

PLUS ONE STATISTICS



**STUDY
MATERIAL**



COMPILED BY:

Dr. Vidhya G Nair

HSST STATISTICS

GHSS FOR GIRLS NADAKKAVU, KOZHIKODE



HIGHER SECONDARY NATIONAL SERVICE SCHEME

CHAPTER 1

STATISTICS- SCOPES AND DEVELOPMENT

Statistics have a crucial role in developing the world.

The word Statistics have been derived from Latin word "Status" or the Italian word "Statista". The meaning of these words is "Political State" or a Government.

Sir Ronald Aylmer Fisher is known as father of modern statistics.

Defintion of Statistics by Croxton and Cowden- "Statistics can be defined as the collection, presentation and interpretation of numerical data."

Functions of Statistics

1. Simplifies complexity.
2. Presents facts in a definite and precise form.
3. Provides comparison.

4. Enlarges human knowledge and experience.
5. Helps in formulating policies, testing hypothesis and forecasting future events.

Scope of Statistics

1. Planning
2. Economics
3. Industry
4. Mathematics
5. Psychology and Education
6. Management Studies

Some applied areas of Statistics are

- Actural Science
- Biostatistics
- Agricultural Statistics

Even if there is a wide application of Statistics in day to day life, Statistics has also some limitations. The misuse of Statistics is the main cause of discredit to this science.

Official Statistics

The Ministry of Statistics and Programme Implementation (MOSPI) has two wings, Statistics and Programme Implementation.

The Statistics Wing called National Statistical Office (NSO) consists of the Central Statistical Office (CSO), the Computer Centre and the National Sample Survey Office (NSSO).

CSO coordinates all statistical activities in the country.

NSSO is the largest organisation in India, conducting regular socio-economic surveys.

It has four divisions- SDRD, FOD, DPD and CPD.

Indian Statistical Institute (ISI)

ISI is a unique institution devoted to the research, teaching and application of Statistics. It is founded by Prof. Prasanta Chandra Mahalanobis in Kolkata. He is known as father of Indian statistics.

29th June, birth day of P C Mahalanobis, is celebrated as National Statistics Day.

ISI publishes a journal 'Sankhya'.

Economics and Statistics Department

The Directorate of Economics and Statistics is the nerve centre of the Kerala State statistical system and it is the nodal agency coordinating all statistical activities in the state.

CHAPTER 2

COLLECTION OF DATA

Data means any measurement, result, fact or observation which gives information.

Data collection is the systematic gathering of data for a particular purpose from various sources.

Statistical investigation includes collection, classification, presentation, analysis and interpretation of data according to well defined procedures. The person authorized to make investigation is known as investigator.

The investigators depute some persons to collect the data from the field. These persons are known as Enumerators. The process is known as Enumeration.

A population consists of all elements, individuals, items or objects whose characteristics are being studied. If data are collected from each and every unit of the population, the investigation is called census.

The representative part of the population is known as sample. The method of collecting data from the sample is known as sampling or sample survey.

The statistical survey may be either by census method or by sampling method.

The factors which can vary from one object to another are called variables.

The variable which can not be numerically measured is called qualitative variable. The variable which are numerically measured is called quantitative variable.

If the variable takes specific values only, it is called discrete variable. A continuous variable takes any value within the defined range of variables.

A nominal scale of measurement is used to name categories such as gender, nation etc. In the ordinal scale of measurement, we can put an order to the data according to the relation among the values of variables. The data regarding a quantitative variable is a cardinal data.

Based on the sources of collection, statistical data may be classified as primary and secondary data.

Primary data collected by the investigator for the first time for his/her own purpose.

Data obtained from existing sources which may be published or unpublished are known as secondary data.

Methods of primary data collection are

1. Direct personal interview
2. Indirect oral investigation
3. Direct observation

4. Telephone interview
5. Mailed questionnaires and schedules
6. Focus group discussion

Questionnaires and schedules are series of questions arranged in a logical order so as to collect information for a specified purpose. A questionnaire is usually mailed by post or email to selected informants. Enumerator approaches personally to the informant and collects information from them.

Sources of secondary data

- Government publications
- Office records in panchayats, municipalities etc.
- Survey reports of various research organizations
- Survey reports in journals, newspapers and other publications
- Websites

CHAPTER 3

CLASSIFICATION AND TABULATION

Classification of data is the process of grouping the data according to some characteristics .

