to test the equality of two means by t-test now we will discuss how to conduct hypothesis test when there are more than two population means.
- In case of more than two means, one can also use t-test for comparing means but the chances of type I error increases.

ANOVA – Analysis of Variance

- In order to avoid this situation, in case of more than two population means, the appropriate test statistic for testing equality of more than two means is **analysis of variance**.
- R. Fisher, the father of statistics, developed a technique called 'experimental design' to establish cause and effect relationship between variables. In fact, ANOVA is an important part of a large 'experimental design' setup.
- In ANOVA, we have a dependent variable which is quantitative in nature and one or more independent variables which are categorical in nature.
- The independent variables which are categorical variables are also called **factors**. Combination of factors or categories is called **treatment.**
- When there is a single independent variable or single factor, it is called one-way ANOVA. If there are two or more factors it is termed as n-way ANOVA.

Introduction to Statistics

One-Way ANOVA

- In one-way ANOVA, we have one dependent variable and one categorical independent variable
- The idea to find how much variation in dependent variable is explained by categorical independent variable and how much variation is not accounted by this independent variable.

One-Way ANOVA

- In fact, we will try to decompose total variation in dependent variable (Y) into variation explained by categorical variable (X) and variation not explained by X, that is, error. SS_Y is the total variation in Y.
- SS_X is the variation in Y that is due ti the variations in the means of group of X. SS_{Error} is the variation in Y that is linked with variation within each category of X.
- The total variation in dependent (Y), denoted by $SS_{Y,}$ is decompose into:

$$\mathbf{SS_Y = SS_X + SS_{Error}}$$

One-Way ANOVA

where

$$SS_Y = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$SS_X = \sum_{j=1}^{c} n(\bar{Y}_j - \bar{Y})^2$$

$$SS_{Error} = \sum_{j}^{c} \sum_{i}^{n} (Y_{ij} - \bar{Y}_j)^2$$

Y_i = individual observation

$\bar{Y}_j$ = average for category j

$\bar{Y}$ = Grand mean

Y_{ij} = ith observation in jth category

Example

The ICICI Bank has three branches in New Delhi, and the management wants to find out whether there is any difference in the average business (Rs. Crores) of the three branches. The following table gives data relating to 8 randomly selected months' business at each branch.

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18
24	27	25
20	32	30
16	34	28
22	31	27
26	29	22

Example: Solution

Test the hypothesis that the average businesses of three branches are equal at 5 percent significant level.

Solution:

In this case our null and alternate hypothesis are:

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: at least one mean is different

The above hypothesis will be tested by the F statistic. First we will compute category mean and Grand mean and then various sums of squares will be computed

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18
24	27	25
20	32	30
16	34	28
22	31	27
26	29	22
$\bar{X}_1 = 20.25$	$\bar{X}_2 = 29.37$	$\bar{X}_3 = 25$
Grand Mean	$\frac{20.25 + 29.37 + 25}{3}$ =24.87	

Example: Solution

Now

$SS_Y = (15-24.87)^2 + (18-24.87)^2 + (21-24.87)^2 + (24-24.87)^2 + (20-24.87)^2 + (16-24.87)^2 + (22-24.87)^2 + (26-24.87)^2 + (30-24.87)^2 + (28-24.87)^2 + (24-24.87)^2 + (27-24.87)^2 + (32-24.87)^2 + (34-24.87)^2 + (31-24.87)^2 + (29-24.87)^2 + (26-24.87)^2 + (24-24.87)^2 + (18-24.87)^2 + (25-24.87)^2 + (30-24.87)^2 + (28-24.87)^2 + (27-24.87)^2 + (22-24.87)^2$

$= 600.26$

$SS_X = 8(20.25-24.87)^2 + 8(29.37-24.87)^2 + 8(25-24.87)^2$

$= 332.89$

$SS_{Error} = (15-20.25)^2 + (18-20.25)^2 + (21-20.25)^2 + (24-20.25)^2 + (20-20.25)^2 + (16-20.25)^2 + (22-20.25)^2 + (26-20.25)^2 + (30-29.37)^2 + (28-29.37)^2 + (24-29.37)^2 + (27-29.37)^2 + (32-29.37)^2 + (34-29.37)^2 + (31-29.37)^2 + (29-29.37)^2 + (26-25)^2 + (24-25)^2 + (18-25)^2 + (25-25)^2 + (30-25)^2 + (28-25)^2 + (27-25)^2 + (22-25)^2$

