University Grants Commission

Subject: Economics Code: 01

Unit – 3: Statistics and Econometrics

Contents

Subunit – 1: Theory of Probability

Sl. No.	Topics
1	3.1.1. Random Experiment
2	3.1.2. Out Come
3	3.1.3. Event
4	3.1.4. Sample's Space
5	3.1.5. Classical definition of Probability
6	3.1.6. Mutually Exclusive event
7	3.1.8. Equally Likely
8	3.1.9. Set Theory
9	3.1.10. De. Morgan's Law
10	3.1.11. Conditional Probability
11	3.1.11.1. Independent event / with replacement / Unconditional probability
12	3.1.11.2. Dependent event / without replacement / Conditional probability
13	3.1.12. Bayes Theorem (State and Proof)
14	3.1.13 Axiom of the probability
15	3.1.14 Mutually independent event
16	3.1.14.1 Pairwise Independent event
17	3.1.15 State and proved the addition theorem / Probability Statement
18	3.1.16 State and prove the theorem of compound probability or conditional probability or multiplication theorem

Subunit – 2: Descriptive Statistics – Measure of Central Tendency & Dispersion, Correlation

Sl. No.	Topics
1	3.2.1 Measure of Central Tendency
2	3.2.1.1. Mean
3	3.2.1.2. Arithmetic Mean
4	3.2.1.3. Geometric Mean
5	3.2.1.4. Harmonic Mean
6	3.2.2. Median
7	3.2.3. Mode
8	3.2.3.1. Range (Absolute measure of Dispersion)
9	3.2.3.2. Co-efficient of Range (Relative measure of Dispersion)
10	3.2.3.3. Quartile Deviation (Absolute measure of Dispersion)
11	3.2.3.4. Co-efficient of Quartile Deviation (Relative measure of Dispersion)
12	3.2.3.5. Mean Deviation (Absolute measure of Dispersion)
13	3.2.3.6. Co-efficient of Mean Deviation (Relative measure of Dispersion)
14	3.2.3.7. Standard Deviation (Absolute measure of Dispersion)
15	3.2.3.8. Co-efficient of Variation (Relative measure of Dispersion)

16	3.2.3.9. Skewness and Kurtosis
17	3.2.3.10. Correlation and Regression Analysis
18	3.2.3.10.1. Correlation Analysis
19	3.2.3.10.2. Correlation and Dependence
20	3.2.3.10.3. Covariance
21	3.2.3.10.4. Correlation Coefficient (r)
22	3.2.3.11. Probable Error
23	3.2.3.12. Rank Correlation
24	3.2.3.13. Coefficient of Determination (R^2)
25	3.2.3.14. Regression Analysis
26	3.2.3.14.1. Explained and Unexplained Variation
27	3.2.3.15. Correlation Ratio

Subunit – 3: Sampling Methods & Sampling Distribution

Sl. No.	Topics
1	3.3.1. Sampling
2	3.3.1.1. Probability sampling
3	3.3.1.2. Non probability sampling
4	3.3.2. Sampling Methods
5	3.3.3. Simple Random Sampling
6	3.3.4. Systematic Sampling

7	3.3.5. Stratified Sampling
8	3.3.6. Over Sampling
9	3.3.7. Probability-Proportional-to-Size Sampling
10	3.3.8. Cluster Sampling
11	3.3.9. Quota Sampling
12	3.3.10. Mini-Max Sampling
13	3.3.11. Accidental Sampling
14	3.3.12. Line-intercept sampling
15	3.3.13. Panel Sampling
16	3.3.14. Snowball Sampling
17	3.3.15. Multistage Sampling
18	3.3.16. Purposive Sampling
19	3.3.17. Errors in sample surveys
20	3.3.18. Sampling Error
21	3.3.19. Non-Sampling Error
22	3.3.20. Standard Error

Subunit – 4: Statistical Inferences, Hypothesis Testing

Sl. No.	Topics
1	3.4.1. Meaning of Statistical Inference
2	3.4.2. Point Estimation
3	3.4.3. Interval Estimation
4	3.4.4. Statistical Hypothesis
5	3.4.5. Statistical Test
6	3.4.6. Region of Acceptance
7	3.4.7. Region of Rejection / Critical Region
8	3.4.8. Critical Value
9	3.4.9. Power of a Test $(1 - \beta)$
10	3.4.10. Level of Significance (α)
11	3.4.11. P-Value
12	3.4.12. Statistical Significance
13	3.4.13. One- Tailed and Two-Tailed Tests
14	3.4.14. Types of Errors
15	3.4.14.1. Type- I Error
16	3.4.14.2. Type- II Error
17	3.4.15. Degrees of Freedom
18	3.4.16. Use of Different Statistical Tests
19	3.4.16.1. Chi-Square (χ^2) Test
20	3.4.16.2. 't'-Test
21	3.4.16.3. 'z'-Test
22	3.4.16.4. 'F'-Test
25	3.4.16.5. Analysis of Variance (ANOVA)

Subunit – 5: Linear Regression Model and their properties - BLUE

Sl. No.	Topics
1	**3.5.1. Stochastic**
2	**3.5.2. Coefficient of Determination (R^2)**
3	**3.5.3. Simple Linear Regression**
4	**3.5.4. Least Squares Method**
5	**3.5.5. Stepwise Regression**
6	**3.5.6. Main approaches**
7	**3.5.7. Regression Fallacy**
8	**3.5.8. Properties of OLS Regression Estimators**

Subunit – 6: Identification Problem and Simultaneous Equation System

Sl. No.	Topics
1	**3.6.1. Definition**
2	**3.6.2. Structural forms**
3	**3.6.2. Reduced form**
4	**3.6.2.1. Seemingly unrelated equation**
5	**3.6.2.2. Recursive equation model**
6	3.6.3. Conditions for Identification Problem
7	3.6.3.1. The order Condition for Identification
8	**3.6.3.2. The rank condition for identification**
10	**3.6.3.3. The properties of the condition of identification**

Subunit – 7: Lagged Variables & Distributed Lag Models

Sl. No.	Topics
1	3.7.1. Meaning
2	3.7.1. Some Examples of Lagged Variable Model
3	3.7.2. Some Models
4	3.7.2.1. Adaptive Expectation Model (by P. Cagan)
5	3.7.2.2. Partial Adjustment Model (by Nerlove)
6	3.7.2.3. Koyck's Geometric Lag Scheme/Koyck's Model

Subunit – 8: Multicollinearity

Sl. No.	Topics
1	3.8.1. Definition of Multicollinearity
2	3.8.2. Consequences of Multicollinearity
3	3.8.3. Partial and Marginal Significance of Regressors
4	3.8.4. Solution for the Incidence of Multicollinearity
5	3.8.5. Relevance of zero mean assumption in linear regression
6	3.8.6. Relevance of the assumption that X is non stochastic
7	3.8.7. Relevance of the least square method in classical linear regression

Subunit – 9: Panel Data and Dummy Variable

1	3.9.1. Example of panel data
2	3.9.2. Least square dummy variable
3	3.9.3. Why Panel data is used?
4	3.9.4. Limitations of fixed effect model
5	3.9.5. Give an example of the regression model having a mixture of quantitative and qualitative variable
6	3.9.6. Interaction Effect
7	3.9.7. Advantages of Dummy variable regression
8	3.9.8. Dummy variable is used in seasonal analysis

Subunit – 10: Time Series Analysis

SL. No.	Topics
1	3.10.1. Time-series processes
2	3.10.2. White noise
3	3.10.3. Unit Root Test (Dickey–Fuller Test)
4	3.10.4 Explain the concept of cointegration in time series analysis:
5	3.10.5 Distinguish between Auto Regressive (AR) and Moving Average (MA)

Subunit – 1: Theory of Probability

3.1.1. Random Experiment:

Random experiment is those experiment whose results depends upon the chance. e.g., Coin tossing, dice throwing etc.

3.1.2. Out Come:

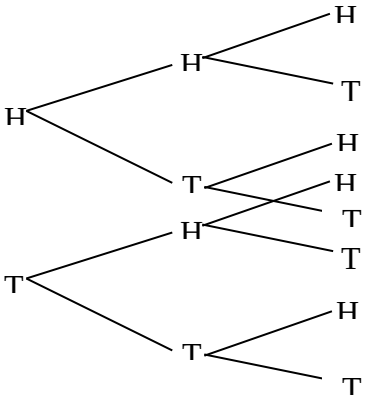
The result of a random experiment is called the outcome.

e.g., (i) If we tossed a coin one time the outcomes are H, T, where H denoted the head and T denoted Tail.

(ii) If we tossed two coins one time or one. Coin two times, the outcomes are

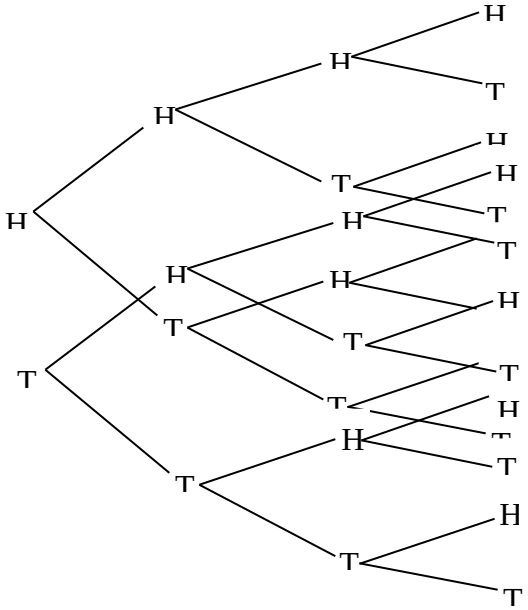
HH, HT, TH, TT.

(iii) If we tossed three coins one time or one coin three times the outcomes are,



H H H
H
H H H
T
H T H
T
H T H
T

(iv) If we tossed four coins one time or one coin four times, the outcomes are



H H H H T H H H
H T H H T T H H
H
H T H T T T H H
T
H T T H T T T H
H
H T T T T T T T

3.1.3. Event:

One or more outcomes together are called the event. Event are two types →
i) Elementary Event: Elementary event are those events which we cannot decompose into several events.

ii) Composite Event: Composite events are those events which we can decompose into several events.
e.g. → If we tossed a coin two times, HH and TT are elementary event. And HT, TH are the composite event.

3.1.4. Sample's Space:

The set of all the possible outcomes of a random experiment is called the sample's space which is denoted by S.
e.g., → If we tossed a coin are time the sample's space is $S = \{H, T\}$.

3.1.5. Classical definition of Probability:

The probability of an event A is denoted by $P(A)$ and according to classical definition.

$$P(A) = \frac{No\ of\ favourable\ cases\ to\ the\ event}{Total\ No.of\ mutually\ exclusive,exhaustive\ and\ equally\ likely\ set\ of\ outcome}$$

3.1.6. Mutually Exclusive event:

Mutually exclusive events are those events which cannot occur simultaneously. So, they are mutually exclusive events.

3.1.7. Exhaustive:

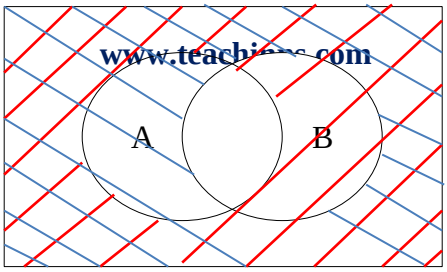
Two more events are said to be exhaustive, if at least one of them will occur.
e.g., → If we tossed a coin one time the events are H, T.
They are exhaustive, because at least one of them must occur.

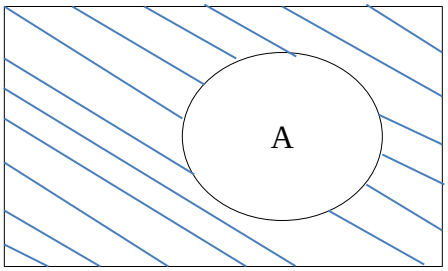
3.1.8. Equally Likely:

Two or more events are said to equally likely if they have equal chances to occur.
e.g., → If we tossed a coin one time, the events are H, T.
$P(H) = \frac{1}{2},\ P(T) = \frac{1}{2}$.
So here the events are equally likely events.

3.1.9. Set Theory:

$U = Union\ (+)$
$\cap = Intersection\ (X)$.
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
But when A and B are mutually exclusive events then $(A \cap B) = 0$
$\therefore P(A \cup B) = P(A) + P(B)$.
$\therefore P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$





3.1.10. De. Morgan's Law:

$P(A' \cup B') = P(A \cap B)'$

$P(A' \cap B') = P(A \cup B)'$

$= 1 - P(A \cup B)$

❖ Show that $P(A') = 1 - P(A)$

$P(A \cup A') = P(S)$

$or, P(A) + P(A') - P(A \cap A') = 1$

$or, P(A) + P(A') = 1 \quad [\because A \text{ and } B \text{ are mutaully exclusive}]$

$or, P(A') = 1 - P(A) \quad (Proved)$

3.1.11. Conditional Probability:

3.1.11.1. Independent event / with replacement / Unconditional probability

Event A and B are said to be independent if $P(AB) = P(A \cap B) = P(A) \cdot P(B)$

3.1.11.2. Dependent event / without replacement / Conditional probability

$P(A \cap B) = P(A) \cdot P(B/A) \qquad [B \text{ given } A = \frac{B}{A}]$

$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\ P(B/A)}{P(B)}$

$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)\ P(A/B)}{P(A)}$

3.1.12. Bayes Theorem (State and Proof):

Statement: An event A can occur only when one of the M.E, exhaustive and equally likely set of events $B_1, B_2, \ldots\ldots\ldots, B_n$ occur.

The unconditional probability $P(B_1), P(B_2), \ldots\ldots\ldots, P(B_n)$, and the conditional probability $P(A/B_1), P(A/B_2), \ldots\ldots\ldots\ldots\ldots, P(A/B_n)$ are the given. Then the probability of event $B_i (i = 1,2, \ldots\ldots, n)$ when A has already occurred

i) $P(B_i/A) = \frac{P(B_i) \cdot P(A/B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)}$

Proof: Event A can occur with $B_i (i = 1,2, \ldots\ldots, n)$ in $B_1A, B_2A, \ldots\ldots\ldots, B_nA$ ways.

So,

$P(A) = P(B_1A) + P(B_2A) + \ldots\ldots\ldots + P(B_nA)$

$P(A) = P(B_1) \cdot P\left(\frac{A}{B_1}\right) + P(B_2) \cdot P\left(\frac{A}{B_2}\right) + \ldots\ldots\ldots + P(B_n) \cdot P\left(\frac{A}{B_n}\right)$

$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A/B_i) \ldots\ldots\ldots\ldots\ldots\ldots (i) \quad [using\ multiplication\ theorem]$

Now,

$P(B_iA) = P(AB_i)$

$P(B_i) \cdot P\left(\frac{A}{B_i}\right) = P(A) \cdot P\left(\frac{B_i}{A}\right)$

$\therefore P\left(\frac{B_i}{A}\right) = \frac{P(B_i) \cdot P\left(\frac{A}{B_i}\right)}{P(A)} = \frac{P(B_i) \cdot P\left(\frac{A}{B_i}\right)}{\sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)}$ [From (i)] (Proved)

❖ **Show that, $P(A\,B) \leq P(A) \leq P(A + B) \leq P(A) + P(B)$**

Ans: Let us consider $P(AB) = P(A) \cdot P\ (B/A)$

Now,
$P(B/A)$ is a probability i.e., $P(B/A) \leq 1$
When it is equal to 1 then $P(AB) = P(A)$……(1)
But when it is less than 1 then, $P(AB) < P(A)$……(2)
Combining (1) and (2) we have, $P(AB) \leq P(A)$……(X)
Now,
$P(A+B) = P(A) + P(B) - P(AB)$
$= P(A) + P(AB) + P(A^cB) - P(AB)$
$= P(A) + P(A^cB)$
Since, $P(A^cB)$ is a probability, i.e., $P(A^cB) \geq 0$
When it is $= 0$ then $P(A+B) = P(A)$……(3)
When it is > 0 then $P(A+B) > P(A)$……(4)
Combining (3) and (4) we get,
$P(A+B) \geq P(A)$……(Y)
Now, $P(A+B) = P(A) + P(B) - P(AB)$
Now, $P(AB)$ is a probability, i.e., $P(AB) \geq 0$
When it is equal to 0 then $P(A+B) = P(A) + P(B)$…… (5)
When it is equal to > 0 then $P(A+B) < P(A) + P(B)$…… (6)
Combining (5) and (6) we have,
$P(A+B) \leq P(A) + P(B)$…… (Z)
Now combining (X), (Y) and (Z) we get,
$P(A\,B) \leq P(A) \leq P(A+B) \leq P(A) + P(B)$ (Proved)

❖ **If event A and B are independent prove that A^c and B^c are also independent.**
Ans:- Since A and B are independent, we can write $P(AB) = P(A) \cdot P(B)$
Now, A^c, B^c will be independent if $P(A^cB^c) = P(A^c)\,P(B^c)$
Now,
$P(A^cB^c) = P(A \cup B)^c$
$= 1 - P(A \cup B)$
$= 1 - [P(A) + P(B) - P(A \cap B)]$
$= 1 - P(A) - P(B) + P(A) \cdot P(B)$
$= [1 - P(A)] - P(B)[1 - P(A)]$
$= [1 - P(A)][1 - P(B)$
$P(A^cB^c) = P(A^c) \cdot P(B^c)$ (Proved)

3.1.13. Axiom of the probability:

Let S be a samples space of a Random experiment. If each event A of the set of all possible events of S, we associate or number $P(A)$, then $P(A)$ called the probability of event A, if the following axioms are satisfied.
i) $P(A) \geq 0$
ii) $P(S) = 1$
iii) For n mutually events of $(A_1, A_2, \ldots\ldots, A_n)$ then
$P(A_1 \cup A_2 \cup \ldots\ldots \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots\ldots \ldots + P(A_n)$.

❖ **Show that $P(A^c) = 1 - P(A)$ with the help of axioms of probability:**

A and A are mutually exclusive and

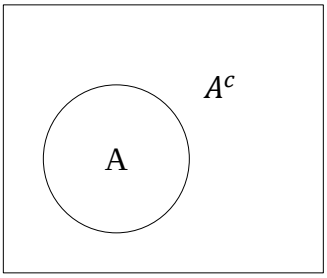
$P(S) = P(A \cup A^c) = P(A) + P(A^c)$ $[By\ axiom\ (3)]$

Now by axiom (2)

$P(S) = 1$

$or,\ P(A) + P(A^c) = 1$

$or,\ P(A^c) = 1 - P(A)$



❖ **Show that $0 \leq P(A) \leq 1$**

By axiom (1) we get, $P(A) \geq 0$

Now, we know that $P(A^c) = 1 - P(A)$

$or, P(A) = 1 - P(A^c)$

So, $P(A^c)$ is a probability then, $P(A^c) \geq 0$

When $P(A^c) = 0$ then $P(A) = 1 \ldots\ldots\ldots\ldots\ldots (i)$

When $P(A^c) > 0$ then $P(A) < 1 \ldots\ldots\ldots\ldots\ldots\ldots (ii)$

Combining (i) and (ii) we have, $P(A) \leq 1$

∴ This is proved that, $0 \leq P(A) \leq 1$

❖ **When $A \subseteq B$ [A *is subset of* B] then prove that $P(A) \leq P(B)$.**

A and $(A^c \cap B)$ are mutually exclusive.

From axiom (3) we have, $P(B) = P(A) + \mathrm{P}(A^c \cap B)$

When $P(A^c \cap B)$ is a probability then $P(A^c \cap B) \geq 0$

Now, when it is > 0 then $P(B) > P(A) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$

and when it is $= 0$ then $P(B) = P(A) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$

From (i) and (ii) we get,

$P(B) \geq P(A)$

∴ $P(A) \leq P(B)$ (proved)

3.1.14. Mutually independent event:

Several events $(A_1, A_2, \ldots\ldots\ldots\ldots, A_n)$ are said to be mutually independent event when the probability of the joint occurrence will be equal to the product of the probability.

e.g., → In cases of 3 events, we have, (A_1, A_2, A_3)

$P(A_1 \cap A_2) = P(A_1)\ P(A_2)$

$P(A_2 \cap A_3) = P(A_2)\ P(A_3)$

$P(A_1 \cap A_3) = P(A_1)\ P(A_3)$

$P(A_1 \cap A_2 \cap A_3) = P(A_1)\ P(A_2)\ P(A_3)$.

3.1.14.1. Pairwise Independent event:

Several events $(A_1, A_2, \ldots\ldots\ldots\ldots, A_n)$ are said to be pairwise independent event when every pair of this events are independent.

e.g.→ For three events, (A_1, A_2, A_3)

$P(A_1 \cap A_2) = P(A_1)\ P(A_2)$

$P(A_2 \cap A_3) = P(A_2)\ P(A_3)$

$P(A_1 \cap A_3) = P(A_1)\ P(A_3)$

3.1.15. State and proved the addition theorem / Probability Statement:

If the two events A and B are mutually exclusive, then the probability of the occurrence of either A or B is given by the sum of their probability, i.e.,
$P(A+B) = P(A) + P(B)$ this is known as addition theorem.
Proof: Let, a random experiment has a possible outcomes, which are mutually exclusive, exhaustive, equally likely and out of $n\ cases\ there\ are\ m$ favourable cases in the event $A\ and\ m$ favourable cases in the $event\ B$. Then,
$P(A) = \frac{m_1}{n}\ \left[\frac{favourable\ cases}{total\ cases}\right]$
$P(B) = \frac{m_2}{n}$
Now, the number of cases favourable to either $A\ or\ B$ is $(m_1 + m_2)$
$\therefore P(A+B) = P\left(\frac{m_1+m_2}{n}\right)$
But $\left(\frac{m_1+m_2}{n}\right) = \frac{m_1}{n} + \frac{m_2}{n} = P(A) + P(B)$
$\therefore$ This proves that $P(A+B) = P(A) + P(B)$ (Proved)

3.1.16. State and prove the theorem of compound probability or conditional probability or multiplication theorem:

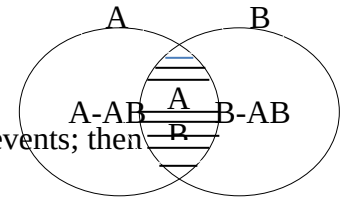
Statement: The probability of occurrence of the event A as well as B, is given by the product of unconditional probability of A and conditional probability. i.e.,

$$P(AB) = P(A) \cdot P(B/A)$$

This is known as multiplication theorem.
Proof: Let, a random experiment has n possible outcomes, which are mutually exclusive, exhaustive and equally likely out of n cases let m cases are favourable to an event A. i.e., $P(A) = \frac{m}{n}$
Out of m cases let m_1 cases be favourable to another event B also,
i.e., $P(AB) = \frac{m_1}{n}$ and $P\left(\frac{B}{A}\right) = \frac{m_1}{m}$
We know, $\frac{m_1}{n} = \frac{m}{n} \cdot \frac{m_1}{n} = P(A) \cdot P\left(\frac{B}{A}\right)$
$\therefore P(AB) = P(A) \cdot P\left(\frac{B}{A}\right)$ (Proved.)

