Business Research Design and Methods

Research Methodology :-

→ Research :- Research is a logical and systematic search for new and useful information on a particular topic.

Research Methodology :- RM is a systematic way to solve a problem. It is a science of studying how research is to be carried out. Essentially, the procedure by which researchers go about their work of describing, explaining and predicting phenomena are called research methodology. It is also defined as the study of methods by which knowledge is gained. Its aim is to give the work plans of research.

→ Objectives of research Methodology

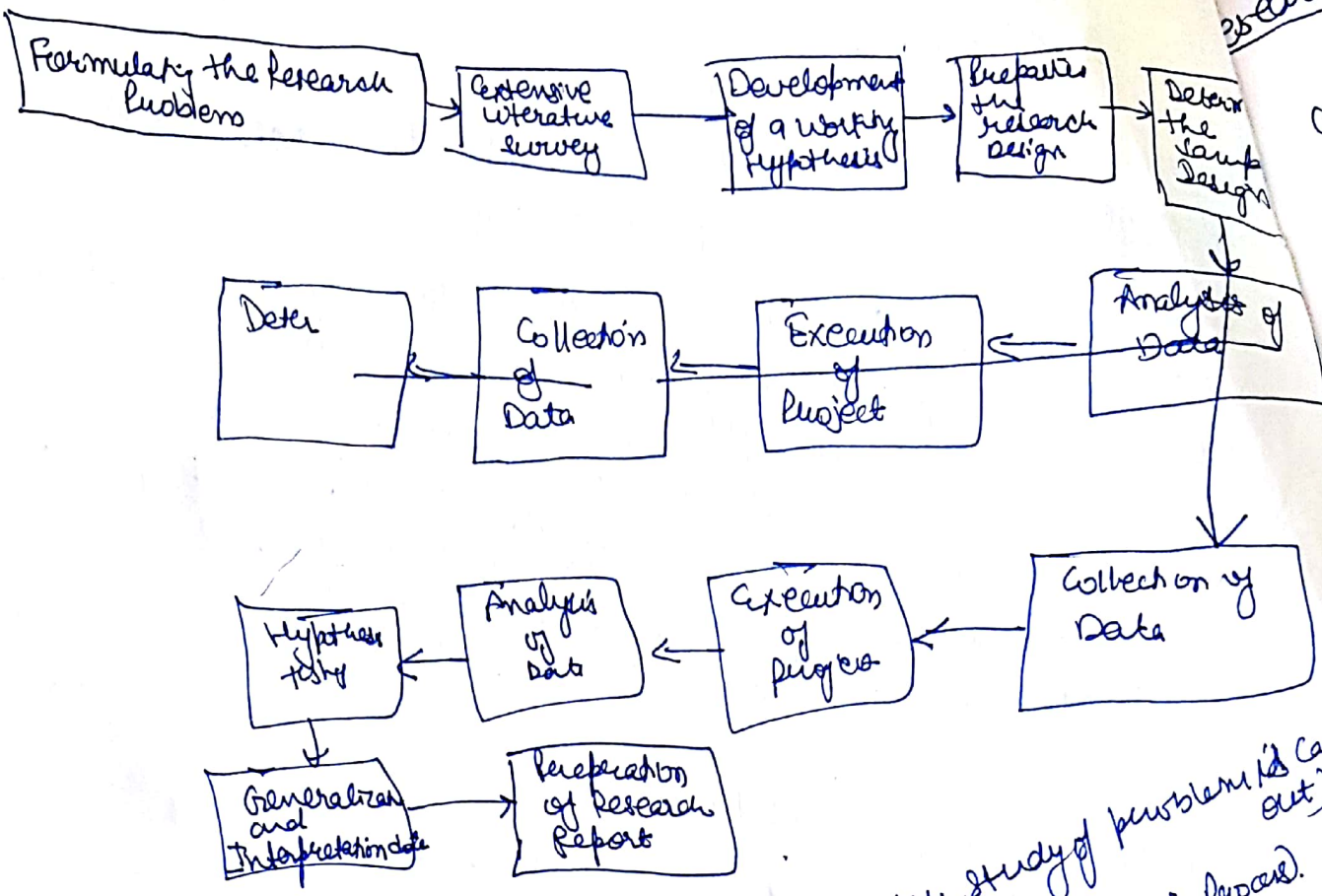
- To gain familiarity or achieve a new insight towards a certain topic
- To verify and test important facts
- To analyze an event, process or phenomenon
- To identify the cause and effect relationship
- To find solutions to scientific, non-scientific and social problems
- To determine the frequency at which something occurs

⇒ Types of research

Fundamental Research (Pure)	Pure Research (Action)
based on question 'Why things happen?'	→ Based on Question 'How things happen?'
- focus on one dicipline	- takes an Inter dicoplinary approach
- descriptive study of research problem	- makes use of problem solving application to solve a specific problem
- Reporting - technical language	- Simple language
- generalisation & formulation of theories	- aims to find solution for problem
- Solve general problem	- deal with practical problem

Scanned by CamScanner

Research Process



⇒ Nature Of Research

- Research is a process (in which in-depth study of problem is carried out)
- Research is properly conducted (helpful in decision making process.)
- The research activities will help us in testing of hypothesis (establish relationship b/w variables)
- The research is a fact finding process.

⇒ Importance (Scope) Of Research

- Research inculcates scientific and inductive thinking and promotes development of logical habits of thinking and organization
- It aids in decision making
- Research plays a dynamic role in several fields and it has increased significance in recent times, it can be related to small business and also to the economy as a whole
- Most of the Government Regulations and Policies are based on and are a result of intensive research
- It significance lies in solving various planning and operational problems
- It involves the study of cause and effect relationship between various variables & help to identify behavior pattern/trends in certain variables

Research Design :- The research Design is the structure with in which research is conducted. A research design is the arrangement of conditions for the collections and analysis of data in a manner that aims to combine to the research purpose. It constitutes the collection, measurement and analysis of data. Research design gives an outline of everything from defining the problem in terms of objective to final analyses of data.

Contents of Research Design

1. Statement of Research objective ie. why the research project is to be conducted
2. Type of data needed
3. Definition of population and Sampling procedure to be followed
4. Time, costs, & responsibility specification
5. Methods, ways, and procedures used for collection of data.
6. Data analysis - tools or methods used to analyze data.
7. Probable output or research outcome and possible actions to be taken based on those outcome.