Types of classification

- Qualitative classification
- Quantitative classification
- Chronological classification
- Geographical classification

Tabulation of data is the method of representing data with the help of a statistical table.

In one way classification only one characteristic is considered for classification. The

table used for one way classification is one way table. If we consider two characteristics at a time for classification for data, it is termed as two way classification of data and the table is two way table.

The number of repetitions of a particular observation in a series is called frequency of the observation.

The series of observations in which items are listed individually is called raw data or individual series. A discrete frequency table is that series in which data are presented in a way that exact measurements of units are clearly shown.

Frequency tables with classes and corresponding frequencies are known as continuous frequency table.

If the lower limit of the first class or upper limit of the last class is not specified, then it is called open end class.

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Total frequency}} = \frac{f}{N}$$

$$\text{Percentage frequency} = \frac{\text{Frequency}}{\text{Total frequency}} \times 100 = \frac{f}{N} \times 100$$

Sum of the relative frequencies is 1 and sum of the percentage frequencies is 100

The number of observations less than or equal to a particular value is called less than cumulative frequency of that value.

The number of observations greater than or equal to a particular value is called greater than cumulative frequency or more than cumulative frequency of that value.

If only one characteristic of the sampling units is measured for the study, it is called uni variate data.

If two characteristics are measured simultaneously from each unit, it is known as bi variate data. The frequency distribution of a bi variate data is called bi variate frequency table

CHAPTER 4

DIAGRAMS AND GRAPHS

Diagrams and Graphs are the methods for simplifying the complexity of quantitative data. Diagrams and Graphs are more attractive and impressive. **Diagrams**

Commonly used diagrams are

1. Bar diagrams
2. Pie diagram

Bar Diagram

There are four types of bar diagram.

1. Simple Bar Diagram
2. Multiple Bar Diagram
3. Sub divided Bar Diagram(Component Bar Diagram)

4. Percentage Bar Diagram

In constructing a pie diagram, the first step is to prepare the data so that the various component values can be transposed into corresponding degrees on the circle using the formulae

$$\text{Angle} = \frac{\text{Item frequency}}{\text{Total frequency}} \times 360$$

The most commonly used graphs for representing a frequency distribution are

1. Histogram
2. Frequency Polygon
3. Frequency Curve
4. Ogives (cumulative frequency curves)
5. Scatter plot

There are two types of ogives. Less than ogive and greater than ogive.

Scatter plot is used to represent a bi variate data.

CHAPTER 5

CENTRAL TENDENCY

The property of the observations in a data to cluster or concentrate around a value is known as Central Tendency.

Measures of central tendency (averages) are the values which gives an idea about the concentration of observations in the central part of the distribution.

Desirable properties of a good average

1. Simple and rigid definition.
2. Simple to understand and easy to calculate.
3. Based on all the observations.

4. Least affected by extreme values.
5. Least affected by fluctuations of sampling.
6. Capable of further mathematical treatment.

The various measures of central tendencies are

1. Arithmetic Mean(AM)
2. Median
3. Mode
4. Geometric Mean(GM)
5. Harmonic Mean(HM)

Arithmetic Mean Arithmetic mean which is also known as mean is denoted by \bar{x}

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{No. of observations}}$$

(i) For a raw Data

$$\bar{x} = \frac{\sum x}{n} \quad (5.1)$$

where n is the no. of observations.

(ii) For a Discrete Frequency Distribution

$$\bar{x} = \frac{\sum fx}{N} \quad (5.2)$$

where $N = \Sigma f$ is the total frequency.

(ii) For a Continuous Frequency Distribution

$$\bar{x} = \frac{\Sigma fx}{N} \quad (5.3)$$

where $N = \Sigma f$ is the total frequency and midpoint of the class is taken as the value of x .

Mathematical properties of AM

1. $\Sigma(x - \bar{x}) = 0$
2. $\Sigma(x - a)^2$ is least when $a = \bar{x}$
3. If each observation is increased by 'a', then the mean is also increased by 'a' and each observation is decreased by 'a', then the mean is also decreased by 'a'.
4. If each observation is multiplied by p , $p \neq 0$, then the mean of the new observations is $p\bar{x}$.