$P(A) = \frac{m}{n}$ $P(B) = \frac{m_1}{n}$
$P(AB) = \frac{m}{n} \cdot \frac{m_1}{n} = \frac{m_1}{n}$

*** If $A\ and\ B$ be any two events corresponding to a random experiment E, then**
$P(A+B) = P(A) + P(B) - P(AB)$
$or, P(A \cup B) = P(A) + P(B) - P(AB)$
Proof:- From the given Venn diagram, it is clear that the events
A-AB, AB and B-AB, are mutually exclusive and also
$A = (A - AB) + AB$
$B = AB + (B - AB)\ \ and$
$A + B = (A - AB) + (AB) + (B - AB)$



Now since $(A - AB), AB\ and\ (B - AB)$ are mutually exclusive events; then
$P(A) = P(A - AB) + P(AB) \ldots\ldots\ldots\ldots\ldots\ldots (i)$
$P(B) = P(AB) + P(B - AB) \ldots\ldots\ldots\ldots\ldots\ldots (ii)$
and $P(A+B) = P(A - AB) + P(AB) + P(B - AB) \ldots\ldots\ldots\ldots\ldots\ldots (iii)$
From (i) and (ii)
$P(A - AB) = P(A) - P(AB) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (iv)$
$P(B - AB) = P(B) - P(AB) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (v)$
From (iii), (iv) and (v), we get
$P(A+B) = P(A) - P(\cancel{AB}) + P(\cancel{AB}) + P(B) - P(AB)$

4. If A, B, C be any three events, then

$P(A+B+C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(CA) + P(ABC)$

Proof:

$$\begin{aligned} P(A+B+C) &= P\{(A+B)+C\} \\ &= P(A+B) + P(C) - P\{(A+B)C\} \\ &= P(A) + P(B) - P(AB) + P(C) - P(AC+BC) \ [by\ distributive\ law] \\ &= P(A) + P(B) - P(AB) + P(C) - P(AC) - P(BC) + P(ACBC) \\ &= P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC) \end{aligned}$$

Subunit – 2: Descriptive Statistics – Measure of Central Tendency & Dispersion, Correlation

3.2.1. Measure of Central Tendency

3.2.1.1. Mean

In mathematics, **mean** has several different definitions depending on the context.

In probability and statistics, **mean** and expected value are used synonymously to refer to one measure of the central tendency either of a probability distribution or of the random variable characterized by that distribution. In the case of a discrete probability distribution of a random variable 'X', the mean is equal to the sum over every possible value weighted by the probability of that value; that is, it is computed by taking the product of each possible value x of X and its probability P(x), and then adding all these products together, giving $\mu = \sum xP(x)$.

An analogous formula applies to the case of a continuous probability distribution. Not every probability distribution has a defined mean; see the Cauchy distribution for an example. Moreover, for some distributions the mean is infinite: for example, when the probability of the value 2^n is $\frac{1}{2^n}$ for n = 1, 2, 3,...

3.2.1.2. Arithmetic Mean

- **Concept:**

The **arithmetic mean** (or **mean** or **average**) is the most commonly used and readily understood measure of central tendency. In statistics, the term average refers to any of the measures of central tendency. The arithmetic mean is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations.

- **Formula:**

1. **Simple Arithmetic Mean:** Symbolically, if we have a data set containing the values X_1, X_2, …, X_n, the arithmetic mean 'AM' is defined by the formula,

$AM = \frac{1}{n}\sum_{i=1}^{n} Xi$

2. **Weighted Arithmetic Mean:** If we have a data set containing the values X_1, X_2, …, X_n have frequencies f_1, f_2, …, f_n then, AM = $\frac{1}{N}\sum_{i=1}^{n}$ fi Xi ; where, N = $\sum_{i=1}^{n}$ fi

3. **Composite Mean:** If two groups contain n_1 and n_2 observations with means $\overline{x_1}$ and $\overline{x_2}$ respectively, then the mean ($\bar{x}$) of the composite group (n_1+ n_2) observations is given by the relation, N$\bar{x}$ = n_1 $\overline{x_1}$ + n_2 $\overline{x_2}$; where, N = n_1+ n_2

Characteristics	Groups		Composite Group
	I	II	
No. of observations	n_1	n_2	N
Mean	$\overline{x_1}$	$\overline{x_2}$	$\bar{x}$

- ## Calculation (Application):

1. Example 5.1: Calculation of Simple Arithmetic Mean:

Let us consider the monthly salary of 10 employees of a firm: 2500, 2700, 2400, 2300, 2550, 2650, 2750, 2450, 2600, 2400

The Simple Arithmetic Mean ($\bar{x}$) is

$$\frac{2500+2700+2400+2300+2550+2650+2750+2450+2600+2400}{10}=2530.$$

2. Example 5.2: Calculation of Weighted Arithmetic Mean:

Table - 5.1: Calculation for Weighted Arithmetic Mean

Price(Rs) per table	Number of table sold(f)	fx
36	14	504
40	11	440
44	9	396
48	6	288
Total	40	1628

Weighted Arithmetic Mean ($\bar{x}$) = $\frac{\Sigma fx}{N}$= 1628/40 = 40.70

Important Note: Although A.M. can be calculated directly, the calculations can be considerably simplified based on the following theory:

(i) If $y_1, y_2, \ldots, y_n$ represent the deviations of $x_1, x_2, \ldots, x_n$ from an arbitrary constant c, then Mean of x = c + Mean of y [i.e., if y = x – c, then $\bar{x} = c + \bar{y}$]

(ii) If $y_1, y_2, \ldots, y_n$ represent the deviations of $x_1, x_2, \ldots, x_n$ from an arbitrary constant c, in units of another constant d, then Mean of x = c + d (Mean of y)

i.e., if $y = \frac{x-c}{d}$ then, $\bar{x} = c + d\bar{y}$

3. Example 5.3:

Calculate the Arithmetic Mean from the following frequency distribution of earners by monthly income:

Income (Rs.):	Below 200	200-399	400-599	600-799	800-999	1000-1199
No. of earners:	25	72	47	22	13	7

Table - 5.2: Calculation for Weighted Arithmetic mean

Class Interval	Frequency	Mid Value (x)	$y = \frac{x-499.5}{200}$	f.y
0 – 199	25	99.5	-2	-50
200 – 399	72	299.5	-1	-72
400 – 599	47	499.5	0	0
600 – 799	22	699.5	1	22
800 – 999	13	899.5	2	26
1000 – 1199	7	1099.5	3	21
Total	186			-53

If y = (x – c) /d, then $\bar{x} = c + d\bar{y}$

Here, c = 499.5 and d = 200

Thus, $\bar{x} = 499.5 + 200 \left(\frac{-53}{186}\right)$ = 499.5 – 56.99 = 442.51Rs.

4. Example 5.4: Calculation of Composite Mean:

There are two branches of an establishment employing 100 and 80 persons respectively. If the arithmetic means of the monthly salaries paid by the two branches are Rs. 275 and Rs. 225 respectively, find the A.M. of the salaries of the employees of the establishment as a whole.

Table 5.3: Mean of Composite Group

Characteristics	Groups I	Groups II	Composite Group
No. of observations	$n_1 = 100$	$n_2 = 80$	N = 180
Mean salary (Rs.)	$\bar{x}_1 = 275$	$\bar{x}_2 = 225$	$\bar{x}$ = ?

$N\bar{x} = n_1 \bar{x}_1 + n_2 \bar{x}_2$

Or, $180\bar{x} = 100 * 275 + 80 * 225 = 45{,}500$

Or, $\bar{x} = 45500 / 180 = 252.78$ Rs.

- **Remarks:**

If the data set is a statistical population (i.e., consists of every possible observation and not just a subset of them), then the mean of that population is called the **population mean**.

On the other hand, if the data set is a statistical sample (a subset of the population), we call the statistic resulting from this calculation a **sample mean**.

The arithmetic mean of a variable is often denoted by a bar, for example as in $\bar{x}$ (read x bar), which is the mean of the n values $x_1, x_2, \ldots, x_n$.

- **Limitations of Arithmetic Mean:**

1. The strongest drawback of AM is that it is very much affected by extreme observations. Two or three very large values of the variable may unduly affect the value of AM.
2. AM cannot be used in the case of open end classes such as <10 or >70 etc.
3. It cannot be determined by inspection nor can it be located graphically.
4. AM cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitatively such as intelligence, honesty, beauty etc.
5. In extremely asymmetrical (skewed) distribution, usually AM is not representative of the distribution and hence is not a suitable measure of locations.

3.2.1.3. Geometric Mean

- **Concept:**

In mathematics, the **geometric mean** is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum).

- **Formula:**

The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers $\{x_i\}_{i=1}^{N}$, the geometric mean is defined as $\left(\prod_{i=1}^{N} x_i\right)^{1/N}$.

For instance, the geometric mean of two numbers, say 2 and 8, is just the square root of their product; that is, $\sqrt{2 \cdot 8} = 4$. As another example, the geometric mean of the three numbers 4, 1, and 1/32 is the cube root of their product (1/8), which is 1/2; that is, $\sqrt[3]{4 \cdot 1 \cdot 1/32} = 1/2$

- **Calculation (Application):**

Example 5.5:

Table 5.4: Calculation for Geometric Mean

Group	Group Index (x)	Weight(f)	log x	f *(log x)
A	118	4	2.0719	8.2876
B	120	1	2.0792	2.0792
C	97	2	1.9868	3.9736
D	107	6	2.0294	12.1764
E	111	5	2.0453	10.2265
F	93	2	1.9685	3.9370
Total	-	20	-	40.6803

Log GM = 1/N $\sum f_i$ (log x_i) = 40.6803/20 =2.0340

GM= antilog 2.0340 = 108.1

G denotes geometric mean.

- **Remarks:**

A geometric mean is often used when comparing different items – finding a single "figure of merit" for these items – when each item has multiple properties that have different numeric ranges. For example, the geometric mean can give a meaningful "average" to compare two companies which are each rated at 0 to 5 for their environmental sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean were used instead of a geometric mean, the financial viability is given more weight because its numeric range is

larger—so a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5). The use of a geometric mean "normalizes" the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

- **Limitations of Geometric Mean:**

1. If any one of the observations is zero and if any one of the observations is negative, GM becomes imaginary regardless of the magnitude of other items.
2. It cannot be calculated if some of values in a series are not known.

3.2.1.4. Harmonic Mean

- **Concept:**

In mathematics, the **harmonic mean** (sometimes called the **sub-contrary mean**) is one of several kinds of average, and in particular one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired.

- **Formula:**

The harmonic mean H of the positive real numbers $x_1, x_2, ..., x_n$ is defined to be

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n \cdot \prod_{j=1}^{n} x_j}{\sum_{i=1}^{n} \frac{\prod_{j=1}^{n} x_j}{x_i}}.$$

From the third formula in the above equation, it is more apparent that the harmonic mean is related to the arithmetic and geometric means.

- **Calculation (Application):**

Example 5.6:

Equivalently, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. As a simple example, the harmonic mean of 1, 2, and 4 is

$$\frac{3}{\frac{1}{1} + \frac{1}{2} + \frac{1}{4}} = \frac{1}{\frac{1}{3}\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{4}\right)} = \frac{12}{7}.$$

- **Remarks:**

The Harmonic mean is a Schur-concave function, and dominated by the minimum of its arguments, in the sense that for any positive set of arguments, $\min(x_1 \ldots x_n) \leq H(x_1 \ldots x_n) \leq n \min(x_1 \ldots x_n)$.Thus, the harmonic mean cannot be made arbitrarily large by adding more big values to the argument set. The harmonic mean is the reciprocal dual of the arithmetic mean for positive inputs:

$$1/H(1/x_1 \ldots 1/x_n) = A(x_1 \ldots x_n)$$

- **Limitations of Harmonic Mean:**

1. The value of harmonic mean cannot be obtained if anyone of the observations is zero.
2. It cannot be calculated if some of the values in a series are unknown.

3.2.2. Median

- **Concept:**

In statistics and probability theory, the **median** is the number separating the higher half of a data sample, a population, or a probability distribution, from the lower half. If each group contains less than half the population, then some of the population is exactly equal to the median. For example, if a < b < c, then the median of the list {a, b, c} is b, and, if a < b < c < d, then the median of the list {a, b, c, d} is the mean of b and c; i.e., it is (b + c)/2.

- **Formula:**

In terms of notation, some authors represent the median of a variable x either as $\tilde{x}$ or as $\mu_{1/2}$, sometimes also M. There is no widely accepted standard notation for the median, so the use of these or other symbols for the median needs to be explicitly defined when they are introduced.

The median is the 2nd quartile, 5th decile, and 50th percentile.

Median (M_d) = $l_1 + \{(N/2) – F\}/f_m \times c$

Where,

l_1 = lower boundary of the median class

N = total frequency

F = cumulative frequency corresponding to l_1

f_m = frequency of the median class

c = width of the median class

- ## Calculation (Application):

1. Example 5.7:

The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one [e.g., the median of (3, 3, 5, 9, 11) is 5]. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values [the median of {3, 5, 7, 9} is (5 + 7) / 2 = 6], which corresponds to interpreting the median as the fully trimmed mid-range. The median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data is contaminated, the median will not give an arbitrarily large result. A median is only defined on ordered one-dimensional data, and is independent of any distance metric.

2. Example 5.8:

Find the median and the median class of the data given below:

Class boundaries:	15-25	25-35	35-45	45-55	55-65	65-75
Frequency :	4	11	19	14	0	2

 # Method 1:

For a grouped frequency distribution,

Median = Value of the variable corresponding to cumulative frequency N/2

Table 5.5: Cumulative Frequency Distribution

Class Boundary	Cumulative Frequency (Less than)
15	0
25	4
35	15
Median	N / 2 = 25
45	34
55	48
65	48
75	50 = N

Since N / 2 = 25 lies between the cumulative frequencies 15 and 34, the corresponding value of the variable, viz., median, must lie in the interval between 35 and 45. The median class is, therefore, (35 – 45).

Now applying simple interpolation

$\frac{\text{Median}-35}{45-35} = \frac{25-15}{34-15}$

Or, $\frac{\text{Median}-35}{10} = \frac{10}{19}$

Or, Median – 35 = $\frac{10}{19} * 10 = \frac{100}{19} = 5.26$

Or, Median = 35 + 5.26 = 40.26

 # Method 2: We first calculate the cumulative frequencies against each class interval as shown below:

Table 5.6: Calculation of Cumulative Frequency

Class Boundaries	Frequency	Cumulative Frequency
15-25	4	4
25-35	11	15 = F
(35-45)	19 = f_m	34
45-55	14	48
55-65	0	48
65-75	2	50 = N

From the last column of table 2.7 it is seen that, N/2 = 25 is more than cumulative frequency 15, but is less than the next cumulative frequency 34; hence the median class is (35 – 45). Hence l_1 = 35, N = 50, F = 15, f_m =19, c = 45 – 35 = 10.

Median = 35 + $\frac{25-15}{19}$ * 10 = 35 + 5.26 = 40.26

- **Remarks:**

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the median (in the case of an even sample size); if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Nonetheless, the value of the median is uniquely determined with the usual definition. At most, half the population has values strictly less than the median, and, at most, half have values strictly greater than the median.

The median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers, e.g., because they may be measurement errors.

- **Limitations of Median:**

1. In case of even number of observations for an ungrouped data, median cannot be determined exactly. We merely estimate it as the arithmetic mean of the two middle terms.
2. Median, being a positional average, is not based on each and every item of the distribution. It depends on all the observations only to the extent whether they are smaller than or greater than it; the exact magnitude of the observations being immaterial.
3. It does not lead itself to algebraic treatment.
4. It is affected by fluctuations of items.
5. It is less stable measure of central tendency than mean.
6. In the case of continuous series, it cannot be calculated exactly.

3.2.3. Mode

- **Concept:**

The mode is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled. The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so the mode is at the peak.

- **Formula:**

In symmetric uni-modal distributions, such as the normal distribution, the mean (if defined), median and mode all coincide. For samples, if it is known that they are drawn from a symmetric distribution, the sample mean can be used as an estimate of the population mode.

Mode = $l_1 + \{d_1/ (d_1+d_2)\} \times c$

Where,

l_1 = lower boundary of the modal class

d_1 = difference of frequencies in the modal class and the preceding class

d_2 = difference of frequencies in the modal class and the following class

c = common width of the classes.

- **Calculation (Application):**

Example 5.9:

Let the monthly profits in rupees of 100 shops are distributed as follows:

Profits per shop (Rs.):	0-100	100-200	200-300	300-400	400-500	500-600
No of shops :	12	18	27	20	17	6

In order to calculate the Mode directly, we find that the largest class frequency, viz., 27, lies in the class 200-300 and hence this is the modal class.

Therefore, in formula, $l_1 = 200$, $d_1 = 27-18= 9$, $d_2 = 27-20 = 7$, $c = 300-200 = 100$.

Mode = $200 + \frac{9}{9+7} \times 100 = 256.25$

- **Remarks:**

Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population. The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions.

The mode is not necessarily unique, since the probability mass function or probability density function may take the same maximum value at several points x_1, x_2, etc. The most extreme case occurs in uniform distributions, where all values occur equally frequently. When a probability density function has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution. Such a continuous distribution is called multimodal (as opposed to uni-modal).

- **Limitations of Mode:**

1. Since mode is the value of X corresponding to the maximum frequency, it is not based on all the observations of the series.
2. Mode is not suitable for further mathematical treatment.
3. As compared with mean, mode is affected to a greater extent by the fluctuations of sampling.

3.2.3. Measure of Dispersion

In statistics, **dispersion** (also called **variability**, **scatter**, or **spread**) denotes how stretched or squeezed a distribution (theoretical or that underlying a statistical sample) is. Common examples of absolute measures of statistical dispersion are the range, quartile deviation, mean deviation, standard deviation and variance.

3.2.3.1. Range (Absolute measure of Dispersion)

- **Concept:**

The range is defined as the difference between the highest and the lowest values in the series. This is the simplest method of measuring dispersion.

- **Formula:**

Range = (H-L) ; where, H = Highest value and L = Lowest value

- **Calculation (Application):**

Example 5.10: The weights of 11 forty-year old men were 148, 154, 158, 160, 161, 162, 166, 170, 182, 195 and 236 pounds.

In the given data, maximum value = 236 and minimum value = 148.

Hence,

Range = 236 – 148 = 88 pounds.

- **Remarks:**

It is easy to understand. Range is the difference between the two extreme values and as such represents the maximum possible difference between any two observations.

- **Limitations of Range:**

1. It is not based on the entire set of data.
2. It is very much affected by fluctuations of sampling. Its value varies widely from sample to sample.
3. It is not possible to find out the range in open-end frequency distribution.
4. It does not present the accurate picture of the series.
5. It is affected by extreme values.

3.2.3.2. Co-efficient of Range (Relative measure of Dispersion)

- **Concept:**

Range as defined above is an absolute measure of dispersion and depends on the units of measurement. Thus if we want to compare the variability of two or more distributions with the same units of measurement, we may use the above formula. However, to compare the variability of the distribution given in different units of measurement we cannot use the above formula but we need a relative measure which is independent of the units of measurement.

- **Formula:**

This relative measure, called the coefficient of range, is defined as follows,

Coefficient of Range = $\frac{H-L}{H+L}$

- **Calculation (Application):**

Example 5.11:

Continuing the above example 8.1, we have in the given data, maximum value = 236 and minimum value = 148.

Hence,

Coefficient of Range = $\frac{H-L}{H+L} = \frac{236-148}{236+148} = \frac{88}{384} = 0.2292$

3.2.3.3. Quartile Deviation (Absolute measure of Dispersion)

- **Concept:**

Quartile Deviation is defined as half the difference between the upper and the lower quartiles.

- **Formula:**

Quartile deviation (Q.D.) is obtained by dividing the difference of the third quartile and the first quartile ($Q_3 - Q_1$) by 2.

Thus, Quartile Deviation = $\frac{Q3 - Q1}{2}$

Where, Q_3 = Third Quartile or Upper Quartile and Q_1 = First Quartile or Lower Quartile

- **Calculation (Application):**

1. **Example 5.12:**

Calculate the quartile deviation from the following:

Class interval:	10-15	15-20	20-25	25-30	30-40	40-50	50-60	60-70	Total
Frequency :	4	12	16	22	10	8	6	4	82

To calculate Quartile Deviation, we have to find Q_1 (first quartile) and Q_3 (third quartile), i.e., values of the variable corresponding to cumulative frequencies N/4 and 3N/4.

Here, total frequency = 82. Therefore, N/4 = 20.5 and 3N/4 = 61.5

Table 5.7: Cumulative Frequency Distribution

Class Boundary	Cumulative Frequency (Less than)
10	0
15	4
20	16
Q_1	N/4 = 20.5
25	32
30	54
Q_3	3N/4 = 61.5
40	64
50	72
60	78
70	82 = N

Applying simple interpolation,

$\frac{Q1 - 20}{25 - 20} = \frac{20.5 - 16}{32 - 16}$ Or, $\frac{Q1 - 20}{5} = \frac{4.5}{16}$

Or, $Q_1 - 20 = \frac{4.5}{16} * 5 = 1.4$

Or, $Q_1 = 20 + 1.4 = 21.4$

Similarly,

$\frac{Q3 - 30}{40-30} = \frac{61.5 - 54}{64-54}$ Or, $\frac{Q3 - 30}{10} = \frac{7.5}{10}$

Or, $Q_3 - 30 = \frac{7.5}{10} * 10 = 7.5$

Or, $Q_3 = 30 + 7.5 = 37.5$

Therefore,

Quartile Deviation = $\frac{Q3 - Q1}{2} = \frac{37.5 - 21.4}{2} = 8.05$

2. Example 5.13:

Calculate the quartile deviation:

Wage per week (Rs.)	No. of wage earners
Less than 35	14
35-37	62
38-40	99
41-43	18
Over 43	7

To calculate quartile deviation at first, we have to determine the quartiles Q_1 and Q_3, which can be done from a cumulative frequency distribution using simple interpolation.

Table 5.8: Cumulative Frequency Distribution

Wages per week (Rs)	Cumulative frequency
34.5	14
Q_1	N/4 = 50
37.5	76
Q_3	3N/4 = 150
40.5	175
43.5	193
…	200 = N

Applying simple interpolation,

$(Q_1 – 34.5) / (37.5 – 34.5) = (50 – 14) / (76 – 14)$

$(Q_3 – 37.5) / (40.5 – 37.5) = (150 – 76) / (175 – 76)$

Solving, $Q_1 = 36.24$ and $Q_3 = 39.74$

Therefore, Quartile Deviation = (39.74 – 36.24) / 2 = 1.75 Rs.

- **Remarks:**

The quartile deviation gives the average amount by which the two quartiles differ from median. For symmetric distribution we have,

$(Q_3 – M_d) = (M_d - Q_1)$

Or, $M_d = \frac{Q3+Q1}{2}$

i.e., median lies half way on the scale from Q_1 to Q_3. Thus, for a symmetric distribution we have,

Q.D. + $Q_1 = \frac{Q3 - Q1}{2} + Q_1 = \frac{Q3+Q1}{2} = M_d$

And, Q_3 - Q.D. = $Q_3 - \frac{Q3 - Q1}{2} = \frac{Q3+Q1}{2} = M_d$

In other words, for a symmetric distribution we have,

$Q_1 = (M_d – Q.D.)$

And, $Q_3 = (M_d + Q.D.)$

Since in a distribution 25% of the observations lie between Q_1 and 25% observations lie above Q_3, 50% of the observations lie between Q_1 and Q_3. Therefore, we can conclude that for a symmetric distribution,

$M_d \pm$ Q.D. covers exactly 50% of the observations.

- **Limitations of Quartile Deviation:**

1. Q.D. is not based on all the observations since it ignores 25% of the data at the lower end and 25% data at the upper end of the distribution, it cannot be regarded as a reliable measure of variability.
2. Q.D. is affected considerably by fluctuations of sampling.
3. Q.D. is not suitable for further mathematical treatment.

3.2.3.4. Co-efficient of Quartile Deviation (Relative measure of Dispersion)

- **Concept:**

Quartile deviation as defined above is only an absolute measure of dispersion. For comparative studies of variability of two distributions we need a relative measure which is known as Coefficient of Quartile Deviation.

- **Formula:**

Coefficient of Quartile Deviation is given by the following formula,

Coefficient of Quartile Deviation = $\frac{Q3 - Q1}{Q3 + Q1}$

- **Calculation (Application):**

Example 5.14:

Continuing the above example 10.1, we have in the given information; Q_1 = 21.4 and Q_3 = 37.5.

Hence,

Coefficient of Quartile Deviation = $\frac{Q3 - Q1}{Q3 + Q1} = \frac{37.5 - 21.4}{37.5 + 21.4} = \frac{16.1}{58.9}$ = 0.2733

3.2.3.5. Mean Deviation (Absolute measure of Dispersion)

- **Concept:**

Mean deviation is defined as the arithmetic average of the deviations of various items from a measure of central tendency, may it be, mean, median or mode. Generally mean deviation is calculated either from mean or from median. Mode is not considered as its value is indeterminate. Between mean and median, the median is supposed to be better than mean because the sum of the deviations from the median is less than the sum of the deviations from the mean. It is also known as first moment of dispersion.