Types of research design

1. Exploratory Research Design
2. Descriptive Research Design
3. Experimental Research Design.

1. Exploratory Research Design :-

- discover & insights to generate possible explanation.
- help in exploring problem or situation.
- break problem into sub par
  help to form hypothesis
- used to increase familiarity with problem under investigation
- researcher new in area, or when problem is of different type

Scanned by CamScanner

Types of Research Design

- Exploratory
- Descriptive
- Causal or experimental

Exploratory

- Exploratory (Qualitative)
  - In depth Interviews (face to face)
- have to find out the hidden facts
- We have lot of data, but not aware about what is our problem
- so it help to find hidden fact and make us aware about what is our problem/objective
- flexible, versatile.

Methods
- literature Review
- Depth interview
- Focus Group
- Case Study
- Pilot Study
- Experience Study

Purpose of Exploratory Design
- clarifying concepts and defining terms
- make problem for more investigation
- more familiarity with problem
- Developing hypothesis
- Establishing priorities for further investigation

Descriptive

Method
- Observation (observe)
- case study (characteristic study)
- survey (collect info from sample)

printed form

2. Descriptive Research Design :- describe characteristic of our population
- concerned with describing problem & its solution.
  - problem the cleared one, should be in hand
  - rests on one or more hypothesis
  - not flexible like exploratory
  - directed to answer the problem.

3. Causal or Experimental Research Design
- determine the cause the and effect relationship
  - typical form of experiment
  - impact of manipulation on independent variable on dependent variable
  - dependent - sales, profit etc
  - independent - price, product etc

Scanned by CamScanner

eg sales increase with advertisement
effect — cause

basic – experiment – to control conditions
– manipulate variables

hypothesis – assumption – keep throughout research
Approve/disapprove

Experimentation – to tell the relationship between experiment
+ve or -ve direction

Basic issues

1. Manipulation of independent variable
- Experimental Treatment
- Experimental Group (artificial)
- Control Group (natural state)

2. Selection and measurement of dependent variables

3 Selection and assignment of Test Unit

i) Test Unit (subject on which we have to conduct experiment)

ii) Sample selection Error (biasness)

iii) Random Sampling

(iv) Matching (assignment to each group)

4) Control over extraneous variable

a) Constant error. (Comes every time we do research)

EG $O_1 X O_2$
CG: O_3 O_4
$(O_2-O_1)-(O_4-O_3)$

X, O, R

Types of ED

$O_1 - O_2$

(2 EG, 2 CG) (Expert show results)

Quasi ED	True ED	Formal
(X O_1) → One shot design	– Pre test Post test control group	→ Completely Randomized
(O_1 X O_2) → One group pre-test → post-test design	– Post test only control group	→ Randomization block design
→ Static group design	– Solomon 4 Group design	→ Factorial Design
[EG → X O_1; CG – O_2; O_2-O_1]	P4B	(Completely Random)

← B/w heterogeneous

[Sampling Design] → Concept, types and their applica... Stages

→ Sample (finite) → population (finite)
Infinite Universe → population element (infinite / one piece of popn)
→ Census

Sampling is the process of how to select the sample.

Why Sampling is required
1. Reduce Cost
2. Reduce time
3. Less effort
4. Easy to structure Test unit
5. Organised convenience

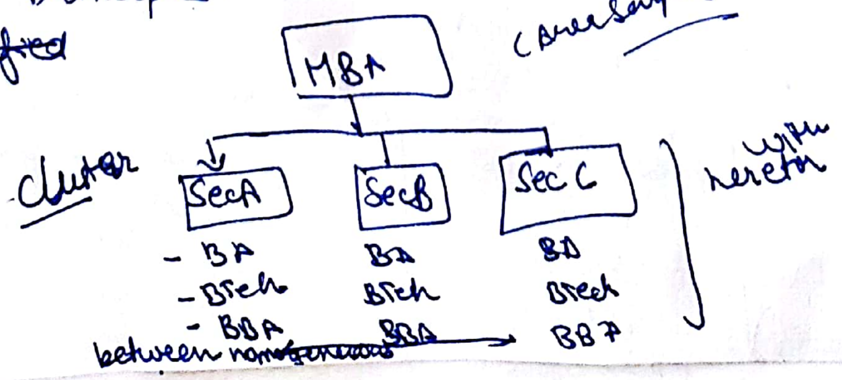
- Studies individual
- Provides accuracy
- Control unlimited data
- Gives greater speed / helps to complete in stipulated time

Steps in Sampling process procedures

- Define the population (element, units, extent & time)
- Specify sampling frame (Telephone directory)
- Specify sampling unit (retailers, our products) (Students, unemployed)
- Specify sampling method / technique
- Determine sampling size
- Specify sampling size (optimum sample)
- Specify sampling plan
- Select the sample.

two Principles
→ Statistical regularity – random
→ large number – (steadiness, stability & consistency)
→ ~~Principles~~.

~~Stratified~~ (Area Sampling)



Stages

Define Target Market (from whom we have to collect data) (On whom I want to conduct research)

↓

Select a Sampling Frame → (prepare list of whom who can take part)

↓

(On what basis collect data) Determine Probability and Non-probability Sampling Techniq

↓

Plan for ~~sample~~ select sample Unit →

↓

Determine sample size.

↓

Select actual Sampling Unit → (final results)

↓

Conduct field work

Sampling Technique

1. Probability Sampling (no biasness - equal chance)
   a) Simple random Sampling
   b) Systematic Sampling
   c) cluster Sampling
   d) Stratified Sampling

Non-Probability Sampling (biasness no-equal chance to get select)
   a) Convineance Sampling
   b) Judgemental Sampling
   c) Snowball
   d) Quota.

Systematic Sampling : procedure → Proceeds with selection of every kth item (k = Population Size / Sample Size)

1st after 10, after 14 Next Next

Cluster → divide on certain criteria (within)

Stratified : homogenous

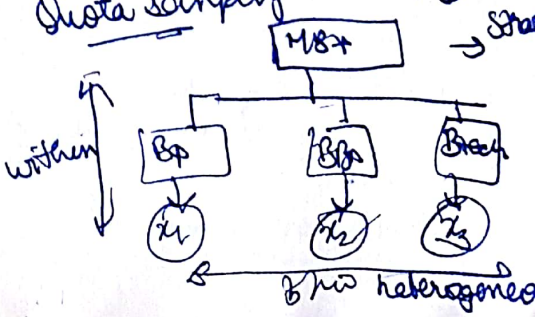
Factor Influencing Sample Size

Stratified - More representative
- Greater accuracy
- Greater Geographical representation

Cluster

Judgmental
Small no. of sample
Business issues
Public policy.