Weighted AM

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w}$$

Combined AM

If \bar{x}_1 and \bar{x}_2 are the means of two groups of n_1 and n_2 observations respectively, the mean of the combined group of $n_1 + n_2$ observations are given by $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

Median

Median is the value of middle most observation in the data.

Median for raw data

When observation are arranged in ascending or descending order of magnitude, Median is the $(\frac{n+1}{2})^{th}$ item in the data where n is the no. of observations.

Median for discrete frequency distribution

Median is the observation having cumulative frequency $\frac{N+1}{2}$, when the observations are arranged in ascending order.

Median for continuous frequency distribution

Median class is the class where $\frac{N}{2}^{th}$ observation lies.

$$Median = l + \frac{(\frac{N}{2} - m)c}{f}$$

where l is the lower limit of the median class, c is the class interval of the median class, f is the median class and m is the cumulative frequency of the class preceding the median class.

Median can be located graphically using ogives.

Mode

Mode of a data is defined as the value that is repeated most often in the data. Mode for a raw data

Mode of a raw data is the observation having the maximum frequency in the data.

Mode for a discrete frequency distribution

Mode is the observation having the highest frequency. Mode for a continuous frequency distribution

Modal class is the class having highest frequency. $Mode = l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$

Where l is the lower limit of the modal class, f_1 is the frequency of the modal class, f_0 is the frequency of the preceding class to the modal class, f_2 is the frequency of the succeeding class to the modal class and c is the class interval of the modal class. Mode can also be locate using histogram.

Empirical relationship between mean, median and mode is:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

Or

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

Geometric Mean (GM)

GM is the n^{th} root of the product of n observations in the data set.

$$GM = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n} = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}}$$

Harmonic Mean (HM)

HM is the reciprocal of the AM of the reciprocals of the given observations.

$$HM = \frac{n}{\sum \frac{1}{x}}$$

Relations among AM, GM and HM

1. For a set of positive values, $AM \geq GM \geq HM$.

When all the observations are the same ,then $AM = GM = HM$

$$2. (GM)^2 = AM \times HM$$

$$\text{or } GM = \sqrt{AM \times HM}$$

Quartiles, deciles and percentiles are the partition values which divide data into several equal parts.

Quartiles divide a data into four equal parts. There are three quartiles, denoted by Q_1 , Q_2 and Q_3 . Q_2 is the median.

Quartiles for raw data

Arrange the n observations in ascending order of magnitude.

Q_1 = value of $\left(\frac{n+1}{4}\right)^{th}$ item in the series

Q_3 = value of $\frac{3(n+1)}{4}^{th}$ item in the series

Quartiles for discrete frequency distribution

Q_1 = observation having cumulative frequency $\frac{(N+1)}{4}$

Q_3 = observation having cumulative frequency $\frac{3(N+1)}{4}$

where N is the total frequency.

Quartiles for continuous frequency distribution

Prepare cumulative frequency. N be the total frequency. Locate the classes having cumulative frequencies $\frac{N}{4}$ and $\frac{3N}{4}$. These classes are called quartile classes.

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m_1\right)c_1}{f_1}$$

$$Q_3 = l_3 + \frac{(\frac{3N}{4} - m_3)c_3}{f_3}$$

where l_1 and l_3 are the lower limits of quartile classes,

f_1 and f_3 are the frequencies of the quartile classes,

c_1 and c_3 are the class intervals of the quartile classes

and m_1 and m_3 are the cumulative frequencies preceding the quartile classes.

Deciles divide the distribution into ten equal parts and there are 9 deciles. Median is the 5th decile

Percentiles divide a distribution into hundred equal parts. There are 99 percentiles.

Median is the 50th percentile.

A box plot is a graph of a data set that consists of a line extending from the minimum value to the maximum value and a box with lines drawn at the Q_1 , the median, and Q_3 .

CHAPTER 6

DISPERSION

Dispersion is the degree of scatter or variation of the variable about a central value.

Measures of dispersion are

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

Range

Range = Highest value - Lowest value

= H- L

Quartile Deviation(QD)

$$QD = \frac{Q_3 - Q_1}{2}$$

Q_1 and Q_3 are explained in the previous chapter for raw data, discrete frequency distribution and continuous frequency distribution.