- **Formula:**

If $x_1, x_2, \ldots, x_n$ are n given observations then the mean deviation (M>D>) about an average A, say, is given by M.D. = $\frac{1}{n}\sum|x-A| = \frac{1}{n}\sum|D|$

Where, |D| = |x-A| read as mod (x-A) is the modulus value or absolute value of the deviation (after ignoring the negative sign).

1. **Mean Deviation about an average A:**

In case of frequency distribution or grouped or continuous frequency distribution, mean deviation about an average A is given by,

M.D. = $\frac{1}{N}\sum f\,|x-A|$

Where, x = value of the variable or mid-value of the class interval, N = $\sum f$

Usually, we obtain the mean deviation (M.D.) about any one of the three averages mean ($\bar{x}$), median (M_d) or mode (M_0).

2. **Mean Deviation about Mean:** = $\frac{1}{N}\sum f\,|x-\bar{x}|$
3. **Mean Deviation about Mean:** = $\frac{1}{N}\sum f\,|x-M_d|$
4. **Mean Deviation about Mode:** = $\frac{1}{N}\sum f\,|x-M_0|$

- ## Calculation (Application):

1. **Example 5.15:**

Mean Deviation about Mean

Find the mean deviation of the following series:

X : 10 11 12 13 14 Total

Frequency: 3 12 18 12 3 48

Table 5.9: Calculations for Mean Deviation about Mean

x	f	f.x	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
Total	48	576	-	36

$\bar{x} = \sum \frac{fx}{N} = \frac{576}{48} = 12$

Thus, Mean Deviation about Mean = $\frac{1}{N}\sum f\,|x-\bar{x}| = \frac{36}{48} = 0.75$

2. Example 5.16:

Mean Deviation about Median

Calculate the Mean Deviation of the following values about the median: 8, 15, 53, 49, 19, 62, 7, 15, 95, 77.

Since there are an even number of observations viz., 10, the median is the average of the two middlemost observations, when arranged in order of magnitude: 7, 8, 15, 15, (19, 49), 53, 62, 77, 95. Thus, Median = (19 + 49)/2 = 34

Table 5.10: Calculations for Mean Deviation about Median

x	|x - Median|
8	26
15	19
53	19
49	15
19	15
62	28
7	27
15	19
95	61
77	43
Total	272

Mean Deviation about Median $= \frac{1}{n} \sum f |x\text{-Median}|$

$= \frac{1}{10} \times 272$

$= 27.2$

- **Remarks:**

1. The sum of absolute deviations (after ignoring the sign) of a given set of observations is minimum when taken about median. Hence mean deviation is minimum when it is calculated from median.
2. For symmetric distribution the range Mean ± M.D. (about mean) or M_d ± M.D. (about median), [as M = M_d for a symmetric distribution] covers 57.5% of the observations of the distribution.

- **Limitations of Mean Deviation:**

1. The strongest objection against mean deviation is that while computing its value we take the absolute value of the deviations about an average and ignore the sings of the deviations.
2. It is not a satisfactory measure when taken about mode or while dealing with a fairly skewed distribution.
3. It is rarely used in sociological studies.
4. It cannot be computed for distribution with open end classes.
5. It tends to increase with the size of the sample through not proportionately and not so rapidly as range.

3.2.3.6. Co-efficient of Mean Deviation (Relative measure of Dispersion)

- **Concept:**

A relative measure of dispersion, called the coefficient of mean deviation is pure number independent of the units of measurement and is useful for comparing the variability of different distributions.

- **Formula:**

1. **Coefficient of M.D.** = $\frac{\textbf{Mean Deviation}}{\textbf{Average about which it is calculated}}$
2. **Coefficient of M.D. about mean** = $\frac{\textbf{Mean Deviation about Mean}}{\textbf{Mean}}$
3. **Coefficient of M.D. about median** = $\frac{\textbf{Mean Deviation about Median}}{\textbf{Median}}$

- **Calculation (Application):**

1. **Example 5.17: Coefficient of M.D. about mean**

Continuing the above example 12.1, we have in the given information; M.D. about mean = 0.75 and mean ($\bar{x}$) = 12.

Thus, Coefficient of M.D. about mean = $\frac{\text{Mean Deviation about Mean}}{\text{Mean}} = \frac{0.75}{12} = 0.0625$

2. **Example 5.18: Coefficient of M.D. about median**

Continuing the above example 12.2, we have in the given information; M.D. about median = 27.2 and median = 34.

Thus, Coefficient of M.D. about median = $\frac{\text{Mean Deviation about Median}}{\text{Median}} = \frac{27.2}{34} = 0.80$

3.2.3.7. Standard Deviation (Absolute measure of Dispersion)

- **Concept:**

Standard deviation (S.D.), usually denoted by the Greek letter σ (sigma) was first suggested by Karl Pearson as a measure of dispersion in 1983. It is defined as the positive square of the deviations of the given observations from their arithmetic mean.

- **Formula:**

Case 1: For Large Sample

 # Formula 1: Standard Deviation for Simple Series:

If $x_1, x_2, \ldots, x_n$ is a set of n observations then its standard deviation is given by,

$\sigma = \frac{1}{n}\sqrt{\sum(x-\bar{x})^2}$; where, $\bar{x} = \frac{\sum x}{n}$, arithmetic mean of given values

 # Formula 2: Standard Deviation for Frequency Distribution:

In case of frequency distribution, the standard deviation is given by,

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

Case 2: For Small Sample

If the observations are small, S.D. can be calculated by using the following relations:

 # Formula 1: Standard Deviation for Simple Series:

$\sigma = \sqrt{[\frac{\sum x2}{n} - (\frac{\sum x}{n})^2]}$

 # Formula 2: Standard Deviation for Frequency Distribution:

$\sigma = \sqrt{[\frac{\sum fx2}{N} - (\frac{\sum fx}{N})^2]}$

Case 3: S.D for Composite Group

If two groups contain n_1 and n_2 observations with means $\bar{x_1}$ and $\bar{x_2}$ and standard deviation σ_1 and σ_2 respectively, then the S.D (σ) of the composite group is given by the relation,

$N\sigma^2 = (n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1 d_1^2 + n_2 d_2^2)$; where, $d_1 = (\bar{x_1} - \bar{x})$, $d_1 = (\bar{x_1} - \bar{x})$, and $\bar{x}$ is the mean of composite group.

Or, $\sigma^2 = \{(n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1 d_1^2 + n_2 d_2^2)\} / N$

Or, $\sigma = \sqrt{[\{(n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1 d_1^2 + n_2 d_2^2)\} / N]}$

Characteristics	Groups		Composite Group
	I	II	
No. of observations	n_1	n_2	N
Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}$
Standard Deviation	σ_1	σ_2	σ

- **Calculation (Application):**

1. Example 5.19:

Calculate the standard deviation from the following series: 20 85 120 60 40

Table 5.11: Calculation of S.D.

x_i	x_i^2
20	400
85	7225
120	14400
60	3600
40	1600
$\sum x_i = 325$	$\sum x_i^2 = 27225$

Here, n = 5

Thus, Standard Deviation (σ) = $\sqrt{[\frac{\sum x2}{n} - (\frac{\sum x}{n})^2]}$

= $\sqrt{[\frac{27225}{5} - (\frac{325}{5})^2]}$

= $\sqrt{\{5445 - (65)^2\}}$

= $\sqrt{(5445 - 4225)}$

= $\sqrt{1220}$

= 34.92

2. Example 5.20:

Calculate the standard deviation of the following data:

Class Limit:	90-99	80-89	70-79	60-69	50-59	40-49	30-39
Frequency:	2	12	22	20	14	4	1

Table 5.12: Calculation of S.D.

Class Interval	Mid-value (x_i)	Frequency (f_i)	x_i^2	$f_i x_i$	$f_i x_i^2$
90-99	94.5	2	8930.25	189	17860.5
80-89	84.5	12	7140.25	1014	85683
70-79	74.5	22	5550.25	1639	122105.5
60-69	64.5	20	4160.25	1290	83205
50-59	54.5	14	2970.25	763	41583.5
40-49	44.5	4	1980.25	178	7921
30-39	34.5	1	1190.25	34.5	1190.25
Total		75		5107.5	359548.75

By formula,

S.D (σ) = $\sqrt{[\frac{\sum fx2}{N} - (\frac{\sum fx}{N})^2]}$

= $\sqrt{[\frac{359548.75}{75} - (\frac{5107.5}{75})^2]}$

= $\sqrt{}$(4793.9833 – 4637.61

= $\sqrt{}$(156.3733

= 12.50

3. Example 5.21: Calculation of Composite Mean:

For a group of 50 boys the mean score and the standard deviation of scores on a test are 59.5 and 8.38 and also for a group of 40 girls the same results are 54 and 8.23 with a combined mean 57.06. Calculate the standard deviation of combined group of 90 children.

Table 5.13: Standard Deviation of Composite Group

Characteristics	Groups		Composite Group
	I	II	
No. of observations	$n_1 = 50$	$n_2 = 40$	N = 90
Mean	$\bar{x}_1 = 59.5$	$\bar{x}_2 = 54$	$\bar{x} = 57.06$
Standard Deviation	$\sigma_1 = 8.38$	$\sigma_2 = 8.23$	σ = ?

Here, $d_1 = 59.5 – 57.06 = 2.44$, $d_2 = 54 – 57.06 = -3.06$

Thus Composite S.D. $(N\sigma^2) = (n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1 d_1^2 + n_2 d_2^2)$

$= 90\sigma^2 = \{50*(8.38)^2 + 40*(8.23)^2\} + \{50*(2.44)^2 + 40*(-3.06)^2\}$

$= (3511.11 + 2709.32) + (297.68 + 374.54)$

$= 6892.76$

Thus,

$\sigma^2 = \frac{6892.76}{90} = 76.59$

$\sigma = \sqrt{76.59} = 8.75$

- **Remarks:**

1. It may be pointed out that although mean deviation could be calculated about any one of the averages (mean, median or mode), standard deviation is always computed about arithmetic mean.
2. The value of the S.D. depends on the numerical value of the deviations $(x_1-\bar{x})$, $(x_2-\bar{x})$, …, $(x_n-\bar{x})$. Thus the value of σ will be greater if the values of x are scattered widely away from the mean. Thus, a small of σ imply that the distribution is homogeneous whereas, the large value of σ imply that it is heterogeneous. In particular, S.D. is zero if each of the deviations is zero i.e., $\sigma = 0$ if and only if,

   $(x_1-\bar{x}) = 0$, $(x_2-\bar{x}) = 0$, …, $(x_n-\bar{x}) = 0$

   or, $x_1 = x_2 = \ldots = x_n = \bar{x}$

- **Limitations of Standard Deviation:**

1. It is not easily computed.
2. It gives more weight to extreme items, because the squares of the deviations which are large in size due to the extremity of items would be still greater.
3. S.D. is also not easy to interpret. It possesses most of the characteristics of an ideal measure of dispersion. Hence it is regarded as the best measure of dispersion.

3.2.3.8. Co-efficient of Variation (Relative measure of Dispersion)

- **Concept and Formula:**

Standard deviation, an absolute measure of dispersion, depends upon the units of measurement. The relative measure of dispersion based on standard deviation is called the coefficient of standard deviation and is given by,

Coefficient of Standard Deviation = $\frac{6}{\bar{X}}$

This is a pure number independent of the units of measurement and thus, is suitable for comparing the variability, homogeneity or uniformity of two or more distributions.

100 times the coefficient of dispersion based on standard deviation is called the coefficient of variation, abbreviated as C.V.

Thus,

Coefficient of Variation (C.V.) = $\frac{6}{\bar{X}} \times 100$

- **Calculation (Application):**

Example 5.22:

From some financial statistics, it is found that the monthly average Electricity Charges is Rs. 2460 and S.D. Rs. 120. The monthly average Direct Wages is Rs. 42000 and S.D. Rs. 1200. State which is the more variable with proper reasons.

The given data are (in Rs.)

Table 5.14: Calculation of Coefficient of Variation

	Electricity Charges	Direct Wages
Mean	2460	42000
S.D.	120	1200

Since the means in the two cases differ widely, although given in the same unit, viz., Rs., it will be inappropriate to employ S.D. for comparing the variability. The coefficient of variation (C.V.) will be more logical here.

C.V. (for Electricity Charges) = $\frac{\text{Rs.}120}{\text{Rs.}2400} * 100 = 4.9$

C.V. (for Direct Wages) = $\frac{\text{Rs.}1200}{\text{Rs.}42000} * 100 = 2.9$

Since the coefficient of variation for Electricity Charges is larger, Electricity Charges are more variable than Direct Wages.

Example 5.23:

The scores of two batsmen, P and Q, in ten innings during a certain season, are as under:

P	32	28	47	63	71	39	10	60	96	14
Q	19	31	48	53	67	90	10	62	40	80

Find which batsman is more consistent in scoring.

Here, we use Coefficient of Variation for measuring dispersion. The batsman with the smaller dispersion is more consistent. The mean and S.D for this series can be calculated by following way:

Table 5.15: Calculations for Mean and S.D

Batsman P			Batsman Q		
x	y = x - 50	y^2	x	y = x - 50	y^2
32	-18	324	19	-31	961
28	-22	484	31	-19	361
47	-3	9	48	-2	4
63	13	169	53	3	9
71	21	441	67	17	289
39	-11	121	90	40	1600
10	-40	1600	10	-40	1600
60	10	100	62	12	144
96	46	2116	40	-10	100
14	-36	1296	80	30	900
Total	-40	6660		0	5968

For Batsman P:

Mean score = 50 + (-40/10) = 46

S.D. = $\sqrt{\{\frac{6660}{10} - (\frac{-40}{10})^2\}} = \sqrt{650} = 25.5$

C.$V_P = \frac{25.5}{46} * 100 = 55$

For Batsman Q:

Mean score = 50 + (0/10) = 50

S.D. = $\sqrt{\{\frac{5968}{10} - (\frac{0}{10})^2\}} = \sqrt{596.8} = 24.4$

C.$V_Q = \frac{24.4}{50} * 100 = 49$

Since for Batsman Q, coefficient of variation is smaller, he is more consistent.

- **Remarks:**

For comparing the variability of two distributions we compute the 'coefficient of variation' for each distribution. A distribution with smaller C.V. is said to be more homogeneous or uniform or less variable than the other and the series with greater C.V. is said to be more heterogeneous or more variable than the other.

3.2.3.9. Variance

- **Concept:**

Variance is the square of the value of standard deviation and is denoted by σ^2.

- **Formula:**

Case 1: For Large Sample

 # Formula 1: Variance for Simple Series:

If $x_1, x_2, \ldots, x_n$ is a set of n observations then its variance is given by,

$\sigma^2 = \frac{1}{n}\sum(x-\bar{x})^2$; where, $\bar{x} = \frac{\sum x}{n}$, arithmetic mean of given values

 # Formula 2: Variance for Frequency Distribution:

In case of frequency distribution, the variance is given by,

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

Case 2: For Small Sample

If the observations are small, S.D. can be calculated by using the following relations:

 # Formula 1: Variance for Simple Series:

$\sigma^2 = \frac{\sum x2}{n} - (\frac{\sum x}{n})^2$

 # Formula 2: Variance for Frequency Distribution:

$\sigma^2 = \frac{\sum fx2}{N} - (\frac{\sum fx}{N})^2$

3.2.3.9. Skewness and Kurtosis

Skewness:

By skewness of a frequency distribution we mean the degree of its departure from symmetry.

A distribution which is not symmetrical is called asymmetrical distribution or skewed distribution. The skewness is said to be positive if the longer tail of the distribution is towards the higher values of the variable and it is negative if the longer tail of the distribution is towards the lower values of the variable.

Other important features are:

(i) In a symmetrical distribution, the mean, median and mode coincide.

(ii) If the distribution is positively skewed, then mean > median > mode.

(iii)If the distribution is negatively skewed, then mean < median < mode.

For our above example the skewness is equal to -1.25068 i.e., the distribution is negatively skewed. Also it is seen that mean (24) < median (26) < mode (27).

Kurtosis:

Another method of describing a frequency distribution is to specify its degree of peakedness or kurtosis. The importance of describing kurtosis is that, two distributions may have the same mean and the same standard deviation and may be equally skew, but one of them may be more peaked than the other.

A positive value of kurtosis indicates that the distribution has high concentration of values near the central tendency and has high tails. In the same way, a negative value of kurtosis indicates that the distribution has low concentration of values in the neighbourhood of the central tendency and has low tails.

Other important features are:

(i) A normal curve is said to be mesokurtic.

(ii) A distribution with positive kurtosis is called leptokurtic.

(iii)A distribution with negative kurtosis is called platykurtic.

3.2.3.10. Correlation and Regression Analysis

3.2.3.10.1. Correlation Analysis

- **Concept:**

The study of relationship between two or more variables is the most important in real life. It is a common knowledge that changes in a variable will affect the other. For example, when the price of the commodity increases, its demand decreases and on the other hand, when the price of a commodity decreases, its demand increases. The statistical deice with the help of which relationships between two or more than two variables are studied is called correlation.

- **Utility of the Study of Correlation:**

(i) With the help of the correlation we can measure the degree of relationship existing between the variables.

(ii) Correlation analyses contribute to the economic behavior, aids in locating the critically important variable and on which other depend.

- ## Types of Correlation:

A. Positive Correlation:

If both the variables move in the same direction, correlation is said to be positive. Some examples of series of positive correlations are:

a. Heights and weights

b. Family income and expenditure on luxury items

c. Price and supply of a commodity and so on

B. Negative Correlation:

If both the variables move in opposite direction, correlation is said to be negative. Some examples of series of negative correlations are:

a. Price and demand of a commodity

b. Volume and pressure of a perfect goes

c. Sale of woolen garments and the day temperature, and so on

C. Linear and Non-linear Correlation:

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values.

For example, let us consider the following Data:

X	1	2	3	4	5
Y	5	7	9	11	13

There are a unit change in the values of X, there is a constant change in the corresponding values of Y (2 units). Mathematically, above data can be expressed by the equation;

$Y = 2X + 3$

In general two variables X and Y are said to be linearly related, if these exists a relationship of the form

$Y = a + bX$

(1)

The relationship between two variables is said to be non-linear of curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but fluctuating rate. In such cases if the data are plotted on the XY-plane we do not get a straight-line curve mathematically speaking, the correlation is said to be non-linear if the slope of the plotted curve is not constant. Such phenomena are common in the data relating to economics and social sciences.

3.2.3.10.2. Correlation and Dependence

- **Concept:**

In statistics, **dependence** is any statistical relationship between two random variables or two sets of data. **Correlation** refers to any of a broad class of statistical relationships involving dependence. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship (i.e., correlation does not imply causation).

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several correlation coefficients, often denoted by **'r'**, measuring the degree of correlation. The most common of this is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation – that is, more sensitive to nonlinear relationships. Mutual information can also be applied to measure dependence between two variables.

3.2.3.10.3. Covariance

- **Concept:**

In probability theory and statistics, **covariance** is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

- **Formula:**

A distinction must be made between (1) the covariance of two random variables, which is a population parameter that can be seen as a property of the joint probability distribution, and (2) the sample covariance, which serves as an estimated value of the parameter.

Let (x_1, y_1), (x_2, y_2),……,(x_n, y_n) be a given set of observations on two variables x and y, the Covariance of x and y, usually represented cov(x,y) and is defined as

$$cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

Expanding the expression on the right, it can be shown that

$$cov(x,y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$$

This form is generally used for calculations.

- **Calculation (Application):**

Example 5.24:

Calculate the covariance between two variables

x:	1	2	3	4	5
y:	5	7	9	11	13

Table: 5.16: Calculations for covariance

x	y	xy
1	5	5
2	7	14
3	9	27
4	11	44
5	13	65
$\sum x = 15$	$\sum y = 45$	$\sum xy = 155$

The covariance of x and y is

$$cov(x,y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$$

$$= \frac{155}{5} - \left(\frac{15}{5}\right)\left(\frac{45}{5}\right)$$

$$= 31 - 3*9$$

$$= 31 - 27$$

$$= 4$$

- **Remarks:**

1. Cov (x, x) = Var (x)
2. Cov (x, y) = Cov (y, x) i.e., symmetric
3. Covariance is unaffected by change of origin.

3.2.3.10.4. Correlation Coefficient (r)

- **Concept:**

A mathematical method for measuring the intensity or the magnitude of linear relationship between two variables was suggested by Karl Pearson, a great British Biometrician and Statistician, and is by far the most widely used method in practice.

A **correlation coefficient** is a coefficient that illustrates a quantitative measure of some type of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values.

- **Types of correlation coefficients:**

(i) Pearson product-moment correlation coefficient:

It is a measure of the strength and direction of the linear relationship between two variables that is defined as the (sample) covariance of the variables divided by the product of their (sample) standard deviations. It is also known as r, R, or Pearson's r.

(ii) Intra-class correlation:

A descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups; describes how strongly units in the same group resemble each other.

(iii) Rank correlation:

The study of relationships between rankings of different variables or different rankings of the same variable

(a) Spearman's rank correlation coefficient, a measure of how well the relationship between two variables can be described by a monotonic function.

(b) Kendall tau rank correlation coefficient, a measure of the portion of ranks that match between two data sets.

(c) Goodman and Kruskal's gamma, a measure of the strength of association of the cross tabulated data when both variables are measured at the ordinal level.

- **Formula:**

Generally correlation coefficient is denoted by **r**.

Let $(x_1, y_1), (x_2, y_2)\ldots (x_n, y_n)$ be a given set of observations on two variables x and y. The correlation coefficient or coefficient of correlation is then defined as

$$\mathbf{r = \frac{cov(x,y)}{\sigma x\, \sigma y}}$$

Where, σ_x and σ_y are the standard deviation of x and y respectively, and cov(x,y) denotes covariance of x and y. This expression is known as Pearson's product-moment formula, and is used as a measure of linear correlation between x and y.

- **Properties of Correlation Coefficient:**

(i) Pearson correlation coefficient cannot exceed 1 numerically. In other words it lies between -1 and +1. Symbolically,

$$-1 \leq r \leq 1$$

In other words, r = +1 implies perfect positive correlation between the variables and r = -1 implies perfect negative correlation between the variables.

(ii) Correlation coefficient is independent of the change of both origin and scale.

(iii) Two independent variables are uncorrelated but converse is not true.

- **Calculation (Application):**

Example 5.25:

Marks of 10 students in Mathematics and Statistics are given below:

Mathematics (X):	32	38	48	43	40	22	41	69	35	64
Statistics (Y):	30	31	38	43	33	11	27	76	40	59

Calculate product-moment correlation coefficient.

Since the correlation coefficient is unaffected by changes of origin, we write x = X-43 and y = Y-38.

Table 5.17: Calculations for Correlation Coefficient

X	Y	x = (X-43)	y = (Y-38)	x^2	y^2	x.y
32	30	-11	-8	121	64	88
38	31	-5	-7	25	49	35
48	38	5	0	25	0	0
43	43	0	5	0	25	0
40	33	-3	-5	9	25	15
22	11	-21	-27	441	729	567
41	27	-2	-11	4	121	22
69	76	26	38	676	1444	988
35	40	-8	2	64	4	-16
64	59	21	21	441	441	441
$\sum X = 432$	$\sum Y = 388$	$\sum x = 2$	$\sum y = 8$	$\sum x^2 = 1806$	$\sum y^2 = 2902$	$\sum xy = 2140$

By formula we have,

$\sigma_x^2 = \frac{1806}{10} - (\frac{2}{10})^2 = 180.56$

$\sigma_y^2 = \frac{2902}{10} - (\frac{8}{10})^2 = 289.56$

$cov(x,y) = \frac{2140}{10} - (\frac{2}{10})(\frac{8}{10}) = 213.84$

Thus,

Correlation Coefficient (r) $= \frac{cov(x,y)}{\sigma x\, \sigma y} = \frac{213.84}{\sqrt{(180.56 * 289.56)}} = 0.94$

3.2.3.11. Probable Error

- **Concept:**

After computing the value of the correlation coefficient the next step is to find the extent to which it is dependable. Probable Error of correlation coefficient, usually denoted by P.E. (r) is an old measure of testing the reliability of an observed value of correlation coefficient is so far as it depends upon the conditions of random sampling.