Quota sampling



Scanned by CamScanner

Scaling techniques including Likert, Thurston, Semantic Differential

Scaling techniques etc.

Measurement Scale :- It is the process of observing and recording the observation that are collected as the part of research effort.

Scale
A series of items which are arranged progressively according to value.

Types of Scale:-
1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

→ Nominal Scale
- Simple Scale
- number or alphabets are written on it
eg no. on cricketers shirt
identify

Ordinal Scale
→ According to value.
eg can say number based

Interval Scale : simple mean
- not for objective & attitude
- based on difference
make interval
make SD that
10-15, 15-20, 20-30, 30-40
(divide in groups rates out base)

Ratio Scale Geometric mean
measure quality = absolute zero
relation between two

(Criteria → Reliability & validity

Likert Scale :-
respondent - agree & disagree
make statement

Scaling Technique

Rating Scale
→ Graphic Rating Scale → (Good, bad, neutral)
→ Itemised Rating Scale (numerical scale) (provide number for description)
→ Likert Scale
→ Semantic Differential Scale → (upto-7 point rating bi-polar scale)
→ Stapel's Scale (+ -) (full statement)
→ Thurstone Scale

Ranking Scale (check or rating order)
→ Method of Rank order
→ Method of Paired comparison (compare paired ranks) 2 ranks, 1 rank, 2 object prefer

Scanned by CamScanner

Techniques of Data collection: Primary and Secondary; classification and Tabulation of Data

Data collection: → It is the process to gather information about the relevant topic of research which is being done by researcher

Source of data

- Primary data (Structured (close-ended), Unstructured (open-ended))
  - Questionnaire
  - Observation
  - Interview
- Secondary Data
  - Internal (data collected within company, invoice, receipts, report etc.)
  - External (Outside data, Literature review, Library, Govt reports etc.)

Significance of primary data.

1. Reliability
2. Availability of wide range of Techniques
3. Control.

Limitation

1. cost 2. Time 3. Large Data

Significance of secondary Data

1. Economic
2. Quickness
3. Availability
4. Useful in Exploratory research.

Classification of Data and Tabular Presentation

→ Qualitative (Attribute, characteristic), Quantitive (numerical data), Temporal (time), Spatial (Geographical)

1) Qualitative

Location	Skilled	Unskilled
Rural.		
Urban		

Tabular form

2) Quantitative

Marks	No of Students
0-10	3
10-20	7
20-30	12
30-40	23
40-50	5
	52

Scanned by CamScanner

3. Temporal → Time

Year	Sales (units)
2010	50,000
2011	70,000
2012	90,000
2013	1,00,000
2014	1,50,000

Where one is main characteristic

4. Spatial Classification → Place/location

Indian Students in different countries in year 2016

Country	No. of Students
USA	50,000
UK	15,000
Japan	17000
Russia	2,000
Australia	25,000

Scanned by CamScanner

Business Research Design and Methods

Introduction to Statistics

Statistics is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them. In general, its investigations and analyses fall into two broad categories called descriptive and inferential statistics.

Descriptive statistics deals with the processing of data without attempting to draw inferences from it. The data are presented in the form of tables and graphs. The characteristics of the data are described in simple terms. Events that are dealt with include everday happenings such as accidents, price of goods, business incomes, epidemics, sports data and population data.

Inferential statistics is a scientific discipline that uses mathematical tools to forecasts and projections by analyzing the given data. This is of use to people employed in such fields as engineering, economics, biology, the social sciences, business, agriculture and communication.

I Population and Sample

A population often consists of a large group of specifically defined elements. For example, the population of a specific country means all the people living within the boundaries of that country.

- Usually, it is not possible or practical to measure data for every element of the population under study. We randomly select a small group of elements from the population and call it a sample. Inferences about the population are then made on the basis of several samples.

Example:-

Scanned by CamScanner

Quantitative and Qualitative Data

Data is quantitative if the observations or measurements made on a given variable of a sample or population have numerical values.

Example: height, weight, number of children, blood pressure, current, ...

Data is qualitative if words, groups & categories represent the observations or measurements.

Example: colors, yes-no answer, blood group.

Quantitative — discrete — if corresponding data values take discrete values
→ continuous — if the data values take continuous values

Example of discrete data: number of children, number of cars.
Example of continuous data: speed, distance, time, pressure.

⇒ Elements

- Collection of data
- Classification of data
- Presentation of data
- Comparison
- Analysis and interpretation

⇒ Scope of Statistics

- Nature of Statistics - Science or art
- Relation with others (subject
- Limitation of Statistics
  - deals within groups only
  - only quantitative data
  - based on estimate and approximation
  - Applied for specific purpose
  - No cause and effect relation)

⇒ Functions of Statistics

- Simplifies Complexity
- Measures periodic Changes
- Proper presentation of facts
- formulation of policies
- Study of relationship
- Comparison and forecasting
- Testing and hypothesis

⇒ Relation with others

- Deals with only quantitative data.
- Doesn't study individual events
- True on an average

Economics
Commerce (business)
Management (help in policy making)
Mathematics & Statistical
Natural Science

l tendency :- It is central value which represent the whole data
A measurement of data that indicate where the middle of the formation lies. ~~Mean Central~~
ntral tendency is a way to describe what is typical for a set of data.
entral tendency does not tell you specifics about the individual pieces of data, but it does give you an overall picture of what is going on in the entire data set. There are three major ways to show central tendency : mean, mode and median.

Mean:
Divide the sum of all given value by total number of values

Median:→ Median is a value which divides the whole set into two parts.

steps
1. Set ascending or descending order
2. ~~Arrange the given d~~

(even) $M = \frac{N+1}{2}^{th}$ item

N = No of item or observation.

Odd: $\frac{\text{Sum of two centre item}}{2}$

$M = \frac{4^{th} + 5^{th}}{2}$

$M = \frac{4^{th} \text{ item} + 5^{th} \text{ item}}{2}$.

Discrete series

(i) Arrange data in ascending order
(ii) Calculate CF → Cummulative frequency

$M = \frac{N+1}{2}$ → cf

N = Σf = Total of frequency

Scanned by CamScanner

Mode :- The mode is the most common number in a set of data. ov

e.g. the mode of 1, 2, 2, 3, 5, 6 is 2

* Measures of Dispersion

Dispersion :- Dispersion is the measure of the variation of the items.
- The degree to which numerical data tend to spread about an averag value is called the variation or dispersion of the data.