$= 267.37$

It can be verified that

$SS_Y = SS_X + SS_{Error}$

Example: Solution

$600.26 = 332.89 + 267.37$

The above null hypothesis can now be tested as follows:

$$F = \frac{\frac{SS_X}{(c-1)}}{\frac{SS_{Error}}{(N-c)}} = \frac{\frac{332.89}{(3-1)}}{\frac{267.37}{(24-3)}} = 13.07$$

Thus, the calculated F value is 13.07. Now we have to compare this calculated value with critical F value. The critical F value for 2 degrees of freedom in numerator and 21 degrees of freedom in denominator is 3.47 for significance level 0.05. Since the calculated value F is greater than the critical F value, we will reject the null hypothesis. This implies that average business of three branches are not equal.

	A	B	C
2	15	30	26
3	18	28	24
4	21	24	18
5	24	27	25
6	20	32	30
7	16	34	28
8	22	31	27
9	26	29	22

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Branch 1	8	162	**20.25**	14.5
Branch 2	8	235	**29.375**	9.696428571
Branch 3	8	200	**25**	14

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	333.25	2	166.625	**13.08695652**	0.000204	3.466800112
Within Groups	267.375	21	12.73214286			
Total	600.625	23				

Introduction to Statistics

N-way Anova

- We can extend the concept of one-way ANOVA to study the effect of more than one factor.
- For example, how do age of the readers (less than 20, 20-50, more than 50) and educational levels (higher secondary, under graduate, post-graduate, M.phil/Ph.D) affect the circulation of a particular newspaper?
- Similarly, if we want to study the effect of students' familiarity with a university (very high, high, medium, low, very low) and image of the university (highly positive, positive, neutral, negative, highly negative) on the preference for the university, n-way anova can be used to determine such effects.

N-way Anova

This also helps the researchers to find the interactions between the factors. Let us consider two factors, namely, X1 and X2 with categories c1 and c2. In this case, total variation is decomposed into:

$$SS_{total} = SS_{X1} + SS_{X2} + SS_{X1X2} + SS_{Error}$$

Where,
SS_{total} = Total Variation
SSX1 = Variance explained by X1
SSX2 = Variance explained by X2
SSX1X2 = Variance jointly explained by X1 and X2

N-way Anova

One can test the significance of the overall effect by F test given below:

$$F = \frac{(SS_{X_1} + SS_{X_2} + SS_{X_1X_2}) / df_1}{SS_{error} / df_2}$$

Where,

df1 = degrees of freedom in the numerator = c1c2 - 1

df2 = degrees of freedom in the denominator = N – c1c2

N-way Anova

One can test the significance of each factor by the F statistic. For example the significance of X1 is tested as follows:

$$F = \frac{\frac{(SS_{X_1})}{(c_1 - 1)}}{\frac{SS_{error}}{(N - c_1 c_2)}}$$

Similarly, you can test the significance of X2 as follows:

$$F = \frac{\frac{(SS_{X_2})}{(c_2 - 1)}}{\frac{SS_{error}}{(N - c_1 c_2)}}$$

N-way Anova

If you are interested in finding the significance of interaction effect, you can test the null of no interaction effect by the following statistic

$$F = \frac{\frac{(SS_{x_1x_2})}{(c_1 - 1)(c_2 - 1)}}{\frac{SS_{error}}{(N - c_1c_2)}}$$

The decision rule is same. When the calculated F value is greater than critical F value reject the null hypothesis.

Example

Thapar University is an ISO certified university which conducts a student survey every year to assess the satisfaction levels of its students.

The students were asked to rate the university on a scale of 1 to 7 (7 representing excellently) on various attributes of quality.

One of the questions of the survey was overall, how well do you think that the Thapar University has prepared you for a bright career in the corporate sector.

Example

The following data give responses of the students to this question. The students were divided by the regional centres and type of courses offered.

		Centres		
		New Delhi	Hyderabad	Mumbai
Course Type	MBA(Finance)	4 2 6 5	6 2 3 5	3 5 2 6
	MBA(Marketing)	4 4 5 6	6 4 5 6	2 3 3 2

Determine whether there are significant differences in the responses using two-way ANOVA to this question at 5 percent significance level.