- **Formula:**

If r is the observed correlation coefficient in a sample of n pairs of observations then its standard errors, usually denoted by,

$$S.E.\ (r) = \frac{(1 - r2)}{\sqrt{n}}$$

Probable error of the correlation coefficient is given by

$$P.E.\ (r) = 0.6745 \times S.E.\ (r)$$

$$= 0.6745 \times \frac{(1 - r2)}{\sqrt{n}}$$

Reason for taking the factor 0.6745 is that in a normal distribution 50% of the observations lie in the range M ± 0.6745 σ, where M is mean and is σ standard deviation.

Calculation (Application):

Example 5.26:

Continuing the above example 23.1, we have Correlation Coefficient (r) = 0.94.

Thus,

Standard Error = $\frac{(1 - r2)}{\sqrt{n}} = \frac{1 - (0.94)2}{\sqrt{10}} = 0.04$

And,

Probable error = 0.6745 × S.E. (r)

= 0.6745 × 0.04

= 0.02698

- ## Usage of Probable Error:

(i) The probable error of correlation coefficient may be used to determine the limits within which the population correlation coefficient may be expected to lie.

r ± P.E. (r)

(1)

This implies that if we take another random sample of the same size n from the same population from which the first sample was taken, then the observed value of the correlation coefficient, say, r_1 in the second sample can be expected to lie within the limits given in (1).

(ii) P.E. (r) may be used to test if an observed value of sample correlation coefficient is significant of any correlation in the population. The following guidelines may be used.

(a) If r < P.E. (r) i.e., if the observed value of r is less than its P.E., then correlation is not at all significant.

(b) If r > σ P.E. (r) i.e., if observed value of r is greater than σ times its P.E., then r is definitely significant.

- ## Remarks:

P.E. can be used only under the following conditions:

(i) The data must have been drawn from a normal population.

(ii) The conditions of random sampling should prevail in selecting sampled observations.

3.2.3.12. Rank Correlation

- **Concept:**

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics such as honesty, beauty, character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Phychologist, developed a formula in 1904 which consists in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under study.

- **Formula:**

Spearman's rank correlation coefficient is given by the formula

$$R = 1 - \frac{6\sum d2}{n\,(n2 - 1)}$$

Where,

d = differences in ranks

n = number of pairs of items

Spearman's rank correlation coefficient lies between -1 and +1, i.e., $-1 \leq r_k \leq 1$

- **Calculation (Application):**

Example 5.27

In a contest, two judges ranked seven candidates in order of their preference as in the following data.

Candidates:	A	B	C	D	E	F	G
Ranks by Judge I	2	1	4	5	3	7	6
Ranks by Judge II	3	4	2	5	1	6	7

Calculate the rank correlation coefficient.

Table 5.18: Calculations for Rank Correlation Coefficient

Candidates	Ranks by		d = (x-y)	d^2
	Judge I (x)	Judge II (y)		
A	2	3	-1	1
B	1	4	-3	9
C	4	2	2	4
D	5	5	0	0
E	3	1	2	4
F	7	6	1	1
G	6	7	-1	1
Total			0	20

Here, n = 7, $\sum d^2 = 20$

Thus, $R = 1 - \frac{6\sum d2}{n\,(n^2 - 1)}$

$= 1 - \frac{6\sum d2}{(n3 - n)}$

$= 1 - \frac{6 * 20}{(7^3 - 7)} = 0.64$

3.2.3.13. Coefficient of Determination (R^2)

- **Concept:**

Coefficient of correlation between two variable series is a measure of linear relationship between them and indicates the amount of variation of one variable which is associated with or is accounted for by another variable. A more useful and rapidly comprehensible measure for this purpose is the coefficient of determination which gives the percentage variation in the dependent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the determination is given by the square of the correlation coefficient, i.e., r^2 or R^2 and pronounced R-square.

- **Formula:**

Coefficient of Determination (R^2) = $\frac{\text{Explained Variance}}{\text{Total variance}} = \frac{\text{ESS}}{\text{TSS}}$

The coefficient of determination is much useful and better measure for interpreting the value of r.

- **Example (Application):**

Example 5.28:

If the value of r = 0.8, we cannot conclude that 80% of the variation in the dependent variable is due to the variation in the considering independent variables. But the coefficient of determination in this case is $r^2 = 0.64$ which implies that only 64% of the variation in the dependent variable has been explained by the considering independent variables and the remaining 36% of the variation is due to other factors.

Example 5.29:

By the same argument while comparing two correlation coefficient, one of which is 0.4 and the other is 0.8 it is misleading to conclude that the correlation in the second case is twice as high as correlation in the first case. The coefficient of determination clearly explains this view point, since in the case r = 0.4, the coefficient of determination is 0.16 and in the case r = 0.8, the coefficient of determination is 0.64, from which we can conclude that correlation in the second case is four times as high as correlation in the first case.

3.2.3.14. Regression Analysis

- **Concept:**

Regression analysis, in the general sense, means the estimation or prediction of the unknown value of the variable from the known value of the other variable. It is one of the very important statistical tools which is extensively used in almost all sciences – natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves. Cost function, production and consumption functions, etc.

Prediction or estimation is one of the major problems in almost all spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, incomes, etc. are of paramount importance to a businessman or economist. Regression analysis is one of the very scientific techniques for making such predictions. In the words of M. M. Blair “Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of data”.

In regression analysis there are two types of variables; dependent variable and independent variable. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values or is used for prediction, is called independent variable. In regression analysis independent variable is also known as predictor or explanatory variable while the dependent variable is also known as explained variable.

- **Usage of Regression Analysis:**

1. Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.
2. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.
3. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.

- **Techniques of Regression:**

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

3.2.3.14.1. Explained and Unexplained Variation

- **Concept:**

In statistics, explained variation measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. Often, variation is quantified as variance; then, the more specific term **explained variance** can be used.

The complementary part of the total variation is called **unexplained** or **residual**.

- **Formula:**

If y_i' represents the estimated value of y from the regression equation of y on x (note that y_i denotes the observed value) when $x = x_i$, i.e. $(y_i' - \bar{y}) = b_{yx}(x_i' - \bar{x})$, then it can be shown that

$\sum (y_i - \bar{y})^2 = \sum (y_i' - \bar{y})^2 + \sum (y_i - y_i')^2$

i.e. Total Variation $[\sum(y_i - \bar{y})^2]$ = Explained Variation $[\sum (y_i' - \bar{y})^2]$ + Unexplained Variation $[\sum(y_i - y_i')^2]$

Thus, it can be shown that $\frac{\text{Explained Variation}}{\text{Total Variation}} = r^2$

- **Example (Application):**

Example 5.30:

Let us consider an empirical model as follows:

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + U$

U is random error term.

Now Y is regressed on a set of explanatory variables $(X_1,\ldots,X_6)$. The estimated regression equation (by using ordinary least square technique) be

RINQ = - 311.2804*** - 0.0031 X_1* + 7.68e-07 X_2** - 0.7673 X_3***
(62.98789) (0.0017231) (3.51e-07) (0.2506054)
+ 0.4908R X_4*** + 286.7195 X_5*** + 87.3818 X_6
(0.0325607) (109.8435) (88.22784)

R-square = 0.8733

Values within parentheses are standard errors

[*** Denote 1% level of significance
** Denote 5% level of significance
* Denote 10% level of significance]

- **Explanation:**

From the above estimated results, we observed that, R^2 (explanatory power) is 0.8733 i.e. 87.33% of total variation of Y is explained by such explanatory variables and hence this 87.33% variation is called explained variation. On the other hand, $(1 - R^2) = (1 - 0.8733) = 0.1267$ i.e. 12.67% variation is called unexplained variation (it may be called residual variation).

3.2.3.15. Correlation Ratio

In statistics, the correlation ratio is a measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole population or sample. The measure is defined as the ratio of two standard deviations representing these types of variation. The context here is the same as that of the intra class correlation coefficient, whose value is the square of the correlation ratio.

Subunit – 3: Sampling Methods & Sampling Distribution

3.3.1. Sampling

- **Concept:**

A finite subset of the population, selected from it with the objective of investigating its properties is called sample and the number of unites in the sample is known as the sample size. Sampling is a tool which enables us to draw conclusions about the characteristics of the populations after studying only those objects or items that are included in the sample.

- **Main objectives of the sampling theory:**

The main objectives of the sampling theory are:

(i) To obtain the optimum results, i.e., the maximum information about the characteristics of the population with the available sources at our disposal in terms of times, money and manpower by studying the sample values only.

(ii) To obtain the best possible estimates of the population parameters.

- **Several stages of sampling process:**

1. Defining the population of concern
2. Specifying a sampling frame, a set of items or events possible to measure
3. Specifying a sampling method for selecting items or events from the frame
4. Determining the sample size
5. Implementing the sampling plan
6. Sampling and data collecting
7. Data which can be selected

3.3.1.1. Probability sampling

- **Concept:**

A probability sample is a sample in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

- **Example:**

We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income.

People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.)

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

- **Probability sampling includes:**

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Probability Proportional to Size Sampling
5. Cluster or Multistage Sampling

These various ways of probability sampling have two things in common:

1. Every element has a known non-zero probability of being sampled.
2. Involves random selection at some point.

3.3.1.2. Non probability sampling

- **Concept:**

Non probability sampling is any sampling method where some elements of the population have no chance of selection or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non-random, non-probability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

- **Example:**

We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

- **Non-probability sampling methods include:**

1. Convenience sampling,
2. Quota sampling
3. Purposive sampling

In addition, non-response effects may turn any probability design into a non-probability design if the characteristics of non-response are not well understood, since non-response effectively modifies each element's probability of being sampled.

3.3.2. Sampling Methods

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

3.3.3. Simple Random Sampling

Simple random sampling is the techniques in which "sample is so drawn that each and every unit in the population has an equal and independent chance of being included in the sample". If the unit selected in any draw is not replaced in the population before making the next draw, then it is known as simple random sampling without replacement and if it is replaced back before making the next draw, then the sampling plan is called simple random sampling with replacement.

A very important and interesting feature of simple random sampling without replacement is that "the probability of selection a specified unit of population at any given draw is equal to the probability of its being selected at the first draw". This implies that in this case from a population of size N, the probability that any sampling unit is included in the sample is 1/N and this probability remains constant throughout the drawing.

3.3.4. Systematic Sampling

Systematic sampling is slight variation of the simple random sampling in which only the first sample unit is selected at random and the remaining units are automatically selected in a definite sequence at equal spacing from one another. This technique of drawing samples is usually recommended if the complete and up-to-date list of the sample units, i.e., the frame is available and the units are arranged in some systematic order such as alphabetical, chronological, geographical order etc.

3.3.5. Stratified Sampling

When the population is heterogeneous with respect to the variable or characteristic under study, then the technique of stratified random sampling is used to obtain more efficient results. Stratification means division into layers or groups.

The criterion used for the stratification of the universe into various strata is known as stratifying factor. Some of the commonly used stratifying factors are age, sex, income, economic status, etc. In many fields of highly skewed distributions, stratification is very effective and valuable tools.

A stratified sampling approach is most effective when three conditions are met:

(i) Variability within strata are minimized

(ii) Variability between strata are maximized

(iii) The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

- **Advantages over other sampling methods**

(i) Focuses on important subpopulations and ignores irrelevant ones.

(ii) Allows use of different sampling techniques for different subpopulations.

(iii) Improves the accuracy/efficiency of estimation.

(iv) Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

- **Disadvantages**

(i) Requires selection of relevant stratification variables which can be difficult.

(ii) Is not useful when there are no homogeneous subgroups.

(iii) Can be expensive to implement.

3.3.6. Over Sampling

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling, the data are stratified on the target and a sample is taken from each stratum so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken compared to a random sample. The results usually must be adjusted to correct for the oversampling.

3.3.7. Probability-Proportional-to-Size Sampling

In some cases, the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability proportional to size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections.

Systematic sampling theory can be used to create a probability proportionate to size sample. This is done by treating each count within the size variable as a single sampling unit. Samples are then identified by selecting at even intervals among these counts within the size variable. This method is sometimes called PPS-sequential or monetary unit sampling in the case of audits or forensic sampling.

- **Example:**

Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school (1011 to 1500).

We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

3.3.8. Cluster Sampling

In this case the total population is divided, depending on the problem under study, into some recognizable subdivisions which are termed as clusters and a simple random sample of these clusters is drawn.

For example, if we are interested in obtaining the income or opinion data in a city, the whole city may be divided into N different blocks or localities (which determine the clusters) and a simple random sample of n blocks is drawn. The individuals in the selected blocks determine the cluster sample.

3.3.9. Quota Sampling

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for several years.

3.3.10. Mini-Max Sampling

In imbalanced datasets, where the sampling ratio does not follow the population statistics, one can resample the dataset in a conservative manner called mini max sampling. The mini max sampling has its origin in Anderson mini max ratio whose value is proved to be 0.5: in a binary classification, the class-sample sizes should be chosen equally. This ratio can be proved to be mini max ratio only under the assumption of LDA classifier with Gaussian distributions.

The notion of mini max sampling is recently developed for a general class of classification rules, called class-wise smart classifiers. In this case, the sampling ratio of classes is selected so that the worst-case classifier error over all the possible population statistics for class prior probabilities would be the best.

3.3.11. Accidental Sampling

Accidental sampling (sometimes known as **grab**, **convenience** or **opportunity sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey were to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing.

- **Several important considerations for researchers using convenience samples include:**

(i) Are there controls within the research design or experiment which can serve to lessen the impact of a non-random convenience sample, thereby ensuring the results will be more representative of the population?

(ii) Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population?

(iii)Is the question being asked by the research one that can adequately be answered using a convenience sample?

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample. Some variants of snowball sampling, such as respondent driven sampling, allow calculation of selection probabilities and are probability sampling methods under certain conditions.

3.3.12. Line-intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a "transect", intersects the element.

3.3.13. Panel Sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for (potentially the same) information several times over a period of time. Therefore, each participant is interviewed at two or more time points; each period of data collection is called a "wave". The method was developed by sociologist Paul Lazarsfeld in 1938 as a means of studying political campaigns. This longitudinal sampling-method allows estimates of changes in the population, for example with regard to chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or to help explain changes in continuous dependent variables such as spousal interaction.

3.3.14. Snowball Sampling

Snowball sampling involves finding a small group of initial respondents and using them to recruit more respondents. It is particularly useful in cases where the population is hidden or difficult to enumerate.

3.3.15. Multistage Sampling

As the name suggests, multistage sampling refers to a sampling technique which is carried out in various stages.

For example, if we are interested in obtaining a sample of, say, n households from a particular state the first stage units may be districts, the second stage units will be households in the villages. Each stage thus results in a reduction of the sample size.

- **Advantages:**

a. Cost and speed that the survey can be done in

b. Convenience of finding the survey sample

c. Normally more accurate than cluster sampling for the same size sample

- **Disadvantages:**

(i) Not as accurate as Simple Random Sample if the sample is the same size

(ii) More testing is difficult to do.

3.3.16. Purposive Sampling

A sample which is selected on the basis of individual judgment of the sampler is called purposive Sampling. There is no special technique for selecting a purposive sample; but the sampler picks out a typical or representative sample according to his own judgment. It all depends on the personal factor and chance is not allowed to play at all.

3.3.17. Errors in sample surveys

Survey results are typically subject to some error. Total errors can be classified into sampling errors and non-sampling errors. The term "error" here includes systematic biases as well as random errors.

- **Sampling errors and biases:**

Sampling errors and biases are induced by the sample design. They include:

(i) **Selection bias**: When the true selection probabilities differ from those assumed in calculating the results.

(ii) **Random sampling error**: Random variation in the results due to the elements in the sample being selected at random.

3.3.18. Sampling Error

In statistics, **sampling error** is incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population. Since the sample does not include all members of the population, statistics on the sample, such as means and quantiles, generally differ from parameters on the entire population. For example, if one measures the height of a thousand individuals from a country of one million, the average height of the thousand is typically not the same as the average height of all one million people in the country. Since sampling is typically done to determine the characteristics of a whole population, the difference between the sample and population values is considered a sampling error.

Exact measurement of sampling error is generally not feasible since the true population values are unknown; however, sampling error can often be estimated by probabilistic modeling of the sample.

3.3.19. Non-Sampling Error

Non-sampling errors are other errors which can impact the final survey estimates, caused by problems in data collection, processing, or sample design. They include:

1. **Over-coverage**: Inclusion of data from outside of the population.
2. **Under-coverage**: Sampling frame does not include elements in the population.
3. **Measurement error**: e.g. when respondents misunderstand a question, or find it difficult to answer.
4. **Processing error**: Mistakes in data coding.
5. **Non-response**: Failure to obtain complete data from all selected individuals.

After sampling, a review should be held of the exact process followed in sampling, rather than that intended, in order to study any effects that any divergences might have on subsequent analysis. A particular problem is that of non-response.

Two major types of non-response exist: unit nonresponse (referring to lack of completion of any part of the survey) and item non-response (submission or participation in survey but failing to complete one or more components/questions of the survey). In survey sampling, many of the individuals identified as part of the sample may be unwilling to participate, not have the time to participate (opportunity cost), or survey administrators may not have been able to contact them. In this case, there is a risk of differences, between respondents and nonrespondents, leading to biased estimates of population parameters. This is often addressed by improving survey design, offering incentives, and conducting follow-up studies which make a repeated attempt to contact the unresponsive and to characterize their similarities and differences with the rest of the frame.[The effects can also be mitigated by weighting the data when population benchmarks are available or by imputing data based on answers to other questions.

Non response is particularly a problem in internet sampling. Reasons for this problem include improperly designed surveys, over-surveying (or survey fatigue), and the fact that potential participants hold multiple e-mail addresses, which they don't use anymore or don't check regularly.

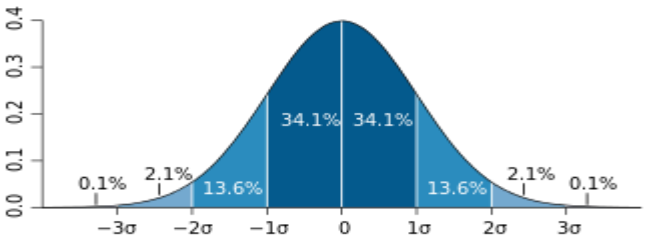
3.3.20. Standard Error

- **Concept:**

For a value that is sampled with an unbiased normally distributed error, the above depicts the proportion of samples that would fall between 0, 1, 2, and 3 standard deviations above and below the actual value.

The **standard error** (**SE**) is the standard deviation of the sampling distribution of a statistic, most commonly of the mean. The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.

Figure 2.1: Area under standard normal curve



For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean, so there is a distribution of sampled means (with its own mean and variance). The **standard error of the mean** (**SEM**) (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

In regression analysis, the term "standard error" is also used in the phrase standard error of the regression to mean the ordinary least squares estimate of the standard deviation of the underlying errors.

Subunit – 4: Statistical Inferences, Hypothesis Testing

3.4.1. Meaning of Statistical Inference

The process of going from the known sample to the unknown population is called statistical inference.

Since sample is only part of a population, the features of the former will generally differ from those of the latter. The question that naturally arises is then: what can be said about the properties of the population from knowledge of the properties of the sample? Although a satisfactory answer to this question may not be found in all cases, in the case of random sampling it can be answered with the help of the probability theory. In sampling theory, we are primarily concerned with this very question.

The basic problem of sampling theory stated above usually presents itself in one of two forms:

(i) Some feathers of the population in which an enquirer is interested may be completely unknown to him, and he may want to make a guess about this feature completely on the basis of a random sample from the population.

(ii) Some information of a tentative nature regarding the feature of the population may be available to the enquirer, and he may want to see whether the information is tenable in the light of the random sample taken from the population.

The first type of problem is the problem of estimation and the second is the problem of testing of hypotheses.

Statistical Inference = Estimation + Testing of Hypotheses

3.4.2. Point Estimation

1. In statistics, **point estimation** involves the use of sample data to calculate a single value (known as a statistic) which is to serve as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter.
2. More formally, it is the application of a point estimator to the data.
3. In general, point estimation should be contrasted with interval estimation: such interval estimates are typically either confidence intervals in the case of frequentist inference, or credible intervals in the case of Bayesian inference.

3.4.3. Interval Estimation

In statistics, **interval estimation** is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter, in contrast to point estimation, which is a single number. Jerzy Neyman (1937) identified interval estimation ("estimation by interval") as distinct from point estimation ("estimation by unique estimate"). In doing so, he recognized that then-recent work quoting results in the form of an estimate plus-or-minus a standard deviation indicated that interval estimation was actually the problem statisticians really had in mind.

The most prevalent forms of interval estimation are:

- Confidence intervals (a frequentist method)
- Credible intervals (a Bayesian method).

Other common approaches to interval estimation, which are encompassed by statistical theory, are:

- Tolerance intervals
- Prediction intervals - used mainly in Regression Analysis
- Likelihood intervals

3.4.4. Statistical Hypothesis

- **Concept:**

A **statistical hypothesis** is a scientific hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A **statistical hypothesis test** is a method of statistical inference used for testing a statistical hypothesis.

A test result is called statistically significant if it has been predicted as unlikely to have occurred by sampling error alone, according to a threshold probability - the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. In the Neyman-Pearson framework, the process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors (type-I and type-II), and by specifying parametric limits on e.g. how much type-I error will be permitted.

- **Null Hypothesis (H_0):**

The random selection of the samples from the given population makes the test of significance valid for us. For applying any test of significance we first set up a hypothesis – a definite statement about the population parameter. Such a statistical hypothesis, which is under test, is usually a hypothesis of no difference and hence is called **Null Hypothesis**. It is usually denoted by H_0. In the words of Prof. R. A. Fisher "Null Hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.

- **Alternative Hypothesis (H_1):**

In statistical hypothesis testing, the **alternative hypothesis** and the null hypothesis are the two rival hypotheses which are compared by a statistical hypothesis test.

In the domain of science two rival hypotheses can be compared by explanatory power and predictive power.

- **Example (Application):**

Example 6.1

In the illustration, the null hypothesis that, the household mean income of a locality is Rs. 5000 ($\mu = 5000$) is tested against 'both-sided alternatives' (say, $\mu > 5000$ or $\mu < 5000$), i.e.

H_0: $\mu = 5000$ against the alternative,

H_1: $\mu \neq 5000$

Thus assuming H_0 to be true, we would be looking for large differences on both sides of the expected value, i.e. in 'both tails' of the distribution. Such tests are, therefore, called "two-tailed test" that will discuss latter in more detail.

Sometimes we are interested in tests for large differences on one side only i.e. in one 'one tail' of the distribution that will also discuss latter in more detail.

For testing the null hypothesis against "one-sided alternatives (right side)" ($\mu > 5000$), i.e.,

H_0: $\mu = 5000$ against the alternative,

H_1: $\mu > 5000$

Test statistic: $Z = \frac{\mu - E(\mu)}{S.E.(\mu)}$

[i.e. $Z = \frac{\text{Observed Value} - \text{Expected Value}}{\text{Standard Error}}$]

Now the calculated value of the statistic Z is compared with 1.645, since 5% of the area under the standard normal curve lies to the right of 1.645. If the observed value of Z exceeds 1.645, the null hypothesis H_0 is rejected at 5% level of significance. If a 1% level is used, we shall replace 1.645 by 2.33. Thus the critical regions for test at 5% and 1% levels are $Z \geq 1.645$ and $Z \geq 2.33$ respectively.

For testing the null hypothesis against "one-sided alternatives (left side)" ($\mu < 5000$) i.e

H_0: $\mu = 5000$ against the alternative,

H_1: $\mu < 5000$

The calculated value of Z is compared with -1.645 for significance at 5% level, and with -2.33 for significance at 1% level. The critical regions are now $Z \leq -1.645$ and $Z \leq -2.33$ respectively.

Table 6.1: Formulation of Null and Alternative Hypothesis

Problem	Null Hypothesis	Alternative Hypothesis
To test whether $\mu = 5000$, or		
(i) μ is different from 5000	H_0: $\mu = 5000$	H_1: $\mu \neq 5000$
(ii) μ is more than 5000	H_0: $\mu = 5000$	H_1: $\mu > 5000$
(iii) μ is less than 5000	H_0: $\mu = 5000$	H_1: $\mu > 5000$

Table 6.2: Type of Test and Critical Region

Alternative Hypothesis	Type of Alternative	Type of Test	Critical Region
H_1: $\mu \neq 5000$	Both-sided	Two-tailed	Both Tails
H_1: $\mu > 5000$	One-sided	One-tailed	Right Tail
H_1: $\mu > 5000$	One-sided	One-tailed	Left Tail

3.4.5. Statistical Test

Statistical test is a procedure whose inputs are samples and whose result is a hypothesis.