⇒ Significance of measuring variation

1. To determine the reliability of an average
2. To serve as a basis for the control of the variability
3. To compare two or more series with regard to their variability
4. To facilitate the use of other statistical measures

⇒ Properties of a good measure of variation

(i) It should be simple to understand
(ii) It should be easy to compute
(iii) It should be rigidly defined
(iv) It should be based on each and every item of distribution
(v) It should be amenable to further algebric treatment.
(vi) It should have sampling stability
(vii) It should not be unduly affected by extreme item

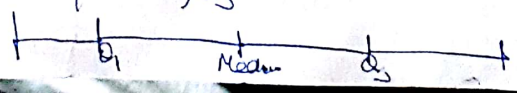
⇒ Measure or methods of dispersion

1. The range
2. The interquartile range and quartile deviation
3. The mean average deviation
4. The standard deviation
5. The Lorenz curve

→ Range = The difference between the greatest and lowest value in the data set.

- Merits and demerits of range.

→ Interquartile range = The difference between the lower (first) Quartile, Q_1 and the upper (third) quartile, Q_3



Scanned by CamScanner

Lower (First) Quartile, Q_1] - Value below which 25% (one-fourth) of the scores lie.

[Upper (Third) Quartile, Q_3] - Value below which 75% (three-fourth) of the scores lie.

→ Box Plot :- Can be created from the lower quartile, Median, Upper Quartile, IQR, and upper and lower Extremes

→ Upper Extreme → Largest data value
→ Lower Extreme → Smallest data value.

Interquartile range = $Q_3 - Q_1$

Quartile deviation = $\frac{Q_3 - Q_1}{2}$

Gives average amount by which two quartiles differ from the media

Coefficient of Q.D = $\frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

⇒ Merits and Limitations of IQR

* Mean absolute deviation :-

Make use of the absolute value to find the distance to find the distance each data point is away from the mean, ~~median~~, found by taking all absolute value differences then averaging them together.

⇒ Standard deviation :- Square root of variance.

→ steps to find standard deviation.

1. Find the mean
2. Subtract the mean from each number
3. Square the difference from step 3
4. Find the sum of square root in step
5. Divide by the number of values
6. ~~Divide by the number of value~~
7. Take the square root

(1 to 5 = variance)

P#13

Scanned by CamScanner

The Coefficient of Variation (CV) is a measure of relative variability the ratio of standard deviation to the mean (average).

$$\text{Coefficient of Variance} = \left(\frac{SD}{\bar{x}}\right) * 100$$

eg. "the S.D is 15% of mean" is a CV

CV useful - compare results of two surveys or tests that have different measures or values.

test A = 90%
B = 25%
more variation related to its mean

CV for population = $\frac{\sigma}{\mu} * 100\%$

CV for sample = $\frac{S}{\bar{x}} * 100\%$

Binomial distribution

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$SD = \sqrt{np(1-p)} = \sqrt{npq}$$

Normal distribution

$$P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-1/2\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$z = \frac{x-\bar{x}}{S}$$

Poisson distribution

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$r = \sqrt{\frac{n-2}{1-r^2}}$$

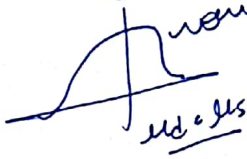
$$\frac{(R_{xy} - R_{xz}) \cdot R_{yz}}{\sqrt{(1-R_{xz}^2)(1-R_{yz}^2)}}$$

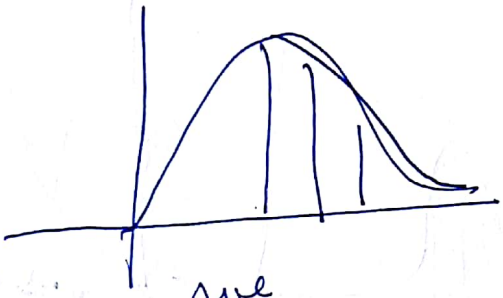
Scanned by CamScanner

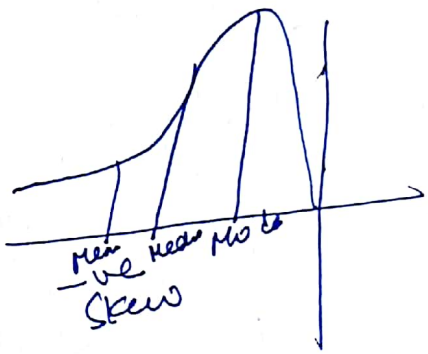
...oness and Kurtosis : Concept and Measures

Skewness :- Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point.

Kurtosis is a measure of wheather the data are heavy-tailed or light-tailed relative to normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.



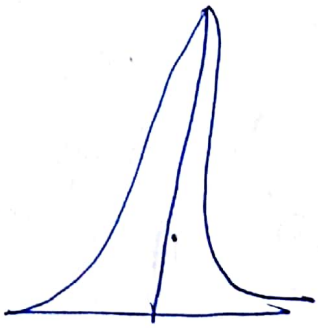




Kurtosis → peakedness of distribution

3 types - differ by amount of peakedness

How values bunched across the centre



μ_4 - fourth moment about mean.

1. Moment coeff of Kurtosis

$$a_4 = \frac{\mu_4}{S^4} = 3 \rightarrow \text{distribution is Meso}$$

SD → S

< 3 platykurtic

> 3 leptokurtic

Scanned by CamScanner

2. Percentile coeff of kurtosis (k)

$= \frac{1}{2}(Q_3 - Q_1)$ → semi-interquartile range.

