Nature and scope of statistics.

a. Definitions of statistics

i. According to **Prof. Horace Secrist** defines "It is the aggregate of facts affected to mark extent by the multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for the predetermined purpose and placed in relation to each other".

ii. According to **Prof. Boddington** defines "It is the science of estimates and probabilities".

iii. According to **Prof. A.L. Bowley** defines, "It may be called the science of counting and may be called the science of averages".

b. Descriptive and inferential statistics

i. **Descriptive statistics**

It is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way.

Characteristics:

- No conclusion can be drawn beyond the analysis data.
- They are simply a way to describe our data.
- It helps in representing data in a meaningful way, which makes it easier for interpretation.
- The data is represented by tables, graphs, and charts.
- It helps in comparison between different data sets and to spot the smallest and largest values or trends or changes over a period of time.

Limitation:

- They allow us to make summations about the people or objects that we have actually measured.
- We cannot use the data we have collected to generalize to other people or objects.

ii. Inferential statistics

Inferential statistics are techniques and methods that allow us to use samples to make generalizations about the populations from which samples were drawn .

Characteristics:

- It is produced through complex mathematical calculations that allows scientists to infer trends about a larger population based on study of samples taken from it.
- To examine the relationships between variables with a sample and then make generalizations or predictions about those variables are related to larger population
- Scientists used techniques like chi-square and t-test which tell them the probability that the results of their analysis of the sample are representative of the population as a whole.
- It starts with a sample and generalizes to a population

Limitation:

- It depends on data provided so it has always some degree of uncertainty.
- It requires the user to make more educated guesses and as so uncertainty is introduced in the result. .

c. Scope of statistics

i. Statistics help in economic planning

1. All economic plans are formulated on the basis of statistical data.
2. The success of the plan is also evaluated with the help of statistics.
3. Economic problems such as production, consumption, wages, prices profits, unemployment, poverty, etc can be expressed numerically
4. The study of production statistics enables us to make a balance between supply and demands
5. The study of consumption statistics enables us to have some idea about the purchasing capacity and standard of living.

ii. Statistics in business and management

1. It helps in formulating policies regarding business and forecasting future trends.
2. Forecasting regarding the future demand of product, market trends and so on depends upon his experience and proper use of statistical methods.
3. To solve uncertainty statistics is required.

iii. Statistics in administration

1. To collect the information regarding the military and fiscal policies
2. To help collect a huge amount of statistics on different aspects of the people and therefore applying required policies for the welfare of the people.

iv. Statistics in research

1. It helps in carrying out different types of research
2. To examined by the help of appropriate statistical tools

v. Statistics and industry

1. It is a widely used inequality tool.
2. It is used to find out whether the product is conforming to the specifications or not.

d. Limitations and distrusts of statistics

i. It is unable to study qualitative characters

   It studies only quantitative characters (i.e. numerical form) of the given problems instead of qualitative characters (i.e. poverty, honesty, intelligence).

ii. Its results are not accurately correct

iii. It is unable to explain individual items

   It always studies a group of values instead of single observation studies the mass of phenomena and the conclusion on certain characteristics obtained.
   All results are approximately correct in statistics.

iv. It is only one of the methods of studying a given problem

   It cannot be of much help in studying the provided problem unless they are supplemented by evidence.

v. It is liable to be misused.

   The only one who has expert knowledge of statistical methods can scientifically handle statistical data. If misued by incompetent, unskilled and inexperienced person, it may lead to false conclusion.

vi. To many methods to study problems

   Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case.

vii. Result are true only on average

   The results are interpolated for which time series or regression or probability can be used. They are not absolutely true.

viii. Its laws are not exact

It only holds true under certain conditions and cannot be universally applied. So it has less practical utility.

ix. It is easily misinterpreted or distorted

e. Characteristics of Statistics

i. Statistics are aggregate of facts.
ii. Statistics must be numerically expressed
iii. Statistics should be collected for a predetermined purpose
iv. Statistics should be collected in a systematic manner
v. Statistics should be capable of being placed in relation to each other.
vi. Statistics are enumerated or estimated according to reasonable standards of accuracy
vii. Reasonable standard of accuracy should be maintained in collection of statistics.

● Data and its collection.

a. Primary and secondary data

Primary Data - The data collected for the first time by the investigator himself from the field of enquiry is called “primary data”. An investigator can collect using different methods for his own purpose of investigation.

Secondary Data - The data which are initially collected by someone but obtained from some published or unpublished sources.

b. Sources of primary and secondary data

Primary source of data

- Direct personal contact

In this method, the investigators collect data by personally contacting the respondents. The investigators must go to the related field and have to meet and interview each and every individual to collect relevant data for his investigation.

Merits:

1. Information collected by this method is more accurate.
2. Responsibility of the data is very high
3. Extra supplementary information can be obtained which may help in drawing conclusions.
4. Proper language and technique can be adopted according to the nature and status of the informant
5. Sensitive type of questions can be asked at such time only when the informants feel at home with the interviewer.

Demerits:

1. It consumes time and money
2. Accurate information can not be obtained due to personal bias
3. This method is not applicable if the field of investigation is not narrow
4. The data will not be reliable if the interviewer is not well-trained, qualified and intelligent.

- **Indirect oral interviews**

In this method, the information is collected by the interviewer from a third person who is directly or indirectly concerned with the information to be collected. It is adopted when informats are unable to give their information directly.