3.4.6. Region of Acceptance

The set of values of the test statistic for which we fail to reject the null hypothesis is called region of acceptance.

3.4.7. Region of Rejection / Critical Region

The set of values of the test statistic for which the null hypothesis is rejected is called critical region or region of rejection.

3.4.8. Critical Value

Critical Value is the threshold value delimiting the regions of acceptance and rejection for the test statistic.

3.4.9. Power of a Test $(1 - \beta)$

The probability of rejecting a false null hypothesis is called **power of a test**. Therefore, power is the probability of drawing a correct conclusion by the test, when the null hypothesis is false. For a specified value of the parameter consistent with the alternative hypothesis,

Power of a Test = 1 – Probability of type-II error

3.4.10. Level of Significance (α)

The maximum size of the type-I error, which we are prepared to risk is known as the level of significance. It is usually denoted by 'α' and is given by:

Prob. [Rejecting H_0 when H_0 is true] = α

- **Remarks:**

Commonly used levels of significance in practice are 1%, 5% and 10%. For example if we adopt 1% level of significance, it implies that in 1 sample out of 100 samples we are likely to reject a correct H_0. In other words, this implies that we are 99% confident that our decision to reject H_0 is correct. Level of significance is always fixed in advance before collecting the sample information.

3.4.11. P-Value

The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

3.4.12. Statistical Significance

In statistics, statistical significance (or a statistically significant result) is attained when a p-value is less than the significance level. The p-value is the probability of obtaining at least as extreme results given that the null hypothesis is true whereas the significance or alpha (α) level is the probability of rejecting the null hypothesis given that it is true. As a matter of good scientific practice, a significance level is chosen before data collection and is usually set to 0.05 (5%). Other significance levels (e.g., 0.01) may be used, depending on the field of study.

Statistical significance is fundamental to statistical hypothesis testing. In any experiment or observation that involves drawing a sample from a population, there is always the possibility that an observed effect would have occurred due to sampling error alone. But if the p-value is less than the significance level (e.g., $p < 0.05$), then an investigator may conclude that the observed effect actually reflects the characteristics of the population rather than just sampling error. Investigators may then report that the result attains statistical significance, thereby rejecting the null hypothesis.

3.4.13. One- Tailed and Two-Tailed Tests

In statistical significance testing, a **one-tailed test** and a **two-tailed test** are alternative ways of computing the statistical significance of a parameter inferred from a data set, in terms of a test statistic. A two-tailed test is used if deviations of the estimated parameter in either direction from some benchmark value are considered theoretically possible; in contrast, a one-tailed test is used if only deviations in one direction are considered possible. Alternative names are **one-sided** and **two-sided** tests; the terminology "tail" is used because the extreme portions of distributions, where observations lead to rejection of the null hypothesis, are small and often "tail off" toward zero as in the normal distribution or "bell curve", pictured above right.

- ## Application of One-Tailed and Two-Tailed Test:

One-tailed tests are used for asymmetric distributions that have a single tail, such as the chi-squared distribution, which are common in measuring goodness-of-fit, or for one side of a distribution that has two tails, such as the normal distribution, which is common in estimating location; this corresponds to specifying a direction. Two-tailed tests are only applicable when there are two tails, such as in the normal distribution, and correspond to considering either direction significant.

In the approach of Ronald Fisher, the null hypothesis H_0 will be rejected when the p-value of the test statistic is sufficiently extreme and thus judged unlikely to be the result of chance. In a one-tailed test, "extreme" is decided before hand as either meaning "sufficiently small" or meaning "sufficiently large" – values in the other direction are considered not significant. In a two-tailed test, "extreme" means "either sufficiently small or sufficiently large", and values in either direction are considered significant. For a given test statistic there is a single two-tailed test, and two one-tailed tests, one each for either direction. Given data of a given significance level in a two-tailed test for a test statistic, in the corresponding one-tailed tests for the same test statistic it will be considered either twice as significant (half the p-value), if the data is in the direction specified by the test, or not significant at all (p-value above 0.5), if the data is in the direction opposite that specified by the test.

For example, if flipping a coin, testing whether it is biased towards heads is a one-tailed test, and getting data of "all heads" would be seen as highly significant, while getting data of "all tails" would be not significant at all ($p = 1$). By contrast, testing whether it is biased in either direction is a two-tailed test, and either "all heads" or "all tails" would both be seen as highly significant data. In medical testing, while one is generally interested in whether a treatment results in outcomes that are better than chance, thus suggesting a one-tailed test; a worse outcome is also interesting for the scientific field, therefore one should use a two-tailed test that corresponds instead to testing whether the treatment results in outcomes that are different from chance, either better or worse. In the archetypal lady tasting tea experiment, Fisher tested whether the lady in question was better than chance at distinguishing two types of tea preparation, not whether her ability was different from chance, and thus he used a one-tailed test.

3.4.14. Types of Errors

3.4.14.1. Type- I Error:

A **type-I error**, also known as an **error of the first kind**, occurs when the null hypothesis (H_0) is true, but is rejected. It is asserting something that is absent, a false hit. A type-I error may be compared with a so called false positive (a result that indicates that a given condition is present when it actually is not present) in tests where a single condition is tested for.

The type-I error rate or significance level is the probability of rejecting the null hypothesis given that it is true. It is denoted by the Greek letter α (alpha) and is also called the alpha level. By convention, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

3.4.14.2. Type- II Error:

A **type-II error**, also known as an error of the second kind, occurs when the null hypothesis is false, but erroneously fails to be rejected. It is failing to assert what is present, a miss. A type-II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss') in a test checking for a single condition with a definitive result of true or false. A Type-II error is committed when we fail to believe a truth. In terms of folk tales, an investigator may fail to see the wolf ("failing to raise an alarm"). Again, H_0: no wolf.

The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals $1-\beta$).

What we actually call type I or type-II error depends directly on the null hypothesis. Negation of the null hypothesis causes type-I and type-II errors to switch roles.

The goal of the test is to determine if the null hypothesis can be rejected. A statistical test can either reject or fail to reject a null hypothesis, but never prove it true.

- **Remarks:**

In any test procedure, four possible mutually disjoint and exhaustive decisions are:

(i) Reject H_0 when actually it is not true.

(ii) Accept H_0 when it is true.

(iii)Reject H_0 when it is true.

(iv)Accept H_0 when it is false.

The decision in (i) and (ii) are correct decision while the decisions (iii) and (iv) are wrong decisions. These decisions may be expressed in the following dichotomous table:

Table 6.3: Rejection Rules for Null Hypothesis

True Statement		Decision from Sample	
		Reject H_0	Accept H_0
	H_0 is True	Wrong (Type-I Error)	Correct
	H_0 is False	Correct (Type-II Error)	Wrong

3.4.15. Degrees of Freedom

What are degrees of freedom?

We can define them as the number of values we can choose freely.

- **Example 6.2:**

Assume that we are dealing with two sample values, '*a'* and '*b'*, and we know that they have a mean of 18. Symbolically, the situation is

$$\frac{a+b}{2} = 18$$

How can we find what values '*a'* and '*b'* can take on in this situation? The answer is that '*a'* and '*b'* can be any two values whose sum is 36, because 36/2 = 18.

Suppose we learn that '*a'* has a value of 10. Now *'b'* is no longer free to take on any value but must have the value of 26, because

If a = 10

Then $\frac{10+b}{2} = 18$

Or, 10 + b = 36

Or, b = 26

This example shows that when there are two elements in a sample and we know the sample mean of these two elements, we are free to specify only one of the elements because the other element will be determined by the fact that the two elements sum to twice the sample mean. Statisticians say, "We have one degree of freedom".

- **Example 6.3:**

Look at another example. There are seven elements in our sample, and we learn that the mean of these elements is 16. Symbolically, we have the situation:

$$\frac{a+b+c+d+e+f+g}{7} = 16$$

In this case, the degrees of freedom, or the number of variables we can specify freely, are (7 - 1 = 6). We are free to give values of six variables, and then we are no longer free to specify the seventh variable. It is determined automatically.

With two sample values, we have one degree of freedom (2 – 1 = 1), and with seven sample values, we have six degrees of freedom (7 – 1 = 6). In each of these two examples, then, we have (n – 1) degrees of freedom, assuming *n* is the sample size.

- **Remarks:**

Thus, as the name suggests the degree of freedom, abbreviated as d.f., denotes the extent of independence (freedom) enjoyed by a given set of observed frequencies. Degrees of freedom are usually denoted by the letter v of the Greek alphabet. Suppose we are given a set of *n* observed frequencies which are subjected to *k* independent constraints, then

d.f. = (Number of frequencies) – (Number of independent constraints on them)

or, v = n – k

3.4.16. Use of Different Statistical Tests

3.4.16.1. Chi-Square (χ^2) Test

- **Concept:**

The square of standard normal variable is called chi-square distribution with 1 degree of freedom. Thus if X is a random variable following normal distribution with mean μ and standard deviation σ then $(\frac{X-\mu}{\sigma})$ is a standard normal variate.

Thus, $(\frac{X-\mu}{\sigma})^2$ is a chi-square distribution with 1 degree of freedom.

- **Application of χ^2 Distribution:**

Chi-square distribution has a number of applications, some of which are enumerated below:

d. Chi-square distribution is used to test goodness of fit.

e. Chi-square distribution is used to test independence of attributes.

f. To test if the population has a specified value of the variance σ^2.

- **Example 6.4: Using χ^2 distribution for large sample**

Application 1: Test for goodness of fit:

A die is thrown 60 times with the following results:

Face:	1	2	3	4	5	6	Total
Frequency	6	10	8	13	11	12	60

Are the data consistent with the hypothesis that the die is unbiased? (Given $\chi^2_{0.01} = 15.09$ for 5 degrees of freedom)

Test:

The null hypothesis is that, H_0: the die is unbiased

Against the alternative, H_1: H_0 is not true (i.e. the die is biased)

Test statistic: $\chi^2 = \sum \{\frac{(f_0 - f_e)2}{f_e}\}$ with (6 – 1) = 5 degrees of freedom.

Where, f_0 is observed frequencies of different classes and f_e is the expected frequencies of different classes.

Then the probability of each face is 1/6, and the expected frequency is 60 × 1/6 = 10 for each.

Observed frequency (f_0):	6	10	8	13	11	12
Expected frequency (f_e):	10	10	10	10	10	10
$(f_0 - f_e)^2$:	16	0	4	9	1	4

Thus, $\chi^2 = \frac{16}{10} + \frac{0}{10} + \frac{4}{10} + \frac{9}{10} + \frac{1}{10} + \frac{4}{10} = 3.4$

Since the observed value of χ^2 is less than the tabulated value 15.09 at 1% level for 5 degrees of freedom, we cannot reject the null hypothesis at 1% level of significance. The conclusion is that the data are in agreement with the hypothesis of an unbiased die.

- ## Example 6.5: Using χ^2 distribution for small sample:

Application 2: Test for a specified value of S.D:

A random sample of size 20 from a normal population gives a sample mean of 42 and a sample standard deviation of 6. Test the hypothesis that the population S.D. is 9. Clearly state the alternative hypothesis you allow for and the level of significance adopted.

Test:

Given that sample size (n) = 20. Sample S.D. (S) = 6, it is required to test the null hypothesis is that, H_0: $\sigma = 9$

Against the alternative, H_1: $\sigma > 9$

Test statistic: $\chi^2 = \frac{(x_i - \bar{x})2}{6^2} = \frac{n S^2}{6^2}$ with (20 – 1) = 19 degrees of freedom.

Thus, $\chi^2 = \frac{n S^2}{6^2} = \frac{20 * 6^2}{9^2} = 8.89$

From χ^2 table we have for 19 degrees of freedom, 5% value of χ^2 = 30.14. Since the observed value of χ^2 viz. 8.89, is less than the tabulated vale at 5% level of significance, we cannot reject the null hypothesis, and conclude that the population S.D. may be 9.

3.4.16.2. 't'-Test

- **Concept:**

Suppose we are interested to test:

g. If the given normal population has a specified value of the population mean, say, μ_0.

h. If the sample mean $\bar{x}$ differs significantly from specified value of population mean.

i. If a given random sample $x_1, x_2, \ldots, x_n$ of size n has been drawn from a normal population with specified mean, μ_0.

Basically, all the three problems are same. We set up the corresponding null hypothesis as follows:

1. H_0: $\mu = \mu_0$ i.e., the population mean is μ_0.
2. H_0: There is no significant difference between the sample mean and the population mean. In other words, the difference between μ_0 and μ is due to fluctuations of sampling.
3. H_0: The random sample has been drawn from the population with mean μ_0.

Under H_0 the test-statistic is

$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$; where, $\bar{x} = \frac{1}{n}\sum x$ and $s^2 = \frac{1}{n-1}\sum(x-\bar{x})$ and it follows t-distribution with (n-1) degrees of freedom.

- **Application of t-Distribution:**

The t-distribution has a number of applications in statistics:

(i) t-distribution is used for the test of significance of single mean, population variance being unknown.

(ii) t-distribution is used for the test of significance of the difference between two sample means, the population variances being equal but unknown.

(iii) t-distribution is used for the test of significance of an observed sample correlation coefficient.

- **Example 6.6: Using t - distribution for small sample**

Application 1: Test for a specified mean (S.D is unknown):

A random sample of size 20 from a normal population gives a sample mean of 42 and sample standard deviation of 6. Test the hypothesis that the population mean is 44.

Test:

The null hypothesis is that, H_0: the population mean (μ) = 44

Against the alternative, H_1: $\mu \neq 44$ [i.e. both sided alternatives]

Since the population standard deviation is not known, we use t-test.

Here n = 20, $\bar{x}$ = 42 and s = 6.

Test statistic: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$ with (n-1) degrees of freedom.

Thus, $t = \frac{42 - 44}{\frac{6}{\sqrt{19}}}$ = - 1.45 [Degrees of freedom = (20-1) = 19]

From the table value of t-distribution, we find for 19 d.f. the percentage points of t are respectively $t_{.025}$ = 2.09 and $t_{.005}$ = 2.86. We have to use a two-tailed test, because the alternatives are both-sided. Since |t| = 1.45 is less than the tabulated value at 5% level of significance corresponding to the two tails, null hypothesis is accepted at 5% level of significance, and we conclude that the population mean may be 44.

- ## Example 6.7: Using t - distribution for small sample

Application 2: Test for equality of two means (S.Ds unknown):

Two types of batteries are tested for their length of life and the following data are obtained:

	No of sample	Mean life in Hours	Variance
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means? Value of *t* for 15 degrees of freedom at 5% level is 2.131.

Test:

It is assumed that two populations are normal distributions with μ_1 and μ_2 and a common S.D. σ. It is also assumed that the two samples are randomly drawn and independent.

The null hypothesis is that, H_0: $\mu_1 = \mu_2$

Against the alternative, H_1: $\mu_1 \neq \mu_2$

Test statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$ with (n_1 + n_2 - 2) degrees of freedom.

$S = \sqrt{\frac{(9*121)+(8*144)}{9+8-2}} = \sqrt{149.4} = 12.2$

Thus, the observed value of t-statistic is

$$t = \frac{600-640}{12.2\sqrt{(\frac{1}{9}+\frac{1}{8})}} = \frac{-40}{12.2*0.486} = -6.7$$

Degrees of freedom = 9 + 8 – 2 = 15

Since the alterative hypothesis is both-sided, the test is two-tailed. The critical region is given by both the tails of t-distribution.

At 5% level, critical region: $|t| \geq 2.131$

Since the observed value of *t* ($|t| = |-6.7| = 6.7$) is larger than the tabulated value (2.131), we reject the null hypothesis at 5% level of significance and we conclude that there is 'significant difference in the two means'.

3.4.16.3. 'z'-Test

- **Concept:**

Fisher's Z-transformation has the following test-statistic:

$$z = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}}$$

- **Application of z-Distribution:**

It has the following applications in statistics:

(i) To test if the correlation coefficient in the population has a specified value.

(ii) To test if two independent sample correlations r_1 and r_2 differ significantly.

- **Example 6.8: Using z - distribution for large sample**

Application 1: Test for a specified proportion:

A die is thrown 400 times and 'six, resulted 80 times. Do the data justify the hypothesis of an unbiased die?

Let us assume that the die is unbiased, i.e. the null hypothesis is that the probability of obtaining a 'six' with the die is 1/6.

The null hypothesis is that, H_0: P = 1/6

Against the alternative, H_1: P ≠ 1/6

Test statistic: $z = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}}$

Since 'six' occurred 80 times out of 400, the observed value of the proportion (P) of 'six' is

P = 80/400 = 0.2

On the assumption that, H_0 is true (i.e. the die is unbiased), the expected value of the proportion of 'six' is equal to 1/6 = 0.167

(S.E. of *P*) $= \sqrt{\frac{\frac{1}{6} * \frac{5}{6}}{400}} = \frac{\sqrt{5}}{120} = 0.0186$

Thus, $z = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}} = \frac{0.2 - 0.167}{0.0186} = 1.77$

When H_0 is true, the statistic *z* follows standard normal distribution. Since the value of z does not fall in the critical region (critical region at 5% level is $|z| \geq 1.96$), it is not significant at 5% level. Thus the null hypothesis is accepted and concludes that the die may be unbiased.

- ## Example 6.9: Using z - distribution for large sample

Application 2: Test for a specified mean:

A sample of 400 male students is found to have a mean height of 171.38 cms. Can it be reasonably regarded as a sample from a large population with mean height 171.17 cms. And S.D. 3.30 cms?

Test:

We assume that the sample really comes from a large population with mean 171.17 and S.D. 3.30.

The null hypothesis is that, H_0: ($\mu = 171.17$, $\sigma = 3.30$)

Against the alternative, H_1: $\mu \neq 171.17$

Test statistic: $z = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}}$

Since the sample size n = 400 is large, the sample mean ($\bar{x}$) is approximately normally distributed.

Observed value ($\bar{x}$) = 171.38

Expected value (μ_0) = 171.17

Standard error of $\bar{x}$ $= \frac{6}{\sqrt{n}} = \frac{3.30}{\sqrt{400}} = 0.165$

Thus, $z = \frac{\text{Observed value} - \text{Expected value}}{\text{Standard Error}} = \frac{171.38 - 171.17}{0.165} = 1.27$

Since the alternative hypothesis is both-sided (i.e. μ is either more than or less than 171.17), the critical region of the test is also two-tailed. At 5% level,

Critical region is $|z| \geq 1.96$.

Since the value of z (1.27) does not fall in the critical region (critical region at 5% level is $|z| \geq 1.96$), it is not significant at 5% level. Thus the null hypothesis is accepted and concludes that the sample may be regarded as having arisen from the given population.

- **Example 6.10: Using z - distribution for small sample**

Application 3: Test for a specified population mean (S.D. known):

A random sample of size 10 is taken from a normal population, whose variance is known to be 7.056 sq. inches. If the observations are (in inches) 65, 71, 64, 71, 70, 69, 64, 63, 67 and 68, test the hypothesis that the population mean is 69 inches. Also obtain 99% confidence limits for the population mean.

The null hypothesis is that, H_0: $\mu = 69$

Against the alternative, H_1: $\mu \neq 69$

Population S.D. (σ) = $\sqrt{7.056}$ = 2.656313 and, n = 10.

For the given data, sample mean $\bar{x} = 67.2$

Test statistic: $z = \frac{\bar{x} - \mu_0}{\frac{6}{\sqrt{n}}}$

$= \frac{67.2 - 69}{\frac{2.656313}{\sqrt{10}}} = \frac{-1.8}{0.84} = -2.14$

Since $|z| = 2.14$ exceeds 1.96 (the critical region at 5% level is $|z| \geq 1.96$), the null hypothesis is rejected and we conclude that the population mean is not 69 inches.

99% confidence limits for μ are $67.2 \pm 2.58 \left(\frac{\sqrt{7.056}}{\sqrt{10}}\right) = 67.2 \pm 2.17 = 65.03$ and 69.37 inches.

3.4.16.4. 'F'-Test

- **Concept:**

F-statistic is the ratio of two independent chi-square variates divided by their respective degrees of freedom.

If y_1 and y_2 are independent chi-square variates with degrees of freedom n_1 and n_2 respectively, then

$F = \frac{\frac{y_1}{n_1}}{\frac{y_2}{n_2}}$ follows F-distribution with (n_1, n_2) degrees of freedom.

- **Application of F-Distribution:**

The F-distribution has a number of applications in statistics:

(i) F-test is used for equality of population variances.

(ii) F-test is used for testing the significance of an observed sample correlation ratio.

(iii)F-test is used for testing the significance of an observed sample multiple correlation.

(iv)F-test is used for testing the linearity of regression.

F-test is used for testing the equality of several population means, i.e., for testing $H_0 = \mu_1 = \mu_2 = \ldots = \mu_k$, (say), for k normal populations.

- **Example 6.11: Using F-distribution for small sample**

Application 1: Test for equality of two S.D.s (means unknown):

The standard deviations calculated from two random samples of sizes 9 and 13 are 2.1 and 1.8 respectively. May the samples be regarded as drawn from normal populations with the same S.D.? (The 5% value of *F* from tables with d.f. 8 and 12 is $F_{.05} = 2.85$)

Here, $n_1 = 9$, $n_2 = 13$, $S_1 = 2.1$ and $S_2 = 1.8$. The 'unbiased' estimates of the variances σ_1^2 and σ_2^2 in the two populations are respectively,

$S_1^2 = (\frac{n_1}{n_1 - 1})S_1^2$ and $S_2^2 = (\frac{n_2}{n_2 - 1})S_2^2$; where S_1 and S_2 are the standard deviations in the samples.

The null hypothesis is that, H_0: $\sigma_1 = \sigma_2$

Against the alternative, H_1: $\sigma_1 > \sigma_2$

Test statistic: $F = \frac{s_1^2}{s_2^2}$ follows F-distribution with d.f. ($n_1 - 1$, $n_2 - 1$).

Thus, $F = \frac{4.96}{3.51} = 1.41$ [where, $S_1^2 = \frac{9*(2.1)^2}{9-1} = 4.96$, $S_2^2 = \frac{13*(1.8)^2}{13-1} = 3.51$]

Degrees of freedom are (8, 12)

Since the observed value of F (1.41) is less than the tabulated value of F at 5% level of significance (2.85) corresponding to d.f. (8, 12), the null hypothesis cannot be rejected at 5% level of significance and, we can conclude that the population S.D.s may be equal.

3.4.16.5. Analysis of Variance (ANOVA)

- **Concept:**

Analysis of variance (ANOVA) has been defined as the statistical technique for the *'separation of variation due to a group of causes from the variation due to other groups'*. Here we shall discuss the simplest use of this technique, namely testing whether the means of a number of populations are equal. The method is based upon an unusual result that the equality of several population means can be tested by comparing the sample variances using F-distribution. It may be recalled that the t-statistic is used for testing whether two population means are equal. The analysis of variance test may therefore be taken as an extension of this test for the case of more than two means.

- **Different Sources of Variation:**

The term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components (two kinds of variances):

(i) "Within-group Variance": It is the average variance of the members of each group around their respective group means, i.e. the mean value of the scores in a sample (as members of each group may vary among themselves).

(ii) **"Between-group Variance":** It represents the variance of group means around the total or grand mean of all groups, i.e. the best estimate to the population mean (as the group means may vary considerably from each other).

- **The Mean Square and Sum Square:**

If the means of all the populations are equal, then the variability 'between' groups would result only from chance and hence would be the same as the variability arising from 'within' groups. On the other hand, if population means are not equal, the variability 'between' groups would be more than the variability 'within' groups.

The measure of variability used in the analysis of variance is called a "Mean Square". This is similar to a variance and is defined as:

$$\textbf{Mean Square} = \frac{\textbf{Sum of squared deviations from mean}}{\textbf{Degree of freedom}}$$

Note that in the t-test for a specified mean, the population variance σ^2 is estimated as the sum of squared deviations from mean divided by the 'sample size minus one'. In the t-test for equality of two means, the common population variance is estimated as the sum of the squared deviation of the two groups of observations from the respective means divided by the 'sum of sample sizes minus two'. These divisors are referred to as "degrees of freedom".