$= \frac{\frac{1}{2}(Q_3 - Q_1)}{(P_{90} - P_{10})}$

$k = 0.263$ → standard value — Meso kurtic

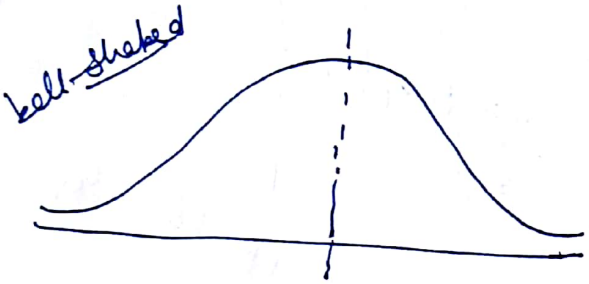
< 0.263 — platykurtic

> 0.263 — leptokurtic

Karl Pearson's coefficient of Skewness

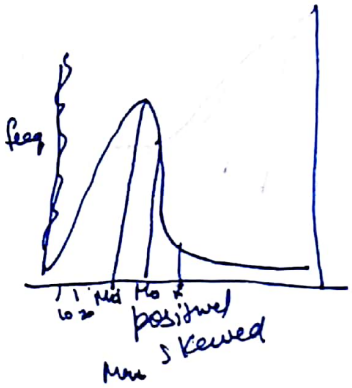
Skewness → off-centre distribution

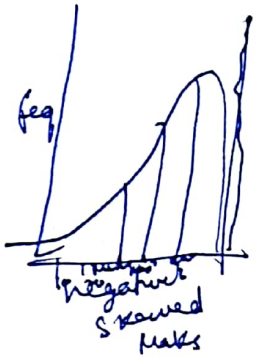
Normal distribution



$\bar{X} = Md = Mode$

$\bar{X} = M_d = M_o$





Karl Pearson

1. Pearsonian coefficient of Skewness

measure relative to dispersion

$Sk_p = \frac{\bar{X} - M_o}{S}$

$\Rightarrow \frac{-}{S.D} \rightarrow +ve$

$\bar{X} - M_o = 3(\bar{X} - M_d)$

$Sk_p = \frac{3(\bar{X} - M_d)}{S}$

Coeff - +ve (+ve skew)

= -ve (-ve skew)

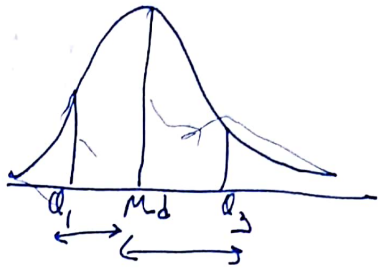
= 0 (i.e. $\bar{X} = M_o$)

0 → Symmetric distribution

Scanned by CamScanner

Quartile coefficient of Skewness

$$SK_Q = \frac{(Q_3 - M_d) - (M_d - Q_1)}{Q_3 - Q_1}$$



Kelly's coefficient of skewed

3 Percentile coefficient of Skewness

Dividing in two equal area

Percentile Skewness $= \frac{(P_{90} - M_d) - (M_d - P_{10})}{P_{90} - P_{10}}$

4. Moment coefficient of Skewness

$$m_r = \frac{\Sigma (x_i - \bar{x})^r}{n}$$

$$\mu_3 = \frac{\Sigma (x_i - \bar{x})^3}{n}$$

moment coefficient $\alpha_3 = \frac{\mu_3}{s^3}$

⇒ Test of Skewness

1) value of mean, median and mode (Not equal)
2) Not bell shaped
3) Positive values of median are not equal to negative value of median
4) Quartile deviation is not different at the right position

0 → Symmetric
+ve - +ve skew
-ve - -ve skew

Skewness - lack of Symmetry
frequency distribution of series goes away from Symmetry
- shows the direction and extent of variation
(limitation of dispersion) →

Page 17.

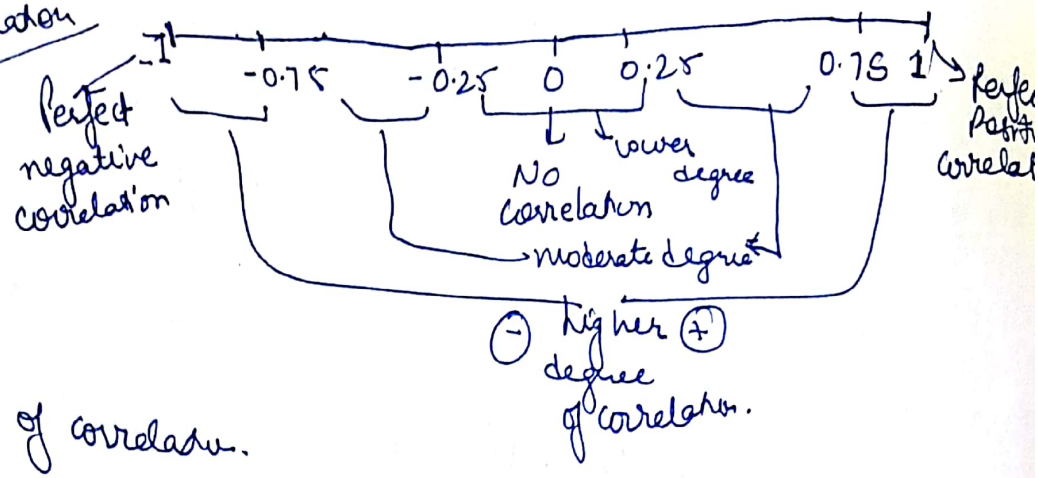
Scanned by CamScanner

Correlation (r)

It is the relationship b/w two or more than two variables

the value of correlation lies between -1 to +1

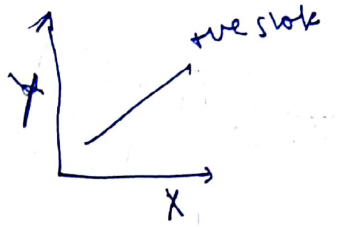
degree of correlation



Types of correlation.

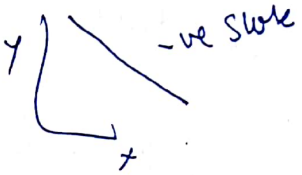
1.) Positive correlation.

X ↑ Y ↑
↓ ↓



② Negative correlation

↑ X ↑ Y ↓
↓ ↑



③ Simple correlation

b/w two variable only

4.) Partial correlation

we have more than two variable, but we study only two variable at a time.

5.) Multiple correlation

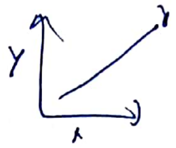
we deal with atleast three variables simultaneously.