Mean Deviation (MD)

MD for raw data

$$MD = \frac{\sum |X - A|}{n}, \text{ where } A \text{ is any average.}$$

MD for discrete frequency distribution

$$MD = \frac{\sum f |X - A|}{N}, \text{ where } A \text{ is any average and } N \text{ is the total frequency.}$$

MD for continuous frequency distribution

$$MD = \frac{\sum f |X - A|}{N}, \text{ where } A \text{ is any average, } N \text{ is the total frequency and } x \text{ is the mid value of the class.}$$

Standard Deviation (SD)

SD for raw data

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

SD for discrete frequency distribution

$$SD \sigma = \sqrt{\frac{\sum f x^2}{N} - \bar{x}^2}, \text{ where } N \text{ is the total frequency.}$$

SD for continuous frequency distribution

$$SD \sigma = \sqrt{\frac{\sum f x^2}{N} - \bar{x}^2}, \text{ where } N \text{ is the total frequency and } x \text{ is the mid value of the class.}$$

Coefficient of variation(CV) and coefficient of QD are relative measures of dispersion.

CV is used to compare the consistency or stability between two or more sets of data.

$$CV = \frac{SD}{Mean} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Covariance is a measure of strength of linear relationship between two variables

Cov(x,y) indicates whether the variables are positively related or negatively related

in a bi variate distribution.

$$\text{Cov}(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \frac{\Sigma xy}{n} - \bar{x} \times \bar{y}$$

CHAPTER 7

SKEWNESS AND KURTOSIS

Skewness means the absence of symmetry in a data set. For a symmetric distribution
Mean= Median= Mode.

There are two types of skewness ‘

1. Positive skeness
2. Negative skewness

For a positively skewed data $Mode < Median < Mean$

For a negatively skewed data $Mean < Median < Mode$.

Measures of skewness

1. Karl Pearsons Coefficient of Skewness

$$S_k = \frac{Mean-Mode}{SD}$$

2. Bowleys coefficient of Skewness

$$S_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

3. Coefficient of Skewness based on moments

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\gamma_1 = \sqrt{\beta_1}$$

μ_3 determines nature of skewness. If $\mu_3 > 0$, the distribution is positively skewed. If $\mu_3 < 0$, the distribution is negatively skewed. If $\mu_3 = 0$, the distribution is symmetric.

Kurtosis

Kurtosis is the measure of peakedness or flatness of the frequency distribution.

There are three types of Kurtosis.

(a) Lepto kurtic

(b) Meso kurtic

(c) Platty kurtic

Measure of Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_2 = \beta_2 - 3$$

If $\beta_2 = 3$ ($\gamma_2 = 0$) the curve is meso kurtic.

If $\beta_2 > 3$ ($\gamma_2 > 0$) the curve is lepto kurtic.

If $\beta_2 < 3$ ($\gamma_2 < 0$) the curve is platty kurtic.

CHAPTER 8

PROBABILITY

Probability is the way of measuring the chances of something to happen.

According to Ya-Lin Chou: "Probability is the science of decision making with calculated risks in the face of uncertainty."

Random experiment

An experiment is called random experiment if it satisfies the following conditions:

- (a) It has more than one outcome.
- (b) It is not possible to predict the outcome in advance.
- (c) It can be repeated any number of times under identical conditions.

Trial: A trial is an action which results in one of several possible outcomes or results.

Sample space: The set of all possible outcome of random experiment is called the sample space. Sample space is usually listed in curly brackets{} and is denoted by S.

Sample point: Each element in the sample space is called a sample point.

Events: An event is a set of outcomes which have some characteristics in common.

Equally likely event: Two or more events which have an equally likely chance or equal probability of occurrence are called equally likely events.

Mutually exclusive (disjoint) events: Events are said to be mutually exclusive if the happening of any one of them excludes the happening of all the others.

Exhaustive events: A set of events is called exhaustive, if all the events together consume the entire sample space.

Basic properties of probability

- The probability is always between 0 and 1.
- The probability of occurrence of an impossible event is 0.
- the probability of something to occur is 1.
- Probability can not be negative.