- **Technique of ANOVA:**

The technique of analysis of variance as a single composite test of significance, for the difference between several group means demands the derivation of two independent estimates of the population variance, one based on variance of group means (between group variance) and the other on the average variance within the groups (within group variance). Ultimately, the comparison of the size of between-groups variance and within-groups variance called F-ratio and is denoted by;

$$\mathbf{F} = \frac{\textbf{Between-groups variance}}{\textbf{Within-groups variance}}$$

$$\text{Or, } \mathbf{F} = \frac{\textbf{Mean Square Between groups (MSB)}}{\textbf{Mean Square Within groups (MSW)}}$$

It is used as a critical ratio for determining the significance of the difference between group means at a given level of significance (viz. 5% or 1%).

- **Steps in Computation (One-way Classified Data):**

(i) Reduce the sample observations by subtracting a suitable constant.

(ii) From these reduced figures, obtain

(a) Total (T_i) for each group

(b) Grand total ($T = \sum T_i$)

(c) Total of squares of all figures ($\sum\sum X_{ij}^2$)

(iii) Calculate

(a) Correction Factor (CF) = $\frac{T^2}{N}$

(b) Total SS = $\sum\sum X_{ij}^2 - CF$

(c) SSB = $\sum(\frac{T_i^2}{n_i}) - CF$

(d) SSW = Total SS – SSB

(iv) Write down the Degrees of Freedom (D.F.):

(a) D.F. for SSB = (k – 1)

(b) D.F. for SSW = (N – k)

(v) Calculate the Mean Squares:

(a) MSB = $\frac{SSB}{k-1}$

(b) MSW = $\frac{SSW}{N-k}$

(vi) Obtain the observed value of F on dividing MSB by MSW: F = $\frac{MSB}{MSW}$

(vii) Consult the F-table and obtain the theoretical values of F at 5% level (say) corresponding to the degrees of freedom (k – 1, N – k).

(viii) If the observed value of F at (vi) equals or exceeds the theoretical value of F at (vii), reject the null hypothesis, and concludes that the population means are not equal. Otherwise, they may be taken to be equal.

- ## Example 6.12:

A random sample of five motor-car tyres is taken from each of 3 brands manufactured by three companies. The lifetime of these tyres is shown below. On the basis of the data, test whether the average lifetime of the 3 brands of tyres are equal or not.

Lifetime of tyres ('000 miles)

Brand	A	B	C
	35	32	34
	34	32	33
	34	31	32
	33	28	32
	34	29	33

Solution:

Each observation is reduced by 30, and shown below:

Table 6.4: Calculations for Analysis of Variance (One-way)

Samples	1	2	3	
	5	2	4	
	4	2	3	
	4	1	2	
	3	-2	2	
	4	-1	3	
Total:	$T_1 = 20$	$T_2 = 2$	$T_3 = 14$	$T = 36$
Total of squares:	82	14	42	$\sum\sum X_{ij}^2 = 138$
Sample size:	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$N = 15$

C.F = $\frac{T^2}{N} = \frac{36^2}{15} = 86.4$

Total SS = $\sum\sum X_{ij}^2 - C.F = 138 - 86.4 = 51.6$

SSB = $\frac{T^2}{n_i}$ - C.F. = $(\frac{20^2}{5} + \frac{2^2}{5} + \frac{14^2}{5})$ - 86.4

= $\frac{(400 + 4 + 196)}{5}$ - 86.4 = (120 – 86.4) = 33.6

SSE = Total SS – SSB = 51.6 – 33.6 = 18

Table 6.5: Analysis of Variance Table (One-way)

Source of variation	**S.S.**	**d.f.**	**M.S.**	**F-values**	
				Observed	**Tabulated**
Between groups	33.6	2	16.8	11.2	F.05 = 3.89
Within groups (Error)	18	12	1.5		F.01 = 6.93
Total	51.6	14	-	-	-

Since the observed value of F exceeds the 1% tabulated value (viz. 6.93), we reject the null hypothesis of equality of population means, and conclude that the average lifetimes of 3 brands of tyres are not equal.

- ## Steps in Computation (Two-way Classified Data):

(i) Reduce the observations by subtracting a constant from each.

(ii) From the reduced figures, we obtain

(a) Totals **(T_i)** for each group classified according to factor A, i.e. for each row.

(b) Totals **(T_j')** for each group classified according to factor B, i.e. for each column.

(c) Grand total ($T = \sum T_i = \sum T_j'$) for all figures.

(d) Total of the squares of all figures ($\sum\sum X_{ij}^2$)

(iii) Calculate the correction factor (CF) and sums of squares:

(a) C.F = $\frac{T^2}{N}$

(b) Total SS = $\sum\sum X_{ij}^2$ – C.F

(c) SSA = $\frac{T_i^2}{k}$ - C.F.

(d) SSB = $\frac{T_j^2}{h}$ - C.F.

(e) SSE = Total SS – SSA - SSB

(iv) Write down the Degrees of Freedom (DF):Calculate the correction factor (CF) and sums of squares:

(a) D.F. for Total SS = hk – 1

(b) D.F. for SSA = h – 1

(c) D.F. for SSB = k – 1

(d) D.F. for SSE = (D.F. for Total SS) – (D.F. for SSA + D.F. for SSB)

(v) Calculate the Mean Squares:

(a) MSA = $\frac{SSA}{DF}$

(b) MSB = $\frac{SSB}{DF}$

(c) MSE = $\frac{SSE}{DF}$

(vi) Obtain the observed values of F on dividing MSA and MSB by MSE:

(a) $F_1 = \frac{MSA}{MSE}$

(b) $F_2 = \frac{MSB}{MSE}$

(vii) Comparing the F-tables and obtain the theoretical values of F at 5% level (say) for the appropriate degrees of freedom.

(a) If the observed value of F_1 exceeds the theoretical value, we conclude that classification according to factor A has a differential effect on the value of the variable. That is, the means of classes by factor A are significantly different. Otherwise, there is no differential effect.

(b) If the observed value of F_2 exceeds the theoretical value, we conclude that classification according to factor B has a differential effect on the value of the variable. If not, then there is no differential effect.

- **Example 6.12:**

Three experiments determine the moisture content of samples of a powder, each man taking a sample from each of 4 consignments. The results are given below:

(a) Perform an analysis of variance on these data and discuss whether there is any significant difference between consignments or between experiments.

Experimenter	Consignment			
	I	II	III	IV
A	9	10	9	10
B	12	11	9	11
C	11	12	10	12

(b) Also, test at 5% level which pairs of experimenters differ significantly, if any. [Given $F_{0.05} = 5.14$ for d.f. (2, 6), $F_{0.05} = 4.76$ for d.f. (3, 6), and $t_{0.025} = 2.45$ for 6 d.f.].

Solution:

Each observation is reduced by 10, and shown below:

Table 6.6: Calculations for Analysis of Variance (Two-way)

Experimenter	Consignment				Total (T_i)
	I	II	III	IV	
A	-1	0	-1	0	-2
B	2	1	-1	1	3
C	1	2	0	2	5
Total (T_j')	**2**	**3**	**-2**	**3**	**T = 6**

Total of the squares of all figures

$\sum\sum X_{ij}^2 = (-1)^2 + 0^2 + (-1)^2 + \ldots + 0^2 + 2^2 = 18$

Correction Factor (CF) = $\frac{T^2}{N} = \frac{6^2}{12} = 3$

Total SS = $\sum\sum X_{ij}^2 - CF = 18 - 3 = 15$

SS between Experimenters = $\frac{T_1^2 + T_2^2 + T_3^2}{4} - CF$

$= \frac{(-2)^2+3^2+5^2}{4} - 3 = 6.5$

SS between Consignments = $\frac{T_1'^2 + T_2'^2 + T_3'^2 + T_4'^2}{3} - CF$

$= \frac{2^2+3^2+(-2)^2+3^2}{3} - 3 = 5.67$

SS due to Error = Total SS – (SS between Experimenters) – (SS between Consignments)

= 15 – 6.5 – 5.67 = 2.83

Table 6.7: Analysis of Variance Table (Two-way)

Source of variation	S.S.	d.f.	M.S.	F-values	
				Observed	Tabulated
Between Experimenters	6.5	2	3.25	6.91	F.05 = 5.14
Between Consignments	5.67	3	1.89	4.02	F.01 = 4.76
Error	2.83	6	0.47	-	-
Total	15.0	11	-	-	-

(a) Since the observed value of F for experimenters (viz. 6.91) is larger than the corresponding tabulated value for d.f. (2, 6) and is significant at 5% level. We therefore conclude that the mean moisture content as determined by the three experimenters are not equal; i.e. there are significant differences between experimenters.

The observed value of F for consignments (viz. 4.02) is less than the corresponding tabulated value for d.f. (3, 6) and hence is not significant at 5% level. We then conclude that the moisture content of the 4 consignments may not be different from one another; i.e. there are no significant differences between consignments.

(b) Critical Differences between totals for experimenters:

C.D. = $\sqrt{0.47} \times \sqrt{2 \times 4} \times 2.45 = 4.8$

We have $T_1 = -2$, $T_2 = 3$, $T_3 = 5$. The differences between T_1 and T_2, T_1 and T_3, T_2 and T_3 are respectively 5, 7 and 2. Since 5 and 7 are larger than C.D., experimenters A & B and A& C differ significantly.

Subunit – 5: Linear Regression Model and their properties – BLUE

3.5.1. Stochastic:

The term ***stochastic*** occurs in a wide variety of professional or academic fields to describe events or systems that are unpredictable due to the influence of a random variable. The word "stochastic" comes from the Greek word *stokhos.*

Researchers refer to physical systems in which they are uncertain about the values of parameters, measurements, expected input and disturbances as "stochastic systems". In probability theory, a purely stochastic system is one whose state is randomly determined, having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely. In this regard, it can be classified as non-deterministic (i.e., "random") so that the subsequent state of the system is determined probabilistically. Any system or process that must be analyzed using probability theory is stochastic at least in part. Stochastic systems and processes play a fundamental role in mathematical models of phenomena in many fields of science, engineering, finance and economics.

3.5.2. Coefficient of Determination (R^2):

In statistics, the **coefficient of determination** denoted $\boldsymbol{R^2}$ or $\boldsymbol{r^2}$ and pronounced **R squared**, is a number that indicates how well data fit a statistical model – sometimes simply a line or a curve. It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model

There are several definitions of R^2 that are only sometimes equivalent. One class of such cases includes that of simple linear regression where r^2 is used instead of R^2. In this case, if an intercept is included, then r^2 is simply the square of the sample correlation coefficient (i.e., r) between the outcomes and their predicted values. If additional explanatory variables are included, R^2 is the square of the coefficient of multiple correlation. In both such cases, the coefficient of determination ranges from 0 to 1.

Important cases where the computational definition of R^2 can yield negative values, depending on the definition used, arise where the predictions that are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data, and where linear regression is conducted without including an intercept. Additionally, negative values of R^2 may occur when fitting non-linear functions to data. In cases where negative values arise, the mean of the data provides a better fit to the outcomes than do the fitted function values, according to this particular criterion.

Coefficient of determination (R^2) = Explained Variations / Total Variations

3.5.3. Simple Linear Regression:

In statistics, **simple linear regression** is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of *n* points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

The adjective *simple* refers to the fact that the outcome variable is related to a single predictor. The slope of the fitted line is equal to the correlation between *y* and *x* corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass (*x*, *y*) of the data points.

Other regression methods besides the simple ordinary least squares (OLS) also exist (see linear regression). In particular, when one wants to do regression by eye, one usually tends to draw a slightly steeper line; closer to the one produced by the total least squares method. This occurs because it is more natural for one's mind to consider the orthogonal distances from the observations to the regression line, rather than the vertical ones as OLS method does.

3.5.4. Least Squares Method:

The method of **least squares** is a standard approach in regression analysis to the approximate solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation.

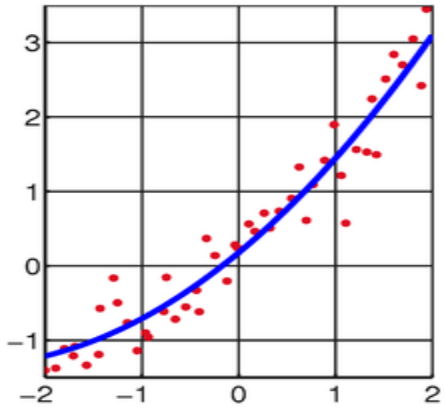
The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. When the problem has substantial uncertainties in the independent variable (the *x* variable), then simple regression and least squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least squares problems fall into two categories: linear or ordinary least squares and non-linear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The non-linear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Polynomial least squares describe the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

When the observations come from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical. The method of least squares can also be derived as a method of moments estimator.

The following discussion is mostly presented in terms of linear functions but the use of least-squares is valid and practical for more general families of functions. Also, by iteratively applying local quadratic approximation to the likelihood (through the Fisher information), the least-squares method may be used to fit a generalized linear model.



For the topic of approximating a function by a sum of others using an objective function based on squared distances, see least squares (function approximation).

The result of fitting a set of data points with a quadratic function.

The least-squares method is usually credited to Carl Friedrich Gauss (1795), but it was first published by Adrien-Marie Legendre.

3.5.5. Stepwise Regression:

In statistics, **stepwise regression** includes regression models in which the choice of predictive variables is carried out by an automatic procedure. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R-square, Akaike information criterion, Bayesian information criterion, Mallows's C_p, PRESS, or false discovery rate.

The frequent practice of fitting the final selected model followed by reporting estimates and confidence intervals without adjusting them to take the model building process into account has led to calls to stop using stepwise model building altogether or to at least make sure model uncertainty is correctly reflected.

3.5.6. Main approaches

The main approaches are:

- **Forward selection**, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.
- **Backward elimination**, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.
- **Bidirectional elimination**, a combination of the above, testing at each step for variables to be included or excluded.

A widely used algorithm was first proposed by Efroymson (1960). This is an automatic procedure for statistical model selection in cases where there is a large number of potential explanatory variables, and no underlying theory on which to base the model selection. The procedure is used primarily in regression analysis, though the basic approach is applicable in many forms of model selection. This is a variation on forward selection.

At each stage in the process, after a new variable is added, a test is made to check if some variables can be deleted without appreciably increasing the residual sum of squares (RSS). The procedure terminates when the measure is (locally) maximized, or when the available improvement falls below some critical value.

3.5.7. Regression Fallacy

The **regression** (or **regressive**) **fallacy** is an informal fallacy. It ascribes cause where none exists. The flaw is failing to account for natural fluctuations. It is frequently a special kind of the post hoc fallacy.

Explanation

Things like golf scores, the earth's temperature, and chronic back pain fluctuate naturally and usually regress towards the mean. The logical flaw is to make predictions that expect exceptional results to continue as if they were average (see Representativeness heuristic). People are most likely to take action when variance is at its peak. Then after results become more normal they believe that their action was the cause of the change when in fact it was not causal.

This use of the word "regression" was coined by Sir Francis Galton in a study from 1885 called "Regression Toward Mediocrity in Hereditary Stature". He showed that the height of children from very short or very tall parents would move towards the average. In fact, in any situation where two variables are less than perfectly correlated, an exceptional score on one variable may not be matched by an equally exceptional score on the other variable. The imperfect correlation between parents and children (height is not entirely heritable) means that the distribution of heights of their children will be centered somewhere between the average of the parents and the average of the population as whole. Thus, any single child can be more extreme than the parents, but the odds are against it.

Examples

When his pain got worse, he went to a doctor, after which the pain subsided a little. Therefore, he benefited from the doctor's treatment.

The pain subsiding a little after it has gotten worse is more easily explained by regression towards the mean. Assuming the pain relief was caused by the doctor is fallacious.

The student did exceptionally poorly last semester, so I punished him. He did much better this semester. Clearly, punishment is effective in improving students' grades.

Often exceptional performances are followed by more normal performances, so the change in performance might better be explained by regression towards the mean. Incidentally, some experiments have shown that people may develop a systematic bias for punishment and against reward because of reasoning analogous to this example of the regression fallacy.

The frequency of accidents on a road fell after a speed camera was installed. Therefore, the speed camera has improved road safety.

Speed cameras are often installed after a road incurs an exceptionally high number of accidents, and this value usually falls (regression to mean) immediately afterwards. Many speed camera proponents attribute this fall in accidents to the speed camera, without observing the overall trend.

Some authors have claimed that the alleged "Sports Illustrated Cover Jinx" is a good example of a regression effect: extremely good performances are likely to be followed by less extreme ones, and athletes are chosen to appear on the cover of *Sports Illustrated* only after extreme performances. Assuming athletic careers are partly based on random factors, attributing this to a "jinx" rather than regression, as some athletes reportedly believed, would be an example of committing the regression fallacy.

3.5.8. Properties of OLS Regression Estimators

Property 1: Linear

This property is more concerned with the estimator rather than the original equation that is being estimated. In assumption A_1, the focus was that the linear regression should be “linear in parameters.” However, the *linear* property of OLS estimator means that OLS belongs to that class of estimators, which are linear in Y, the dependent variable. Note that OLS estimators are linear only with respect to the dependent variable and not necessarily with respect to the independent variables. The *linear* property of OLS estimators doesn’t depend only on assumption A_1 but on all assumptions A_1 to A_5.

Property 2: Unbiasedness

If you look at the regression equation, you will find an error term associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number. Therefore, before describing what unbiasedness is, it is important to mention that unbiasedness property is a property of the estimator and not of any sample.

Unbiasedness is one of the most desirable properties of any estimator. The estimator should ideally be an unbiased estimator of true parameter/population values.

Consider a simple example: Suppose there is a population of size 1000, and you are taking out samples of 50 from this population to estimate the population parameters. Every time you take a sample, it will have the different set of 50 observations and, hence, you would estimate different values of βa_0 and βa_i. The unbiasedness property of OLS method says that when you take out samples of 50 repeatedly, then after some repeated attempts, you would find that the average of all the βa_0 and βa_i from the samples will equal to the actual (or the population) values of βa_0 and βa_i.

Mathematically,

$E(b_o) = \beta_o$

$E(b_i) = \beta_i$

Here, 'E' is the expectation operator.

In layman's term, if you take out several samples, keep recording the values of the estimates, and then take an average, you will get very close to the correct population value. If your estimator is biased, then the average will not equal the true parameter value in the population.

The unbiasedness property of OLS in Econometrics is the basic minimum requirement to be satisfied by any estimator. However, it is not sufficient for the reason that most times in real-life applications, you will not have the luxury of taking out repeated samples. In fact, only one sample will be available in most cases.

Property 3: Best: Minimum Variance

This property is what makes the OLS method of estimating α and β the best of all other methods. When there is more than one unbiased method of estimation to choose from, that estimator which has the lowest variance is best. (Variance is a measure of how far the different α and β are from their mean; the variance is the average distance of an element from the average.)

An estimator (a function that we use to get estimates) that has a lower variance is one whose individual data points are those that are closer to the mean. This estimator is statistically more likely than others to provide accurate answers. The OLS estimator is one that has a minimum variance.

This property is simply a way to determine which estimator to use.

- An estimator that is unbiased but does not have the minimum variance is not good.
- An estimator that has the minimum variance but is biased is not good
- An estimator that is unbiased and has the minimum variance of all other estimators is the best (efficient).

The OLS estimator is an efficient estimator.

Property 4: Asymptotic Unbiasedness

This property of OLS says that as the sample size increases, the biasedness of OLS estimators disappears.

Property 5: Consistency

An estimator is said to be consistent if its value approaches the actual, true parameter (population) value as the sample size increases. An estimator is consistent if it satisfies two conditions:

a. It is asymptotically unbiased

b. Its variance converges to 0 as the sample size increases.

Both these hold true for OLS estimators and, hence, they are consistent estimators. For an estimator to be useful, consistency is the minimum basic requirement. Since there may be several such estimators, asymptotic efficiency also is considered. Asymptotic efficiency is the sufficient condition that makes OLS estimators the best estimators.

Subunit – 6: Identification Problem and Simultaneous Equation System

3.6.1. Definition

A system describing joint dependence of variables is called a system of simultaneous equation.

In simultaneous equation system there is two-way causation, i.e. Y= $f(X)$ and $X = f(Y)$ i.e. there is a joint dependence of variables.

3.6.2. Structural forms

A structural model is a complete system of equations which describe the structure of the relationship of the economic variables. Structural equation express the endogenous variables as functions of the other endogenous variables, predetermined variables and the disturbances (random variable).

As an illustration we shall use the following simple model for a closed economy –

$C_t = \alpha_0 + \alpha_1 Y_t + u_1$

$I_t = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_2$

$Y_t = C_t + I_t + G_t$

The first equation is a consumption function. The second is an investment function and third is a definitional equation. The system is complete in that it contains three endogenous variables, C_t, I_t, Y_t in three equation. The model contains two predetermined variables, government expenditure G and lag income, Y_{t-1}.

3.6.2. Reduced form

The reduced form of a structural model is the model in which the endogenous variables expressed as a function of the predetermined variables only. The reduced form is obtained in two ways. The first is to express the endogenous variables directly as functions of the predetermined variables and the second for obtaining the reduced form of a model is to solve the structural system of endogenous variables in terms of the predetermined variables, the structural parameters and disturbances.

Let the structural form equations:

$C_t = \alpha_1 Y_t + u_1 - - - (1)$

$I_t = \beta_1 Y_t + \beta_2 Y_{t-1} + u_2 - - - (2)$

$Y_t = C_t + I_t + G_t - - - (3)$

(a) Substitute C_t and I_t in equation (3)

$Y_t = \alpha_1 Y_t + u_1 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_2 + G_t$

$Y_t = \alpha_1 Y_t + \beta_1 Y_t + \beta_2 Y_{t-1} + G_t + (u_1 + u_2)$

$(1 - \alpha_1 - \beta_1) Y_t = \beta_2 Y_{t-1} + G_t + u_1 + u_2$

$$Y_t = \left(\frac{\beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{u_1 + u_2}{1 - \alpha_1 - \beta_1}\right)$$

This is the reduced form of the third structural equation.

(b) Substitute Y_t in the consumption function: -

$$C_t = \alpha_1 \left[\left(\frac{\beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{u_1 + u_2}{1 - \alpha_1 - \beta_1}\right)\right] + u_1$$

$$C_t = \left(\frac{\alpha_1 \beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{\alpha_1}{1 - \alpha_1 - \beta_1}\right) G_t + \frac{\alpha_1 u_1 + \alpha_1 u_2}{1 - \alpha_1 - \beta_1} + u_1$$

$$C_t = \left(\frac{\alpha_1 \beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{\alpha_1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{u_1 + \alpha_1 u_2 - \beta_1 u_1}{1 - \alpha_1 - \beta_1}\right)$$

This is the reduced form equation of the consumption

(c) Substitute Y_t into investment equation: -

$$I_t = \beta_1 \left[\left(\frac{\beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{u_1 + u_2}{1 - \alpha_1 - \beta_1}\right)\right] + \beta_2 Y_{t-1} + u_2$$

$$= C_t = \left(\frac{\beta_1 \beta_2}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{\beta_1}{1 - \alpha_1 - \beta_1}\right) G_t + \frac{(u_1 + u_2)\,\beta_1}{1 - \alpha_1 - \beta_1} + \beta_2 Y_{t-1} + u_2$$

$$= \left(\frac{\cancel{\beta_1 \beta_2} + \beta_2 - \alpha_1 \beta_2 - \cancel{\beta_1 \beta_2}}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{\beta_1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{\beta_1 u_1 + \cancel{u_2 \beta_1} + u_2 - \alpha_1 u_2 - \cancel{\beta_1 u_2}}{1 - \alpha_1 - \beta_1}\right)$$

$$= \left(\frac{\beta_2 (1 - \alpha_1)}{1 - \alpha_1 - \beta_1}\right) Y_{t-1} + \left(\frac{\beta_1}{1 - \alpha_1 - \beta_1}\right) G_t + \left(\frac{u_2 + \beta_1 u_1 - \alpha_1 u_2}{1 - \alpha_1 - \beta_1}\right)$$

This is the reduced form of the investment function.