Scanned by CamScanner

Linear correlation :-

Situation in which any change in one variable leads to change in other variable in same ratio

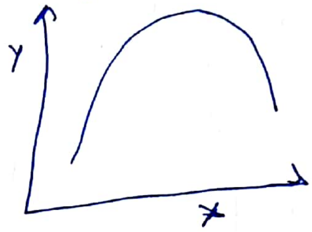
X	Y	X/Y
1	10	1/10
2	20	1/10
3	30	1/10
4	40	1/10
5	50	1/10



7. Non-linear correlation

change in variable, ratio tends to change

X	Y	X/Y
1	10	1/10
2	18	1/9
3	24	1/8
4	28	1/7
5	30	1/6



8. Methods to measure correlation

- Karl pearson's coef of correlation
  - Actual mean method
  - Assumed mean method
- Spearman's rank correlation
  - when ranks are given
  - when ranks are not given
  - when ranks are equal

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$$

$x (X - \bar{X})$ & $y (Y - \bar{Y})$

Assumed mean method

$$r = \frac{N \Sigma dxdy - \Sigma dx . \Sigma dy}{\sqrt{N \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \Sigma dy^2 - (\Sigma dy)^2}}$$

$dx = (X - A)$

$dy = (Y - A)$

A is assumed mean

Scanned by CamScanner

Regression Analysis :-

Regression is a statistical measure used in finance, investing and other disciplines that attempts to determine the strength of relationship between one dependent variable (Y) & series of another changing variable (independent variable)

two types

- Linear Regression
  - use one independent variable to explain or predict the outcome of the dependent variable Y
- Multiple regression.
  - uses two or more independent variable to predict the outcome

Multiple regression :

$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_f x_f + u.$$

Linear regression: $Y = a + bx + u$

a = intercept
b = slope
u = regression residual

Y = the variable that you are trying to predict (dependent variable)

X = independent variable (the variables that you are using to predict Y).

Scanned by CamScanner

main objective of linear regression

- Establish if there is a **relationship** between two variables
  - +ve relation – move together
  - -ve relation – one ↑ other ↓
- More specifically, establish if there is a statistically significant relationship b/w two
- Example: Income & Spending, wage and gender, student height and Exam Score

- Forecast new observations
  - Can we use what we know about the relationship to forecast unobserved values?
  - Examples: what will are sales be over the next quarter
    What will the ROI of a new store opening be contingent on store attributes

→ Variable Role
- dependent variable (Y)
- Independent variable (X)

The Magic: ↑ Linear Equation

$y = a + bx$

$y = mx + b$

state word.

$y = \beta_0 + \beta_1 x$ → coeff. or slop of x

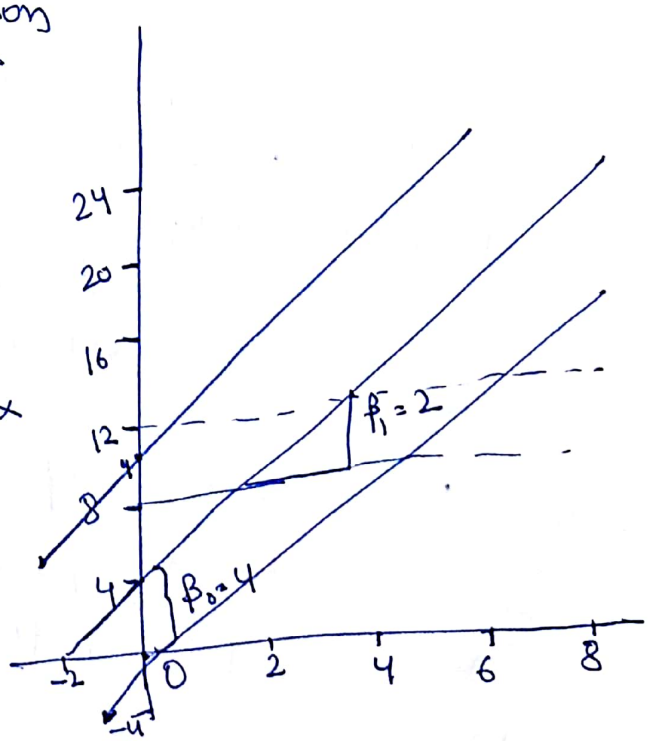
↑ Intercept

$y = \beta_0 + \beta_1 X$

$y = 4 + 2x$

$y = 9 + 2x$

$y = -2 + 2x$



Scanned by CamScanner

Simple linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y — dependent variable
- β_0 — constant or intercept
- β_1 — x's slope or coef.
- x — predicted var
- ε — Error term.

Slope (b)

$$b = r\frac{S_y}{S_x}$$

$$a = \bar{y} - b\bar{x}$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}}$$

Elementary Probability Theory : Concepts, Definitions and Rules

Probability Theory

Sample Space (S) = Set of all possible outcomes in an experiment.

Event = one particular outcome $E \subseteq S$ $E \subseteq S$

Basic probability
How do we find probability of one (or more) things event happen..

ex) expert : flip a coin
event. Get head

Notation $P(E) \rightarrow P(H) = \frac{1}{2}$ or 50%

ex role a dice exp.
Get a 3 event

$$P(3) = \frac{1}{6}$$

Scanned by CamScanner

ex) experiment : Roll 2 dice
event : sum of 7.

$P(\text{Sum of 7}) = \frac{6}{36} = \frac{1}{6}$.

General

$P(E) = \frac{\text{no. of fav. outcomes}}{\text{Total no. of outcomes}}$

Symbols $P(E) = \frac{n(E)}{n(S)}$

$0 < P(E) < 1$

$P(E) = 0$ | E = impossible

$P(E) = 1$ | E = guaranteed

Calculate Probability of Compound Event
↳ OR, NOT, AND

Mutually exclusive events
depended one another
if one happen, other will not happen vice-versa

non-mutual have common things

if $E1 \cap E2 = \phi$ = mutually exclusive
$E1 \cap E2 \neq \phi$ = non-mutually exclusive

$P(H) = \frac{FC}{\text{Total Cases}}$

$(FC + UC)$ → P#10

Scanned by CamScanner

sure event, the probability = 1.

Probability Distributions: Binomial, Poisson and Normal Distributions

A probability distribution is a statistical function that describe all possible value and likelihoods that random variable can take within a given range.

Discrete - countable/finite.

Random variable - Outcome based on chance.

A variable, X, that has value for each outcome for a procedure, that is determined by chance.

Probability distribution :- A table that gives the probability for each value of a random variable.

Ex. DIE : roll a die

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Discrete random variable :- a variable with a countable or finite no. of value.

e.g. Student in a class

Continuous random variable :- a variable with infinite number of possible value

History

GRAM FROM PROB. DIST.

HORIZONTAL : Value of random variable

VERTICAL : Probability

Ex: WEIGHTED DIE

X	P(X)
1	.05
2	.15
3	.35
4	.30
5	.10
6	.05
	1.00

NOTE.

$0 \le P(X) \le 1$

$\Sigma P(X) = 1.00$

⇒ Mean, Variance & SD

Mean

$\mu = \frac{\Sigma(X \cdot F)}{N}$

$\Sigma\left[\frac{X \cdot F}{N}\right]$

$\mu = \Sigma[X \cdot P(X)]$ → Mean expected value

Variance

$\sigma^2 = \Sigma[X^2 \cdot P(X)] - \mu^2$

S.D.