Mathematical or classical definition of probability

If a random experiment results in 'n' exhaustive , mutually exclusive and equally likely outcomes out of which 'm' are favourable to the occurrence of an

event 'A'. Then the probability of occurrence of A, usually denoted by $P(A)$ is given by

$$P(A) = \frac{\text{Number of favourable cases}}{\text{number of exhaustive cases}} = \frac{m}{n}$$

or

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{N(A)}{N(S)}$$

Algebra of events

- Complement of A: \bar{A} or A^c or A' → not A
- A or B → At least one of the event A or B occurs.
- A and B → Simultaneous occurrence of A and B
- A' and B' → neither A nor B
- A and B' → only A
- (A and B') or (A' and B) → exactly one among A and B

Addition rule of probability

Rule 1

When two events A and B are mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B)$$

Rule 2

When two events A and B are not mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Frequency approach of Probability

If after n repetitions of an experiment, where n is very large, an event A is observed to occur in m of these, then the $P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}$

Given the frequency of the distribution, the probability is computed as

$$P(A) = \frac{\text{frequency of the class}}{\text{total frequency in the frequency distribution}}$$

Axioms on Probability

Axiom 1: Non-negativity

For any event A, $P(A) \geq 0$

Axiom 2: Certainty

If S is the sample space $P(S)=1$

Axiom 3: Additivity

For two mutually exclusive events A_1 and A_2 ,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

CHAPTER 9

CONDITIONAL PROBABILITY

$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)}, \text{ provided } P(B) \neq 0$$

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}, \text{ provided } P(A) \neq 0$$

Multiplication Theorem

$$P(A \text{ and } B) = P(A).P(B/A) = P(B).P(A/B)$$

Independent and dependent events

Two events are said to be independent if the occurrence of one event do not affect the occurrence of the other. Otherwise the events are said to be dependent.

If $P(A/B) = P(A)$ and $P(B/A) = P(B)$, then A and B are independent.

Multiplication Theorem for Independent Events

If A and B are independent, multiplication theorem becomes

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Total Probability Theorem

If A_1 and A_2 are two mutually exclusive and exhaustive events and A is any other event which can occur along with A_1 and A_2 , then total probability theorem states that

$$P(A) = P(A_1)P(A/A_1) + P(A_2)P(A/A_2)$$

If A_1 , A_2 and A_3 are three mutually exclusive and exhaustive events and A is any other event which can occur along with A_1 , A_2 and A_3 , then total probability theorem states that

$$P(A) = P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3)$$

Bayes' Theorem

If A_1 and A_2 are two mutually exclusive and exhaustive events and A is any other event which can occur along with A_1 and A_2 , then

$$P(A_1/A) = \frac{P(A_1)P(A/A_1)}{P(A_1)P(A/A_1) + P(A_2)P(A/A_2)}$$

$$P(A_2/A) = \frac{P(A_2)P(A/A_2)}{P(A_1)P(A/A_1) + P(A_2)P(A/A_2)}$$

This theorem can be extended if we have three mutually exclusive and exhaustive events A_1 , A_2 and A_3 , then

$$P(A_1/A) = \frac{P(A_1)P(A/A_1)}{P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3)}$$

$$P(A_2/A) = \frac{P(A_2)P(A/A_2)}{P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3)}$$

$$P(A_3/A) = \frac{P(A_3)P(A/A_3)}{P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3)}$$

CHAPTER 10

SAMPLING TECHNIQUES

In census, data is collected from each and every unit of the population. The method of collecting data from the sample is known as sampling or sample survey.

Sampling errors are seen in sample surveys due to the fact that only a part of the population is used for enquiry. Sampling errors decrease as sample size increases.

Errors other than sampling errors in a survey are called non sampling errors.

Methods of Sampling

- Non Probability Sampling
- Probability sampling

Non probability sampling

- Convenience sampling
- Judgement sampling
- Quota sampling

Probability sampling

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling
- Multi Stage Sampling

Simple Random Sampling(SRS)

SRS is a probability sampling in which each unit in the population has an equal chance of being included in the sample.

- Simple Random Sampling Without Replacement (SRSWOR)
- Simple Random Sampling With Replacement (SRSWR)

If a population consists of N units and a sample of n units to be taken, the possible number of samples in SRSWOR is NC_n and in SRSWR is N^n