First Method: -

Let, $\pi_{11} = \frac{\alpha_1\beta_2}{1-\alpha_1-\beta_1}$, $\pi_{12} = \frac{\alpha_1}{1-\alpha_1-\beta_1}$

$$\pi_{21} = \frac{\beta_2(1-\alpha_1)}{1-\alpha_1-\beta_1}, \pi_{22} = \frac{\beta_1}{1-\alpha_1-\beta_1}$$

$$\pi_{31} = \frac{\beta_2}{1-\alpha_1-\beta_1}, \pi_{31} = \frac{1}{1-\alpha_1-\beta_1}$$

The reduced form equation be,

$$\left.\begin{aligned} C_t &= \pi_{11}Y_{t-1} + \pi_{12}G_t + v_1 \\ I_t &= \pi_{21}Y_{t-1} + \pi_{22}G_t + v_2 \\ Y_t &= \pi_{31}Y_{t-1} + \pi_{32}G_t + v_3 \end{aligned}\right\}$$

3.6.2.1. Seemingly unrelated equation

Let the simultaneous equation in general form can be written as follows-

$$\gamma_{11}Y_{1t} + \gamma_{12}Y_{2t} + \cdots + \gamma_{1G}Y_{Gt} + \beta_{11}X_{1t} + \cdots + \beta_{1k}X_{kt} = \epsilon_{1t}$$

$$\gamma_{21}Y_{1t} + \gamma_{22}Y_{2t} + \cdots + \gamma_{2G}Y_{Gt} + \beta_{21}X_{2t} + \cdots + \beta_{2k}X_{kt} = \epsilon_{2t}$$

$$\gamma_{G1}Y_{1t} + \gamma_{G2}Y_{2t} + \cdots + \gamma_{GG}Y_{Gt} + \beta_{G1}X_{Gt} + \cdots + \beta_{Gk}X_{kt} = \epsilon_{Gt}$$

Put it in matrix form,

$$\underbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} \cdots \gamma_{1G} \\ \gamma_{21} & \gamma_{22} \cdots \gamma_{2G} \\ \vdots & \vdots \quad \vdots \\ \gamma_{G1} & \gamma_{G2} \cdots \gamma_{GG} \end{bmatrix}}_{\Gamma} \underbrace{\begin{bmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{Gt} \end{bmatrix}}_{Y_t} + \underbrace{\begin{bmatrix} \beta_{11} & \beta_{12} \cdots \beta_{1k} \\ \beta_{21} & \beta_{22} \cdots \beta_{2k} \\ \vdots & \vdots \quad \vdots \\ \beta_{G1} & \beta_{G2} \cdots \beta_{Gk} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{kt} \end{bmatrix}}_{X_t} = \underbrace{\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{Gt} \end{bmatrix}}_{\epsilon_t}$$

The seemingly unrelated equation be

$$\Gamma = \begin{bmatrix} \gamma_{11} & 0 & \cdots & 0 \\ 0 & \gamma_{22} & \cdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{GG} \end{bmatrix}$$

In this case each endogenous variable appears in on the one equation. In this case we do not have simultaneous equation system. Instead we have seemingly unrelated equation.

3.6.2.2. Recursive equation model: -

A model is called recursive if its structural equation can be ordered in such a way that the first includes only predetermined variables in the right-hand side; the second equation contains predetermined variables and the first endogenous variable (of the first equation) in the right-hand side; and so on.

e.g. $Y_1 = f(X_1, X_2, \dots, X_R, U_1)$

$Y_2 = f(X_1, X_2, \dots, X_R, Y_1, U_2)$

$Y_3 = f(X_1, X_2, \dots, X_R, Y_1, Y_2, U_3)$

And so on.

In this case matrix is a triangular matrix.

Where $\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \cdots \gamma_{1G} \\ \gamma_{21} & \gamma_{22} \cdots \gamma_{2G} \\ \vdots & \vdots \quad \vdots \\ \gamma_{G1} & \gamma_{G2} \cdots \gamma_{GG} \end{bmatrix}$

i.e. $= \begin{bmatrix} \gamma_{11} & 0 & \dots\dots & 0 \\ \gamma_{21} & \gamma_{22} & \dots\dots & 0 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{G1} & \gamma_{G2} & \gamma_{G3} \cdots & \gamma_{GG} \end{bmatrix} \begin{bmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \\ \vdots \\ Y_{Gt} \end{bmatrix}$

the solution of Y_{1t} is completely determined by equation (1). The solution of Y_{2t} is determined by equation (1) and equation (2). The solution of Y_{Gt} is determined by all equations.

But we cannot identify from this scatter diagram how function can be estimated from market transaction data.

This problem is identification problem.

Problem of Estimation: -

Let us consider a model

$C_t = \alpha + \beta Y_t + u_t ----(1)$

$Y_t = C_t + I_t ----(2)$

The above model is a structural form model.

Putting C_t in equation (2)

$Y_t = \alpha + \beta Y_t + I_t + u_t$

$Y_t = \left(\frac{\alpha}{1-\beta}\right) + \frac{I_t}{1-\beta} + \frac{u_t}{1-\beta} ----(3)$

And

$C_t = \alpha + \beta Y_t + u_t$

$C_t = \alpha + \beta Y_t + I_t + u_t$

$or, \quad C_t = \left(\frac{\alpha}{1-\beta}\right) + \left(\frac{\beta}{1-\beta}\right) I_t + \frac{u_t}{1-\beta} - - - -(4)$

Equation (3) and equation (4) are reduced form model.

C_t and Y_t are endogenous variables and I_t is pre-determined variable.

Regress Y_t on I_t in equation (3) we get

$\left(\frac{\alpha}{1-\beta}\right) and \left(\frac{1}{1-\beta}\right)$

Regress C_t on I_t in equation (4) we get

$\left(\frac{\alpha}{1-\beta}\right) and \left(\frac{\beta}{1-\beta}\right)$

Let $Cov(Y_t, u_t) = E\left[\left(Y_t - E(Y_t)\right)\left(u_t - E(u_t)\right)\right]$

$Y_t = \frac{\alpha}{1-\beta} + \frac{I_t}{1-\beta} + \frac{u_t}{1-\beta}$

$E(Y_t) = \frac{\alpha}{1-\beta} + \frac{I_t}{1-\beta} \qquad [as\ E(u_t) = 0]$

$\therefore \quad Y_t - E(Y_t) = \frac{u_t}{1-\beta}$

$\therefore \quad Cov(Y_t, u_t) = E\left[\frac{u_{t^2}}{1-\beta}\right]$

$= \frac{1}{1-\beta} \ . E(u_{t^2}) = \frac{\sigma u^2}{1-\beta} \neq 0$

∴ OLS technique cannot be applied.

3.6.3. Conditions for Identification Problem:

There are two conditions which must be fulfilled for an equation to be identified.

3.6.3.1. The order Condition for Identification:

In the model of G simultaneous equation with G endogenous variables and K pre-determined variables, an equation which include g endogenous variables and k pre-determined variable is identified if the number of pre-determined variables excluded from that equation $(K-k)$ which not less than the number of endogenous variable included in that equation less one i.e. $K-k \geq g-1$.

This is called the order condition.

This is necessary condition.

3.6.3.2. The rank condition for identification:

In the model of G simultaneous equation an equation is identified if and only if matrix Δ which is constructed from coefficient of all the variables excluded from that specific equation but included in other equations of the model, as a rank required the number of equation less one

i.e. $Rank(\Delta) = G-1$

This is necessary and sufficient condition.

3.6.3.3. The properties of the condition of identification:

The order condition of identification is the necessary condition while the rank condition of identification is necessary as well as sufficient condition. In other word for an equation to be identified the rank condition must be satisfied but whether the equation is exactly or over identified that is determined by order condition.

1.) <u>Over identified</u>:

When $(K-k) > (g-1)$ and rank $(\Delta) = G-1$ then we call it the equation is over identified.

2.) <u>Exactly identified/ just identified:=</u>

$(K-k) = (g-1)$ and rank $(\Delta) = G-1$

3.) Under identified:

(a) $(K-k) < (g-1)$

(b) $(K-k) \geq (g-1)$ but rank $(\Delta) < G-1$

Example of order condition: -

Example 1:

Demand: $Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + \epsilon_{1t} - - - (1)$

Supply: $Q_t = \beta_0 + \beta_1 P_t + \epsilon_{2t} - - - (2)$

Check: - Order check:

Here, $K = 1,\ G = 2$

For demand function: - $k = 1,\ g = 2$

$\therefore (K-k) = 1-1 = 0\ and\ (g-1) = 2-1 = 1$

$\therefore (K-k) < (g-1)$

$\Rightarrow$ **The demand function is under identified**

For supply equation: - $k = 0,\ g = 2$

$\therefore (K-k) = 1-0 = 1$

$(g-1) = 2-1 = 1$

$\therefore (K-k) = (g-1)$

$\Rightarrow$ **Supply function is just or exactly identified**

Rank (Δ) check: -

$Q_t - \alpha_1 P_t - \alpha_0 - \alpha_2 Y_t = \epsilon_{1t}$

$Q_t - \beta_1 P_t - \beta_0 - 0.Y_t = \epsilon_{2t}$

$Rank\ of\ row = 1$ as $\Delta = [-\alpha_2]_{1\times 1}$

$\therefore G-1 = 2-1 = 1$

$\therefore Rank\ of\ \Delta = G-1$

$\therefore$ Rank order condition is satisfied.

$\therefore$ Supply function is just identified.

Example 2: -

Let us given the following macro-Economic model-

Consumption function: - $C_t + \alpha_0 + \alpha_1 Y_t + \alpha_2 C_{t-1} + \epsilon_{1t}$

Investment function: -$I_t = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + \beta_3 I_{t-1} + \epsilon_{2t}$

Tax function: $T_t = \gamma_0 + \gamma_1 Y_t + \epsilon_{3t}$

Income Identity: - $Y_t = C_t + I_t + G_t$

The model is complete because the total no. of endogenous variables $G = 4(C_t, Y_t, I_t, T_t)$ and the total number of endogenous variables $K = 4(C_{t-1}, Y_{t-1}, I_{t-1}, G_t)$ are equal.

Identification status of Consumption function: -

Order check: -

Here, $G = 4,\ K = 4$

And $g = 2,\ k = 1$

$\therefore K - k = 4 - 1 = 3,\ \ g - 1 = 2 - 1 = 1$

$\therefore (K - k) > g - 1$

Rank (Δ) check: -

$C_t - \alpha_1 Y_t + 0I_t + 0T_t - \alpha_0 - \alpha_2 C_{t-1} + 0Y_{t-1} + 0I_{t-1} + 0G_t = \epsilon_{1t}$

$0.C_t - \beta_1 Y_t + I_t + 0T_t - \beta_0 - 0C_{t-1} - \beta_2 Y_{t-1} - \beta_3 I_{t-1} + 0G_t = \epsilon_{2t}$

$0.C_t - \gamma_1 Y_t + 0I_t + 1T_t - \gamma_0 + 0C_{t-1} + 0.Y_{t-1} + 0I_{t-1} + 0G_t = \epsilon_{3t}$

$-C_t + 1Y_t - I_t + 0T_t + 0 + 0C_{t-1} + 0Y_{t-1} + 0I_{t-1} - G_t = 0$

First equation: $\Delta = \begin{bmatrix} 1 & 0 & -\beta_2 & -\beta_3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -1 \end{bmatrix}_{3\times 5}$

$\Delta_1 = \begin{bmatrix} 1 & 0 & -\beta_2 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 1(0-0) - 0(\ ) - p_2(0+1)$
$= -\beta_2 \neq 0$

$\therefore Rank\ of\ \Delta = 3$

$G - 1 = 4 - 1 = 3$

$\therefore Rank\ of\ \Delta = G - 1$ ∴ Rank condition is satisfied.

∴ consumption function is over identified.

Identification status of Investment Function: -

Order check: -

$I_t = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + \beta_3 I_{t-1} + \epsilon_{2t}$

$G = 4, \qquad K = 4$

$g = 2, \qquad k = 2$

$or, K - k = 4 - 2 = 2 \quad g - 1 = 2 - 1 = 1$

$\therefore (K - k) > (g - 1)$

Rank check: -

$$\Delta = \begin{bmatrix} 1 & 0 & -\alpha_2 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 \end{bmatrix}_{3\times 4}$$

$$\Delta_2 = \begin{vmatrix} 1 & 0 & -\alpha_2 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{vmatrix} = \alpha_2 \neq 0$$

$Rank\ of\ \Delta = 3$

$\therefore G - 1 = 4 - 1 = 3$

$\therefore Rank\ (\Delta) = G - 1$

So, investment function is over identified.

Identification status of Tax function: -

Order check: -

$G = 4, \quad K = 4$

$g = 2, \quad k = 0$

$\therefore K - k = 4 - 0 = 4 \quad g - 1 = 2 - 1 = 1$

$\therefore (K - k) > g - 1.$

Rank check: -

$$\Delta = \begin{pmatrix} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & \beta_2 & p_3 & 0 \\ -1 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

$$\Delta_3 = \begin{vmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{vmatrix} \begin{matrix} = 1(0-0) - 0(\,) - 1(0+1) \\ = 1 \neq 0 \end{matrix}$$

$\therefore Rank\ of\ \Delta = 3$

$\therefore G - 1 = 4 - 1 = 3$

$\therefore Rank(\Delta) = G - 1$

$\therefore$ Rank condition is satisfied.

$\therefore$ Tax function is over identified.

Subunit – 7: Lagged Variables & Distributed Lag Models

3.7.1. Meaning

Distributed lag models are model which include lagged values of the exogeneous variables and lagged values of the dependent variables among the set of explanatory variables.

The general form of a distributed lag model with only lagged exogeneous variables is

$$Y_t = a + b_0X_t + b_1X_{t-1} + b_2X_{t-2} + \cdots + b_3X_{t-3} + \cdots + u_t$$

They are called distributed lag models because the influence of the explanatory variable on the dependent variable is distributed over a number of past values of X. the number of lags, S, may be either finite or infinite. However, in order to avoid explosive values of Y_t, we assume that the b's have a finite sum

$$\sum_{i=0}^{s} bi < \infty$$

The average lag is defined as the weighted average of all the lags involved with weights being the relative size of the respective b coefficients.

$$Average\ lag = \frac{\sum_{i=0}^{s} ibi}{\sum_{i=0}^{s} bi} = \sum_{i=0}^{s} i\ \frac{bi}{\sum_{i=0}^{s} bi}$$

Lagged values of the variables are important explanatory variables in most economic relationships because economic behaviour in any one period is to a great extent determined by past experience and past patterns of behaviour.

3.7.1. Some Examples of Lagged Variable Model: -

1. The consumption function:

Recent versions of the consumption function postulate that the current level of consumption depends on past levels of consumption, due to 'habit persistence', on current income and past levels of income and other factors

$$C_t = f(C_{t-1}, C_{t-2}, \dots, Y_t, Y_{t-1}, Y_{t-2}, \dots, X_{1t}, X_{2t} \dots)$$

2. <u>The demand for durables</u>: -

The demand for durables (D_d) depends, among others, on present income(Y_t), past levels of income(Y_{t-s}) which determine the amount saved for the acquisition of the durables, on the stocks of durables or equivalently on past acquisitions of durables $(S_{d,}t-1)$, prices(P_t) and so on.

$D_d = f(Y_t, Y_{t-1}, \dots, S_d, t-1, P_t)$

3. <u>the demand for inventory investment:</u> -

The firm usually define their inventories on the basis of the actual sales of the three-past period (and other factors)

$I_t = f(X_{t-1}, X_{t-2}, X_{t-3} \dots)$

4. <u>The investment function</u>: -

Investment projects depends on past outputs, on expectations about future profits, on capital stock and other factors.

$I_t = f(X_t, X_{t-1}, X_{t-2}, \dots, \pi_t, K_{t-1}, r_{t-4}, \dots)$

where, $X = level\ of\ output$

$\pi = profit$

$K = capital$

$r = interest\ rate$

3.7.2. Some Models

3.7.2.1. Adaptive Expectation Model (by P. Cagan):

This model is based on the following behavioural hypothesis. The value of y in any one period t depends not on the actual value of Xt, but on the 'expected' or 'permanent' level of x at time t, say X_{t^*}.

Suppose that the quantity demanded is determined by the expected price P_t rather than the actual price

$Q_t = b_0 + b_1 P_{t^*} + u_t - - - (1)$

Similarly, according to Friedman's 'permanent income hypothesis' the level of consumption C_t is determined by the 'expected' income or by the permanent income Y_{t^*}

$C_t = C_1 Y_{t^*} + u_t - - - (2)$

Now the 'expected' variables are exante variables which are not observable.

Let us first write the 'adaptive expectation' model in the general form

$Y_t = b_0 + b_1 X_{t^*} + u_t - - - (3)$

Since X_{t^*} is not directly observable, we postulate that expectation concerning its value are formed on the 'adaptive' rule.

$X_{t^*} - X_{t-1^*} = \gamma(X_t - X_{t-1^*}) \, ,0 < \gamma \leq 1 - - - (4)$

Where, $\gamma = expectation\ coefficient.$

$or, \quad X_{t^*} = X_{t-1^*} + \gamma(X_t - X_{t-1^*})$

We next proceed with the following transformation leading to the substitution of the unobservable variable X_{t^*} in the original model (equation 1)

$Y_t = b_0 + b_1 X_{t^*} + u_t$

Solving for X_{t^*} we obtain

$$X_{t^*} = -\frac{b_0}{b_1} + \frac{1}{b_1} Y_t - \frac{1}{b_1} u_t - - - (5)$$

Lagging one period

$$X_{t-1^*} = -\frac{b_0}{b_1} + \frac{1}{b_1} Y_{t-1} - \frac{1}{b_1} u_{t-1}$$

Substituting X_{t^*} and X_{t-1^*} in the 'adaptive expectations' equation (4) we obtain

$$\left(-\frac{b_0}{b_1} + \frac{1}{b_1} Y_t - \frac{1}{b_1} u_t\right) - \left(-\frac{b_0}{b_1} + \frac{1}{b_1} Y_{t-1} - \frac{1}{b_1} u_{t-1}\right)$$

$$or, \quad \cancel{\frac{b_0}{b_1}} + \frac{1}{b_1} Y_t = \gamma \left[X_1 - \left(\cancel{-\frac{b_0}{b_1}} + \frac{1}{b_1} Y_{t-1} - \frac{1}{b_1} u_{t-1}\right)\right]$$

$$= \gamma X_t + \gamma \frac{b_0}{b_1} - \gamma \frac{1}{b_1} Y_{t-1} + \gamma \frac{1}{b_1} u_{t-1}$$

$$or, \quad Y_t - u_t - Y_{t-1} + u_{t-1} = b_1 \gamma X_t + \gamma b_0 - \gamma Y_{t-1} + \gamma u_{t-1}$$

$$or, \quad Y_t = (\gamma b_0) + (\gamma b_1) X_t + (1 - \gamma) Y_{t-1} + [u_t - (1 - \gamma) u_{t-1}]$$

3.7.2.2. Partial Adjustment Model (by Nerlove):

Nerlove (and others) in an attempt to avoid the estimation difficulties which, arise with Koyck's assumption, postulated the following model which is based on a different behavioural hypothesis. There is a desired level of Y in period t, say Y_{t^*}, which depends on the level of X in period t, X_t that is

$$Y_{t^*} = b_0 + b_1 X_t + u_t - - - (1)$$

The model (1) cannot be measured because the desired quantity Y_{t^*} is not observable. To replace it we must postulate some behavioural principle i.e. some specific rule for decision making by the investing firm. The 'stock adjustment principle' implies the following behavioural pattern.

The gradual adjustment process may be expressed in the 'so-called adjustment equation'.

$Y_t - Y_{t-1} = \gamma(Y_{t^*} - Y_{t-1}) + u_t - - - (2) \qquad 0 < \gamma \leq$

Where, $Y_t - Y_{t-1} =$ actual change in capital stock (i.e. realised investment in period t)

$Y_{t^*} - Y_{t-1} =$ desired change in capital stock (desired investment)

$\gamma =$ adjustment coefficient.

This behavioural rule reals: the achieved (realised) by the firm in any one period change $(Y_{t^*} - Y_{t-1})$

Substituting $Y_{t^*} = b_0 + b_1 X_t + u_t$ into the adjustment equation (2) we obtain

$Y_t - Y_{t-1} = \gamma[(b_0 + b_1 X_t + u_t) - Y_{t-1}] + v_t$

$or, \qquad Y_t - Y_{t-1} = \gamma b_0 + \gamma b_1 X_t + \gamma u_t - \gamma Y_{t-1} + v_t$

$or, \qquad Y_t = (\gamma b_0) + (\gamma b_1) X_t + (1-\gamma) Y_{t-1} + (v_t + \gamma u_t)$

Which reads as follows: the capital stock in any one period t depends partly in the level of output in that period and partly on the existing capital stock at the beginning of the period.

3.7.2.3. Koyck's Geometric Lag Scheme/Koyck's Model:

This is one of the most popular distributed lag models in applied research. Koyck's distributed lag model assumes that the weights (lag coefficients) are declining continuously following the pattern of a geometric progression. The original model includes only exogeneous lagged variables.

$Y_t = a_0 + b_0 X_t + b_1 X_{t-1} + b_2 X_{t-2} + \cdots + u_t - - - (1)$

Where, $u \sim N\ (0, \sigma u^2)$

Koyck's geometric lag-scheme implies that more recent values of X exist a greater influence on Y than remotes value of X. in particular the lag coefficients of this model decline in the form of a geometric progression

$b_1 = \lambda b_0$

$b_2 = \lambda^2 b_1$

And in general, $b_i = \lambda^i b_0 \qquad 0 < \lambda < 1$

Substituting in the original model we obtain

$Y_t = a_0 + b_0 X_t + (\lambda b_0) X_{t-1} + (\lambda^2 b_0) X_{t-2} + \cdots + u_t - - - (2)$

Lagging by one period

$Y_{t-1} = a_0 + b_0 X_{t-1} + (\lambda b_0) X_{t-2} + (\lambda^2 b_0) X_{t-3} + \cdots + u_{t-1} - - - (3)$

Multiplying through by λ and subtracting from the equation (2) we obtain

$Y_t - \lambda Y_{t-1} = a_0(1-\lambda) + b_0 X_t + (u_t - \lambda u_{t-1})$

$or, \qquad Y_t = a_0(1-\lambda) + b_0 X_t + \lambda Y_{t-1} + v_t$

Where, $(v_t = u_t - \lambda u_{t-1})$

Subunit – 8: Multicollinearity

3.8.1. Definition of Multicollinearity:

Our important assumption of the CLRM is that the explanatory variables are not related to each other. But when the explanatory variables are related to each other then we call it presence of multicollinearity.

3.8.2. Consequences of Multicollinearity:

If the inter correlation between the explanatory variables is perfect ($r_{xi,xj} = 1$) then,

(a) The estimates of the coefficients are indeterminate.

(b) The standard errors of these estimates become infinitely large.

Proof of (a): -

Suppose that the relation to be estimated is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + U$

Where, β_1 is the partial regression coefficient (P.R.C) of X_1 and β_2 is the partial regression coefficient of X_2 and that X_1 and X_2 are related with the exact relation $X_2 = KX_1$, where K is any arbitrary constant.