$\sigma = \sqrt{\Sigma\sqrt{[X^2 P(X)]} + \mu^2}$

X	P(X)	X · P(X)	X^2	$P(X)^2$
1	.05	.05		
2	.15	.30		
3	.35	.		
4	.30	1.20		
5	.10	.50		
6	.05	.30		
	1.00	$\Sigma(X \cdot P(X)) = 3.4$		

$\mu = 3.4$

$\sigma^2 \rightarrow 1.34$

$\sigma = \sqrt{1.34}$.

Scanned by CamScanner

Binomial distribution

two possible Outcome

Success/failure

p is complement of q

3 = n = no. of trails

2 = X = no. of time Success (0 - n)

Probability of success $p = .90$

Prob. of not getting success $q = 1 - p = .10$

$$P(X) = \frac{n!}{(n-x)!\,x!} \cdot p^x q^{n-x}$$

$P(X=8) = \frac{13}{(13-8)!\,8!} (.5)^8 (.5)^5$

$n = 13$

$p = .5$

$q = .5 = 1 - .5$ $P(x=8) = .157$

$P(\text{at least } 8) = P(X \geq 8)$

$= P(8) + P(9) + P(10) + P(11) + P(12) + P(13)$

$P(\text{More than } 8) = P(X) > 8 =$

$P(9) + P(10) + P(11) + P(12) + P(13)$

$P(\text{at most } 8) = P(X \leq 8)$

value between 0 & 8

$P(\text{fewer than } 8) = P(X < 8)$

Scanned by CamScanner

Binomial evaluates the probability of an event occuring several times over a given number of ~~trials~~ trials and given the event's probability in each trial.

Poisson distribution :- helps in describing the chance of occurence of a number of events in some given interval or given space conditionally that the value of average number of occurence of even is known. → major use only condition of PPD

$$P(x,\mu) = \frac{(e^{-\mu})(\mu^x)}{x!}$$

μ = avg no. of success
x = no. of occuring success
e = 2.71828

→ Normal PD :-

68% - one SD of mean
95% - two SD of mean
99.7% - 3 SD of mean

M = Md = Mo = equal
total curve are = 1

Hypothesis testing

Types of Hypothesis :-

Null Hypothesis H_0 :

Alternative Hypothesis H_a :

Errors in hypothesis test

Type I error - reject null Hypothesis

Type II error (β) - failure to reject a false null hypothesis

Scanned by CamScanner

Procedure in the hypothesis

1.) State the null hypothesis and alternate hypothesis
2.) Select the appropriate test statistics and level of significance.
3.

testing hypothesis of mean = Z-Statistics

Z-Statistic — σ (S.D) known.
$n > 30$
normal data

t-test — σ unknown
normality or $n > 30$

4) Level of significance

0.10 - political polling
0.05 - consumer research project
0.01 - quality assurance

'α' = probability of rejecting the null hypothesis when it is true.

eg. 'α' = 0.05 → 5% risk of concluding that a difference exist when there is no actual difference

3.) State the decision rule

null hypothesis H_0 - accept or reject.
CV of test - determined by α

4) Compute the appropriate test statistics & make decision.

Z-statistics, $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$, t-Statistics $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

compare t-value & CV
5) Interpret the decision

Scanned by CamScanner

Chi-Square test

don't depend on normal distribution to interpret findings

Purpose

i) to test hypothesis of no association between two or more groups, population or criteria (ie check independence b/w 2 variables)

(ii) to test how likely the observed distribution of data fits with the distribution that is expected.

$$\chi^2_2 \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Expected

Total = 29+24+22+19+21+18 +19+20+23+18 +20+23

O – observed freq.

E – Exp. freq.

Expected 256 / 12 → Total no. of categories = 256

= 21.333

Category	observed	Expected	Residual (obs - Exp)	(obs-Exp)²	(obs - Exp)²/Exp
~~Aries~~	29	21.333	7.667	58.782889	2.755
Taurus	24	21.333	2.667	7.112889	0.333424
Gemini	22	21.333	0.667	0.44889	0.021042
Cancer	19	21.333	-2.333	5.442889	0.255139
leo	21	21.333	-0.333	0.110889	0.005198
Virgo	18	21.333	3.333	11.108889	0.520837
Libra	19	21.333	-2.333	5.442889	0.255139
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricon	18	21.333	-3.333	11.108889	0.520837
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

component = (obs - Exp)² / Exp

Scanned by CamScanner

→ A chi-square test for independence shows how categorical variables are related.
- Variation depends upon - how data collected
  - how hypothesis recorded

→ Chi-Square hypothesis test is appropriate if you have:
- Discrete Outcomes (categorical)
- Dichotomous variable.
- Ordinal variable.

→ Test a chi-square Hypothesis : steps

Sample: 1. 11 degrees of freedom
2. Chi-Square test statistic of 5.094.

Step 1: Take the chi-square statistic
Find p-value in chi-square tables
- Closest value for df = 11 & 5.094 is b/w .900 & .950
p-value 0.9256

Step 2: Use p-value.
Small p-value (1% - 5%) - reject null hypothesis
large p-value (92.56%) - not reject null hypothesis

→ The chi-square distribution:
- Special case of gamma distribution

Let's have random sample taken from ND.
The χ^2 - distribution is the distribution of the sum of these random samples squared

df(k) = no. of samples being summed

eg. 10 sample from ND
then, df = 10

(also skewed in χ^2)

- always right skewed.
high degree of freedom looks like N.D.

P#5

(continuation on next P#4)

Scanned by CamScanner

oduction to analysis of Variance (ANOVA)

Introduction
- Different Types of ANOVA
- Assumptions of ANOVA
- Hypothese in ANOVA
- Main Effect / Interaction Effects
- Post Hoc Test
- The F Distribution

ANOVA is a statistical Method. It is used to compare the means of two or more groups

- Factors (Variables)
- levels

0mg	50 mg	100mg
9	7	4
8	6	3
7	6	2
8	8	4
8	8	4
9	7	3
8	6	2

Factor : "Dosage"
levels - "0mg", "50mg", "100mg"

Types of ANOVA

→ One way ANOVA - one factor with at least two levels, levels are independent

→ Repeated - Measures ANOVA - one factor with atleast two levels, levels are dependent.

Day 1	Day 2	Day 3
9	7	4
8	6	3
7	6	2
8	7	3

Scanned by CamScanner

Factorial Anova — Two or more factor (each of which with least two levels), levels can be either independent, depend Or both (mixed).