Solution

We will test the following hypotheses concerning to two-way ANOVA.

Our null and alternate hypotheses for row effects are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: at least one mean is different

Our null and alternate hypotheses for column effects are:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: at least one mean is different

For interaction effects:

H_0: There is no interaction effect

H_1: There is interaction effect

Excel: Two Factor Without Replication

	New Delhi	Hyderabad	Mumbai
MBA (Finance)	4	6	3
	2	2	5
	6	3	2
	5	5	6
MBA (Marketing)	4	6	2
	4	4	3
	5	5	3
	6	6	2

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
MBA (Finance)	3	13	4.33333	2.33333
	3	9	3	3
	3	11	3.66667	4.33333
	3	16	5.33333	0.33333
MBA (Marketing)	3	12	4	4
	3	11	3.66667	0.33333
	3	13	4.33333	1.33333
	3	14	4.66667	5.33333
New Delhi	8	36	4.5	1.71429
Hyderabad	8	37	4.625	2.26786
Mumbai	8	26	3.25	2.21429

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	10.625	7	1.51786	0.64885	0.71015	2.7642
Columns	9.25	2	4.625	1.9771	0.17528	3.73889
Error	32.75	14	2.33929			
Total	52.625	23				

Excel: Two Factor With Replication

	New Delhi	Hyderabad	Mumbai
MBA (Finance)	4	6	3
	2	2	5
	6	3	2
	5	5	6
MBA (Marketing)	4	6	2
	4	4	3
	5	5	3
	6	6	2

Anova: Two-Factor With Replication

SUMMARY	New Delhi	Hyderabad	Mumbai	Total
MBA (Finance)				
Count	4	4	4	12
Sum	17	16	16	49
Average	4.25	4	4	4.08333
Variance	2.916666667	3.333333333	3.33333	2.62879
MBA (Marketing)				
Count	4	4	4	12
Sum	19	21	10	50
Average	4.75	5.25	2.5	4.16667
Variance	0.916666667	0.916666667	0.33333	2.15152
Total				
Count	8	8	8	
Sum	36	37	26	
Average	4.5	4.625	3.25	
Variance	1.714285714	2.267857143	2.21429	

Excel: Two Factor With Replication

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.041666667	1	0.041667	0.021277	0.885649	4.413873
Columns	9.25	2	4.625	2.361702	0.122798	3.554557
Interaction	8.083333333	2	4.041667	2.06383	0.155967	3.554557
Within	35.25	18	1.958333			
Total	52.625	23				

Introduction to Statistics

Difference between Anova two-factor with replication and without replication

- The fundamental difference **between Anova two-factor with replication** and without replication is that the sample size is different.
- **with-replication**,
  - The total number of samples is mostly uniform.
  - If that is the case, the means are calculated independently.
  - This type of data is also known as balanced data.
  - But if the sample size is not uniform, the analysis becomes difficult. It is better to get the sample size uniform to get faster results.

Difference between Anova two-factor with replication and without replication

- **without replication**,
  - the sample observation size is one.
  - It means that there is only a single observation for each combination of nominal variables.
  - Here, the analysis can be done using the means of both the variables as well as the total mean of considering every observation as a single cluster.
  - The F-ratio can then be calculated by the remainder mean and the total mean.

Agenda

- Overview to Hypothesis Testing
- Chi-Square Test
- Test for continuous Data
- Non-Normal Data
- Correlation and Regression

Discussion

Jenny manages the Claims Processing and Inbound Customer Service teams for a medical insurance company. She wants to analyse the performance of her team and is looking at their issue resolution time for the past 6 months. On plotting the data, she discovers that the data does not follow a normal distribution.

What is the problem with Jenny's data? How can she resolve the normality problem?

Approach to Non-Normal Data

If data is not normal:

- Identify reasons for non-normality and, if possible, address those reasons
- Use techniques that do not require normality of data

Reason for Non-Normality

1. Extreme values
2. Overlap of two or more processes
3. Insufficient data discrimination
4. Sorted data
5. Values close to zero or a natural limit
6. Data follows a different distribution

Hypothesis Tests Summary