Let $S = \sum ei^2 = \sum(Y - \alpha - \beta_1 X_1 - \beta_2 X_2)^2$

$$\therefore \frac{\delta S}{\delta \alpha} = 0 \Rightarrow \sum (Y - \alpha - \beta_1 X_1 - \beta_2 X_2)(-1) = 0$$

$$or, \bar{Y} = \alpha + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 - - - (1)$$

$$\frac{\delta S}{\delta \beta_1} = 0 \Rightarrow \sum (Y - \alpha - \beta_1 X_1 - \beta_2 X_2)(-X_1) = 0$$

$$\Rightarrow \sum YX_1 = \alpha \sum X_1 + \beta_1 \sum {X_1}^2 + \beta_2 \sum X_1 X_2 - - - (2)$$

$$\frac{\delta S}{\delta \beta_2} = 0 \Rightarrow \sum (Y - \alpha - \beta_1 X_1 - \beta_2 X_2)(-X_2) = 0$$

$$\Rightarrow \sum YX_2 = \alpha \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum {X_2}^2 - - - (3)$$

Eliminating ∞

$$\sum yx_1 = \beta_1 \sum {x_1}^2 + \beta_2 \sum x_1 x_2$$

$$\sum yx_2 = \beta_1 \sum x_1 x_2 + \beta_2 \sum {x_2}^2$$

$$\hat{\beta}_1 = \frac{\begin{vmatrix} \sum yx_1 & \sum x_1x_2 \\ \sum yx_2 & \sum {x_2}^2 \end{vmatrix}}{\begin{vmatrix} \sum {x_1}^2 & \sum x_1x_2 \\ \sum x_1x_2 & \sum {x_2}^2 \end{vmatrix}} = \frac{\sum yx_1 \sum {x_2}^2 - \sum x_1x_2 \sum yx_2}{\sum {x_1}^2 \sum {x_2}^2 - (\sum x_1x_2)^2}$$

$$\hat{\beta}_2 = \frac{\begin{vmatrix} \sum yx_1 & \sum {x_1}^2 \\ \sum yx_2 & \sum x_1x_2 \end{vmatrix}}{\begin{vmatrix} \sum {x_1}^2 & \sum x_1x_2 \\ \sum x_1x_2 & \sum {x_2}^2 \end{vmatrix}} = \frac{\sum yx_1 \sum x_1x_2 - \sum yx_2 \sum {x_1}^2}{\sum {x_1}^2 \sum {x_2}^2 - (\sum x_1x_2)^2}$$

Substituting Kx_1 for x_2 we obtain

$$\hat{\beta}_1 = \frac{K^2(\sum yx_1)(\sum {x_1}^2) - K^2(\sum x_1y)(\sum {x_1}^2)}{K^2(\sum {x_1}^2)^2 - K^2(\sum {x_1}^2)^2} = \frac{0}{0}$$

And

$$\hat{\beta}_2 = \frac{K(\sum x_1y)(\sum {x_1}^2) - K(\sum x_1y)(\sum {x_1}^2)}{K^2(\sum {x_1}^2)^2 - K^2(\sum {x_1}^2)^2} = \frac{0}{0}$$

Therefore, the parameters are indeterminate: there is no way of finding separate values of each coefficient.

Proof of (b):

If $r_{xi,xj} = 1$ the standard errors of the estimates become infinitely large. In the two variable model,

$Y = \alpha + \beta_1X_1 + \beta_2X_2 + u$

If X_1 and X_2 are perfectly correlated ($X_2 = KX_1$) the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be

$$Var(\hat{\beta}_1) = \sigma u^2 \frac{\sum {x_2}^2}{\sum {x_1}^2 \sum {x_2}^2 - (\sum x_1x_2)^2}$$

And

$$Var(\hat{\beta}_2) = \sigma u^2 \frac{\sum {x_1}^2}{\sum {x_1}^2 \sum {x_2}^2 - (\sum x_1x_2)^2}$$

Substituting KX_1 for X_2 we obtain,

$$Var(\hat{\beta}_1) = \frac{\sigma u^2 K^2 \sum {x_2}^2}{K^2(\sum {x_1}^2)^2 - K^2(\sum {x_1}^2)^2}$$

$$= \frac{\sigma u^2 \sum {x_1}^2}{0} = \infty$$

And

$$Var(\hat{\beta}_2) = \frac{\sigma u^2 \sum {x_1}^2}{K^2((\sum {x_1}^2)^2 - (\sum {x_1}^2)^2)}$$

$$= \frac{\sigma u^2 \sum {x_1}^2}{0} = \infty$$

Thus, the variance of the estimates become infinite unless $\sigma u^2 = 0$.

On the other hand

In the presence of high but imperfect multicollinearity there is no exact linear relationship among the explanatory variables.

The variance and covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ are given by,

$$Var(\hat{\beta}_1) = \frac{\sigma u^2}{\sum x_1{}^2(1 - r_{12}{}^2)}$$

$$Var(\hat{\beta}_2) = \frac{\sigma u^2}{\sum x_2{}^2(1 - r_{12}{}^2)}$$

And

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{r_{12}{}^2 \sigma u^2}{(1 - r_{12}{}^2)(\sum x_1{}^2)(\sum x_2{}^2)}$$

$r_{12}{}^2$ is the squared correlation coefficient between x_1 and x_2 and $0 < r_{12}{}^2 < 1$

So, as $r_{12}{}^2$ tend to 1, the variance of the estimates increases. Therefore, the estimates are biased.

3.8.3. Partial and Marginal Significance of Regressors:

Partial Significance:

Suppose a population regression function is $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

Where, Y_i is the dependent variable and X_{2i} & X_{3i} are explanatory variables and u_i is stochastic disturbance term.

The coefficient β_2 and β_3 are called partial regression coefficients. The significance of the partial regression coefficients are as follows-

β_2 measures the change in mean value of Y per unit change in X_2 holding the value of X_3 is constant. Like wise β_3 measures the change in mean value of Y per unit change in X_3 holding the value of X_2 constant.

Marginal Significance:

Suppose $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$

The explanatory variables X_2 and X_3 are statistically significant on the basis of the 't' test, on the basis of 'F' test collectively if the both the regressors have significant effect on the explained variable Y. now if we add another variables to this model and assess its significance and test whether the addition of the variable to the model increases ESS (R^2) significantly in relation to the RSS. This is called marginal significance of an explanatory variable.

3.8.4. Solution for the Incidence of Multicollinearity:

If multicollinearity has serious effects on the coefficient estimates of important factors one should adopt one of the following corrective solutions-

(I) Application of methods incorporating extraneous quantitative information.

The most important of these methods are-

(a) The method of restricted least square

(b) The method of pooling cross section and time series data (which is actually a special case of restricted least square)

(c) Dasbin's version generalised least squares.

(d) the mixed estimation technique, proposed by Theil and Goldberger.

(II) Increase of the size of the sample.

(III) Substitution of Lagged variables for other explanatory variables in distributed lag models.

(IV) Introducing of additional equations in the model.

(V) Application of the principle component method.

3.8.5. Relevance of zero mean assumption in linear regression:

An important assumption of the ordinary least square is that the mean of the random error term is equal to Zero, $i.e. E(u) = 0$ when the regression equation is $Y = \alpha + \beta X + u$

Where, $Y =$ dependent variable

$X =$ independent variable

$u =$ random error term.

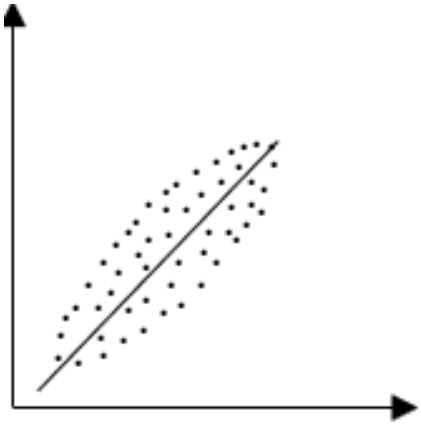
The essence of the zero mean assumption is that we take it as axiomatically true that the positive and negative values of u have a sum equal to zero. This assumption is imposed due to stochastic nature of economic relationships which otherwise it would be impossible to estimate with the common rules of mathematics. By assuming that u has a zero mean, the expected (mean) value of Y is

$E(Y) = (\alpha + \beta X) + E(u)$

$= \alpha + \beta X \quad [Since\ E(u) = 0]$

It can be interpreted as the linear relationship which 'on the average' holds between X & Y.

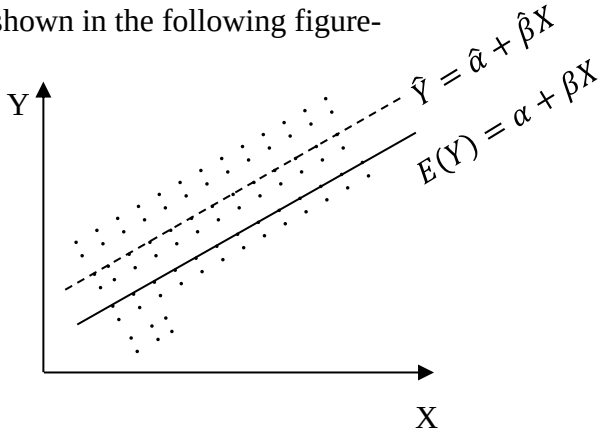
It implies that the observations of Y and X must be scattered around the line in a random day.



Only under this assumption will our estimate line be a good approximation of the true line.

<u>If this assumption is violated</u>,

The alternative possible assumptions are either $E(u) > 0\ or,\ E(u) < 0$ (the mean of u is positive or negative). Then this would imply that the observations of Y and X would lie above or below the true line, shown in the following figure-



The estimated line would be $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ which is obviously not a good approximation to the true line $E(Y) = \alpha + \beta X$. The estimate of true line is achieved by fitting a regression line which passes through the observations, if the true line below or above the observations which would imply that

$E(u) > 0 \; or \; E(u) < 0$ respectively. This is the reason why we assume $E(u) = 0$ at the outset of our estimation procedure.

3.8.6. Relevance of the assumption that X is non stochastic:

Let us consider the classical linear regression model is $Y_i = \alpha + \beta X_1 + u_i$

Here we assume that X is non stochastic, it implies that the values of explanatory variables are fixed in repeated samples.

In other words, the explanatory variables has a limited randomness.

The error term u_i may be occurring due to observational error. We are also assuming that the disturbances u_i is related to Y not only to X, i.e. X_i & u_i are uncorrelated. We know that even though 'X' has no within sampling fluctuation but there might be fluctuations across samples. That across sampling fluctuation of X may relate with the sampling fluctuation of error term. So there might be some correlation between X_i & u_i.

The non-stochastic specification implies that

(I) There is a one-way dependency relation and that dependency may not be perfectly linear.

(II) Y is dependent not only on X but also on some other variables.

(III) There exists observational error in Y and not in X.

(IV) The variability of X is independent of U.

(V) The disturbance term conditional upon is identically & symmetrically or is at least uncorrelated way distributed with identical mean zero. Where identically distribution implies homoscedasticity and uncorrelated distribution implies non-auto correlation.

If this assumption is violated i.e. if we draw an infinite number of sample for a fixed set of values of X, then OLS estimators are biased even asymptotic and BLUE property not satisfied.

3.8.7. Relevance of the least square method in classical linear regression:

There are various econometric methods that can be used to derive estimates of the parameters of economic relationships from statistical observations. The relevance of the least square method in classical linear regression model is given below-

Firstly, the parameters estimate obtained by ordinary least squares have some optimal properties –

(1) u_i is normally distributed with mean zero and constant variance σu^2

i.e. $u_i \sim N(0, \sigma u^2)$

(2) the explanatory variables are non-stochastics.

(3) Non-auto correlation assumption.

(4) Homoscedasticity assumption.

(5) There is no multicollinearity problem.

Secondly the computational procedure of OLS is fairly simple as compared with other econometric techniques and the data requirements are not excessive.

Thirdly the least square method has been used in a wide range of economic relationships with fairly satisfactory result and despite the improvement of computational equipment and of statistical information which facilitated the use of other more elaborate econometric techniques, OLS is still one of the most commonly employed methods in estimating relationships in econometric models.

Forthly, the mechanise of least square are simple to understand.

Fifthly OLS is an essential component of most other econometric techniques. In fact, all other techniques involve the application of the least squares' method, modified in some respects.

Subunit – 9: Panel Data and Dummy Variable

3.9.1. Example of panel data:

Let us suppose that real gross investment of a company (Y) depends on the real value of the firm (X_2) And real capital stock (X_3).

Suppose we have four companies -I, II, III and IV. Data for each company on the three variables are available for the period 1935 to 1954 i.e. 20 years. Pooling or combining all the 80 observations we can write the Grunfeld invest function as

$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$

$(where, i = 1,2,3,4$

$t = 1, \dots\dots, 20)$

Where, i stands for cross sectional unit and t stands for the time period.

3.9.2. Least square dummy variable:

In fixed effect model when the slope coefficients are constant but the intercept varies across individuals then we can write the model as

$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} - - - (1)$

In literature this is known as fixed effect model, we call it effect because each individual's intercept does not vary over time i.e. it is time invariant. To allow for the intercept to vary across companies we can write the equation (2) in dummy differentiate intercept form

$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$

Where,

$D_{2i} = 1$ if the company is 2nd company

$= 0$, otherwise.

$D_{3i} = 1$ if the company is 3rd.

$= 0$, otherwise.

$D_{4i} = 1$ if the company is 4th.

$= 0$, otherwise.

Since we are using dummies to estimate the fixed effects, in literature it is known as least square dummy variable (LSDV).

3.9.3. Why Panel data is used?

(I) since panel data relate to individuals' overtime, there is much heterogeneity in this data.

(II) Panel data give more information of data, more variability and less collinearity among variables, more degrees of freedom and more efficiency.

(III) Panel data can better detect and measure effects that simply cannot be observed in pure cross section or pure time series data.

(IV) Economies of scale and technological change can be better handled by panel data.

(V) Panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregate.

3.9.4. Limitations of fixed effect model:

Although FEM is easy to apply it has several limitations-

(I) Introducing too many dummy variables reduces the degree of freedom.

(II) With so many variables within the model there is always the possibility of multicollinearity problem.

(III) All the results in the FEM are based on the assumption that $u_{it} \sim N(0, \sigma^2)$ as the index i refers to cross section units and t refers the time series observations, we might have problem of heteroskedasticity and auto correlation.

3.9.5. Give an example of the regression model having a mixture of quantitative and qualitative variable:

The regression with a mixture of quantitative and qualitative variables is called 'ANVOVA' model.

Let us suppose that, $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$ where,

$Y_i =$ the average salary of public-school teachers and $X_i =$ spending on public school per pupils.

Here,

$D_{2i} = 1$ if the state is in north-east.

$= 0$, otherwise.

$D_{3i} = 1$ if the state is in south.

$= 0$ otherwise.

3.9.6. Interaction Effect:

Let $Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$

Suppose $Y =$ Hourly wage in Rs.

$X =$Education

$D_2 = 1$, if female

$= 0$, otherwise/male

$D_{3i} = 1$, if non white

$= 0$, otherwise.

Different effect of gender dummy is constant across races. Similarly, differential effect of race dummy is constant across sexes. Now question is a female non-white worker may earn lower wages than a male white worker i.e. there may be interaction between the two qualities variables.

i.e. $Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$

where, $\alpha_2 =$ the differential effect of being a female.

$\alpha_3 =$ differential effect of being non-white

$\alpha_4 =$ differential effect of being female non-white.

3.9.7. Advantages of Dummy variable regression:

The advantages of dummy variable are given below-

(I) We need to run only the single regression equation.

(II) The dummy variable approach not only tells us whether the two regressions are different but also points out sources of differences, i.e. whether it is due to intercept or slope or both.

$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$

If α_2 is statistically significant then we accept the hypothesis that the two regressions have not the same intercept. If β_2 is statistically significant we reject the hypothesis that the two regression lines are parallel. The test of stability can be made by the usual F-test.

3.9.8. Dummy variable is used in seasonal analysis:

A time series can be represented by the following equation –

$TS = s + c + t + u$

Where $Ts =$ time series

$s =$ seasonal component

$c =$ cyclical component

$t =$ trend component

$u =$ random component

Often it is desirable to remove the seasonal factor from the time series. The process of removing the seasonal component from time series is known as deseasonalization. Suppose we have quarterly data of sells of refrigerators from 1978 to 1995.

$Y_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_t$

$D_1 = 1$, if 1st quarter.

$= 0$, otherwise.

$D_2 = 1$, if 2nd quarter.

$= 0$, otherwise.

$D_3 = 1$, if 3rd quarter.

$= 0$, otherwise.

$D_4 = 1$, if 4th quarter.

$= 0$, otherwise.

Subunit – 10: Time Series Analysis

3.10.1. Time-series processes:

A ***time series*** is a sequence of observations on a variable taken at discrete intervals in time. We index the time periods as 1, 2, …, *T* and denote the set of observations as

$$(y_1, y_2, ..., y_T) .$$

We often think of these observations as being a finite sample from a ***time-series stochastic process*** that began infinitely far back in time and will continue into the indefinite future:

Each element of the time series is treated as a random variable with a probability distribution. As with the cross-section variables of our earlier analysis, we assume that the distributions of the individual elements of the series have parameters in common.

we may assume that the variance of each y_t is the same and that the covariance between each adjacent pair of elements $\text{cov}(y_t, y_{t-1})$ is the same. If the distribution of y_t is the same for all values of *t*, then we say that the series *y* is ***stationary***, which we define more precisely below. The aim of our statistical analysis is to use the information contained in the sample to infer properties of the underlying distribution of the time-series process (such as the covariances) from the sample of available observations.

3.10.2. White noise

The simplest kind of time-series process corresponds to the classical, normal error term of the Gauss-Markov Theorem. We call this kind of variable ***white noise***. If a variable is white noise, then each element has an identical, independent, mean-zero distribution. Each period's observation in a white-noise time series is a complete "surprise": nothing in the previous history of the series gives us a clue whether the new value will be positive or negative, large or small.

Formally, we say that ε is a white-noise process if

$$E(\varepsilon_t) = 0, \forall t,$$

$$\text{var}(\varepsilon_t) = \sigma^2, \forall t,$$

$$\text{cov}(\varepsilon_t, \varepsilon_{t-s}) = 0, \forall s \neq 0.$$

Some authors define white noise to include the assumption of normality, but although we will usually assume that a white-noise process ε_t follows a normal distribution we do not include that as part of the definition. The covariances in the third line of equation (1.1) have a special name: they are called the ***autocovariances*** of the time series. The *s*-order autocovariance is the covariance between the value at time *t* and the value *s* periods earlier at time *t* – *s*. Fluctuations in most economic time series tend to persist over time, so elements near each other in time are correlated. These series are ***serially correlated*** and therefore cannot be white-noise processes. However, even though most variables we observe are not simple white noise, we shall see that the concept of a white-noise process is extremely useful as a building block for modeling the time-series behavior of serially correlated processes.

3.10.3. Unit Root Test (Dickey–Fuller Test)

A simple AR (1) model is

$y_t = py_{t-1} + u_t$ Where, y_t is the variable of interest, t is the time index, p is a coefficient, and u_t is the error term. A unit root is present if p=1. The model would be non-stationary in this case.

The regression model can be written as $\Delta y_t = (p-1)y_{t-1} + u_t = \delta y_{t-1} + u_t$

where Δ is the first difference operator. This model can be estimated and testing for a unit root is equivalent to testing $\delta=0$ (where $\delta \equiv$ p-1). Since the test is done over the residual term rather than raw data, it is not possible to use standard t-distribution to provide critical values. Therefore, this statistic t has a specific distribution simply known as the Dickey–Fuller table. There are three main versions of the test:

1. Test for a unit root: $\Delta y_t = \delta y_{t-1} + u_t$
2. Test for a unit root with drift: $\Delta y_t = a_0 + \delta y_{t-1} + u_t$
3. Test for a unit root with drift and deterministic time trend: $\Delta y_t = a_0 + a_1 t + \delta y_{t-1} + u_t$

Each version of the test has its own critical value which depends on the size of the sample. In each case, the null hypothesis is that there is a unit root, $\delta=0$. The tests have low statistical power in that they often cannot distinguish between true unit-root processes ($\delta=0$) and near unit-root processes (δ is close to zero). This is called the "near observation equivalence" problem.

The intuition behind the test is as follows. If the series y is stationary (or trend-stationary), then it has a tendency to return to a constant (or deterministically trending) mean. Therefore, large values will tend to be followed by smaller values (negative changes), and small values by larger values (positive changes). Accordingly, the level of the series will be a significant predictor of next period's change, and will have a negative coefficient. If, on the other hand, the series is integrated, then positive changes and negative changes will occur with probabilities that do not depend on the current level of the series; in a random walk, where you are now does not affect which way you will go next.

It is notable that $\Delta y_t = a_0 + u_t$ may be rewritten as $y_t = y_0 + \sum_{i=1}^{t} u_i + a_0 t$ with a deterministic trend coming from $a_0 t$ and a stochastic intercept term coming from $y_0 + \sum_{i=1}^{t} u_i$,, resulting in what is referred to as a *stochastic trend*.[2]

There is also an extension of the Dickey–Fuller (DF) test called the augmented Dickey–Fuller test (ADF), which removes all the structural effects (autocorrelation) in the time series and then tests using the same procedure.

3.10.4 Explain the concept of cointegration in time series analysis:

The econometric use of the term equilibrium refers to any long run relationship among non-stationary variables. Cointegration does not require that the long run relationship be generated by market forces or by the behavioural rules of individuals. The concept of cointegration is introduced by Eagle and Granger (1987).

Their formal analysis begins by considering a set of economic variables in long-run equilibrium when

$\beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_n x_{nt} = 0$

Where vector $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$

And vector $X_t = (x_{1t}, x_{2t}, \ldots x_{nt})$

The components of the vector x_t are said to be cointegrated of order (d,b) denoted by

$x_t \sim CI(d, b)$ if –

(I) all components of x_t are integrated of order d.

(II) There exists a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ such that the linear combination

$\beta x_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_n x_{nt}$ is integrated of order $(d - b)$ where $b > 0$.

Cointegration refers to variables that are integrated of the same order of course. This do not imply that all integrated variables are cointegrated. If two variables are integrated of different orders, they cannot be cointegrated.

Suppose x_t is $I(d_1)$ i.e. x_{1t} is integrated of order d_1 and x_{2t} is integrated of order d_2 i.e. $I(d_2)$ where, $d_2 > d_1$. In such case there cannot be a cointegration relation between x_{1t} and x_{2t}.

However, if x_{1t} and x_{2t} are $CI(2,1)$ there exist a linear combination of the form $\beta_1 x_{1t} + \beta_2 x_{2t}$ which is integrated of order $I(1)$. It is possible that this linear combination of x_{1t} and x_{2t} is cointegrated with $I(1)$ variable. Lee and Granger use the form multi-cointegration to refer this type of circumstances.

3.10.5 Distinguish between Auto Regressive (AR) and Moving Average (MA)

Let us consider a model as,

$y_t = a + by_{t-1} + u_t$ where, u_t is white noise disturbance term. This scheme is known as a first order autoregressive scheme, usually denoted by $AR(1)$. It can be interpreted as the regression of y_t on itself lagged one period. Similarly, if the model is $y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + u_t$ where, u_t is also white noise disturbance term. Then this scheme is known as a second order autoregressive scheme and is denoted by $AR(2)$.

And in general,

$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_q y_{t-q} + u_t$ is a qth order autoregressive scheme and is denoted by $AR(q)$.

On the other hand if we consider a model as,

$y_t = a + b_1 u_t + b_2 u_{t-1}$ where, u_t is white noise disturbance term.

Here y_t is equal to a constant term plus a moving average of current and past error terms. y_t follows first order moving average scheme and is denoted by $MA(1)$.

If y_t follows –

$y_t = a + b_1 u_t + b_2 u_{t-1} + b_3 u_{t-2}$, it is an $MA(2)$ process and in general if $y_t = a + b_1 u_t + b_2 u_{t-1} + b_3 u_{t-2} + \cdots + b_q u_{t-(q-1)}$, it is called $MA(q)$ process.

(II) AR process is simply a linear combination of lagged dependent variables. On the other hand, MA process is simply a linear combination of white noise error terms.

(III) In case of $AR(m)$ autocorrelation converges to a particular value but partial auto correlation will be cut off after 'm' lags.

On the other hand, in case of $MA(K)$ process autocorrelation will cut off after 'K' lags but partial autocorrelation will be damped in nature, because $MA(K) \longrightarrow AR(\infty)$.

(IV) We have estimated the order of AR process from the cut-off point of the partial autocorrelation but we have also estimated the order of MA process from the cut-off point of autocorrelation.

(V) Under certain condition we may find that $AR(1)$ transform to a $MA(\infty)$ process and $MA(1)$ transform to a $AR(\infty)$ proccess.

(VI) In one sense we can say $AR(m)$ process as $ARMA(m, 0)$ process and on the other hand $MA(m)$ process as $ARMA(0, m)$ process.

(VII) $AR(P)$ process can also be written as $ARIMA(P, 0,0)$ and $MA(P)$ process can be written as $ARIMA(0,0, P)$.

(VIII) AR process can also be written as $|b| < 1$, it is a stationary but MA process can also be nonstationary.