		Day 1	Day 2	Day 3
Men →		9	7	4
		8	6	3
		7	6	2
Women →		8	7	3
		8	8	4
		8	7	3

Assumption in ANOVA

1. Normality of Sampling Distribution of Means
The distribution of sample means is normally distributed.
2. Independence of Errors
Errors between cases are independent of one another.
3. Absence of Outliers.
Outlying scores have been removed from the data set.
4. Homogeneity of Variance
Population variances in different levels of each independent variable are equal.

Hypotheses:

ANOVA with one factor ("A", three levels):

H_0: $\mu_{A1} = \mu_{A2} = \mu_{A3}$ ← Main effect

H_1: not all means are equal.

Scanned by CamScanner

NOVA with two factors ("A" & "B"; each with 3 levels):

with $H_0 = \mu_{A1} = \mu_{A2} = \mu_{A3}$

here H_1 = not all mean are equal.

$H_0 = \mu_{B1} = \mu_{B2} = \mu_{B3}$

H_1 : not all means are equal

H_0: interaction absent ← Interaction Effect

H_1 = interaction present. ($A \times B$)

Main Effects in ANOVA

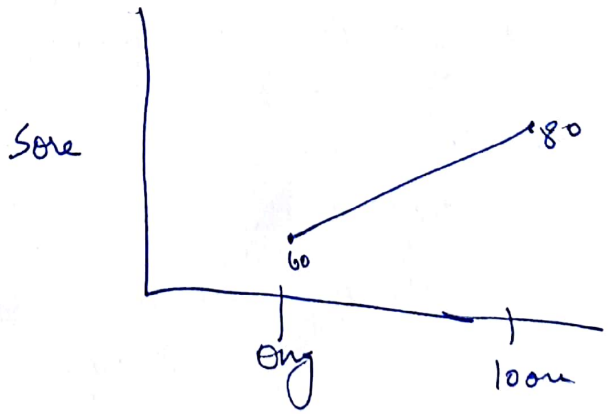
$H_0 = \mu_{A1} = \mu_{A2} = \mu_{A3}$

H_0: not all means are equal

Pretend we are comparing the test score of people who have received a medication (100mg dosage group) and people who have not received a medication (0mg dosage group). The 0mg condition has a mean of 60, while the 100 mg condition has a mean of 80. This could be represented in a graph like this:

0mg - 60

100mg - 80



Scanned by CamScanner

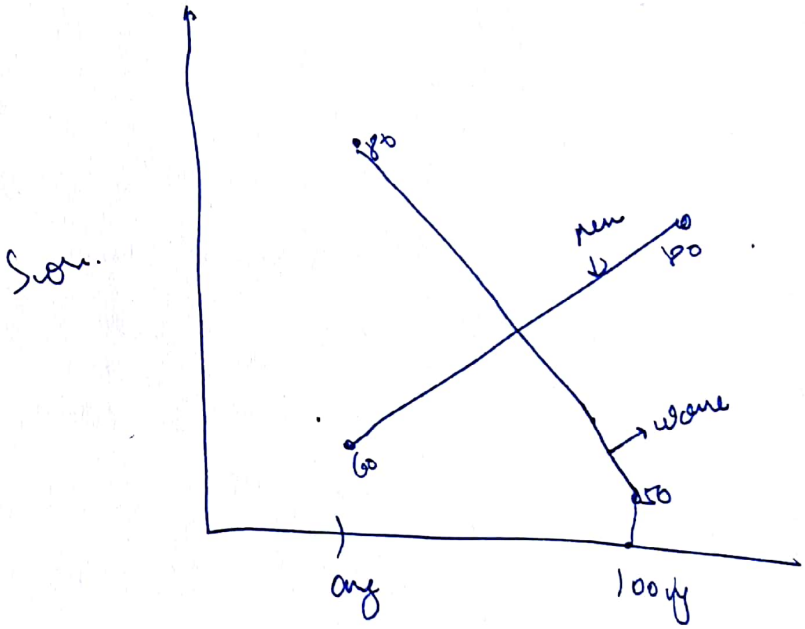
Interaction Effects in ANOVA

H_0: $\mu_{A1} = \mu_{A2} = \mu_{A3}$	H_0: $\mu_{B1} = \mu_{B2} = \mu_{B3}$	H_0: interaction absent
H_1: not all means are equal	H_1: not all means are equal	H_1: interaction absent

2-way ANOVA

Here we have a Factorial ANOVA with two factors: dosage (0mg & 100mg) and gender (men & women). In the 0mg dosage condition, men have a mean of 60 while women have a mean of 80. In the 100 mg dosage condition, men have a mean of 80 while women have a mean of 50. This could be represented in a graph like this.

	M	W
0 mg:	60	80
100 mg:	80	50



⇒ Post-Hoc Analysis in ANOVA

H_0: $\mu_{A1} = \mu_{A2} = \mu_{A3}$

H_1: not all means are equal.

If we reject the null hypothesis, all we know is that there is a difference somewhere among the groups.

Additional test called post hoc test can be done to determine where difference lies.

Scanned by CamScanner

The F distribution in ANOVA-

When doing an ANOVA, we calculate an "F" statistics. It is similar to other statistics such as "Z" & "t".

$$F: \frac{\text{Treatment difference + Random Differences}}{\text{Random Differency}}$$

If there are no treatment differences (that is, if there is no actual effect), we expect F to be 1. If there are treatment differences, we expect F to be greater than 1.

The F statistic has its own one-tailed distribution, much like how the 'z' and 't' statistics have their own separate distributions.

P.T.O.

Scanned by CamScanner

#. Interaction Effects in Anova

Chi-square test (continuation) (a statistical method assessing the goodness of fit between a set of observed values and those expected theoretically).

⇒ Uses

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation
- Independence of two criteria of classification of qualitative variables
- Relationship between categorical variable (contigency table).
- Sample variance study when the underlying distribution is normal
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The Chi-square test (a goodness of fit test).

⇒ Chi distribution

It distributes Square root of a variable distributed according to Chi-square distribution with $df = n > 0$ degrees of freedom over probability density function of:

$$f(x) = 2^{(1-n/2)} x^{(n-1)} e^{(-(x^2)/2)} / \Gamma(n/2)$$

⇒ Chi-Square p-values

- tells us the result is significant or not.

(tell us deviation b/w O & E values is by chance or because of some fac

Two things needed

1. Df. (no. of categories − 1)
2. alpha level (α).
   usual (α) = 0.05 (5%)
   could be 0.01 or 0.10

Scanned by CamScanner