Merits

1. It saves money, time and labour
2. A wide area can be taken as the field of investigation
3. The opinion and suggestions of experts can be solicited

Demerits

1. Exact information may not be obtained due to the doubtful information given by witnesses.

2. The investigator can twist the facts, if he is a biased person.

- **Informations from correspondents**

  In this method, local agents known as “correspondents” are appointed in the front frields under study. The necessary information is available from those pointed correspondents to the investigator. This method is more suitable if the normations are to be collected from a wide area.

  Merits

  1. Information of wide area can be obtained
  2. Regular information are available
  3. Results are obtained easily and cheaply
  4. Qualitative information is obtained due to the appointment of local agents.

  Demerits

  1. If the appointed local agents are personally biased, information obtained may not be accurate and reliable.

- **Mailed questionnaire**

  In this method, a list of questions relating to the investigation, is prepared and sent by post to the various informants. The informants are requested to fill up the questionnaire and are sent back to the enquiry office with the time mentioned.

  Merits

  1. Real information are obtained as the questionnaires are filled by informant
  2. Information are obtained quickly and cheaply
  3. If the informants are spread over a wide geographical area and the information are to be collected from a wide area, then this method is suitable

  Demerits

  1. This method is suitable only for those regions where people are educated and cooperative.
  2. Most of the questionnaires are not returned back by the informants due to their non-responsibilities.
  3. The results may not be accurate due to the misunderstanding of the given set of questions.

- **Questionnaire sent through enumerators.**

In this method, local agents are appointed and trained properly. Then the questionnaires are sent to the informants through the enumerators but not by post The enumerators visit door to door along with their questionnaires and the information given by the informants are noted. The data collected by the enumerators are sent back to the investigator for further processing of data.

Merits

1. This method is suitable even for uneducated informants
2. The chances of responsibility is high due to the personal contact between enumerator and informant
3. Enumerators can ask some additional questions relating to the investigation.

Demerits

1. It is a very laborious, expensive and time consuming method.
2. This method is not free from the biases of the enumerators.

- **Internet**
- **Telephone**

Secondary source of data

- **Published sources**
    1. Official publication published by
        a. Government such as report of C.B.S
        b. Reports of international organisations such as W.H.O, U.N.O, World Bank
    2. Semi-official publications, published by various local organizations like Nepal Rastra Bank.
    3. Non-governmental publications such as,
        a. Reports of N.G.O and I.N.G.O
        b. Reports of Nepal chamber of commerce
        c. Publications of individual intellectuals and scholars
        d. Financial and economic journals
        e. Reports of trade associations, magazines, etc.

- **Unpublished sources**

All the information may not be published but may be suitable for the purpose of investigation.

a. Reports of private offices
b. Hospital records
c. Material collected by researchers
d. Records of campus administrator,etc

c. Methods of data collection: census method, sample method

Census	**Sampling**
A systematic method that collects and records the data about the members of the population.	Sampling refers to a portion of the population selected to represent the entire group, in all its characteristics.
Enumeration is complete	Enumeration is partial
Study of each and every unit of population	Study of only a handful of units of the population
It is time consuming process	It is a fast process
It is expensive method	It is an economical method
Results are reliable and accurate	Results are less reliable and accurate due to the margin of error in data collection
Error not present	Error depends on the size of the population
Appropriate for population of heterogeneous nature	Appropriate for population of homogeneous nature

d. Compilation of administrative records

If a service decision is challenged in court, the court may ask us to provide an administrative record. We prepare this by using the decision file (it is compiled and maintained by an employee during the decision making process, containing the complete story of our decision making process.)

- ## Classification and tabulation of data

    a. **Classification procedure: qualitative and quantitative classification**

    b. **Tabulation of data**

- ## Diagrammatic and graphical presentation of data

    a. **Importance and limitations**

    b. **Types of diagrammatic representations: Bar diagram, pie diagram, pictogram**

    c. **Types of graphics representations: histogram, frequency polygon, frequency curve, cumulative frequency curve (Ogive)**

- ## Measures of central tendency (MEAN = ϖ)

    a. **Arithmetic mean**

Simple	Weighted			Combined
	frequency	weight	Mid-point (continuous)	
ΣX / n	ΣfX / N	ΣwX / N	Σfm / N	$(n_1\varpi_1 + n_2\varpi_2)/n_1 + n_2$

Deviation method (change of origin and scale)

Step 1: find midpoint from the continuous data.
Step 2: Assume mean(A). d=m-A.
Step 3: h means width, d’ = (m-A)/h.
Step 4: find fd’

Use the following formula to find mean

Mean = $A + (\Sigma fd')/N \ * h$

Corrected Mean

Step 1: Incorrect $\Sigma X \ = \ n\,\varpi \ = \ n * incorrect\ mean$

Step 2: Correct $\Sigma X = Incorrect\ \Sigma X - incorrect\ items + correct\ items$

Step 3: Corrected mean $(\varpi) = Corrected\ \Sigma X/n$

b. Geometric mean

Simple	Frequency Distribution
Antilog $[(1/n)\Sigma logX]$	Antilog $[(1/N)\Sigma f.logX]$

c. Harmonic mean

Simple	Frequency Distribution
$n/\Sigma(1/X)$	$n/\Sigma(f/X)$

Note: AM $\geq$ GM $\geq$ HM.

d. The median: quartiles, deciles and percentiles

Ungrouped Data	Grouped Data(frequency distribution)	Continuous Series
(n+1)th/2	(N+1)th/2	(N/2)th, then L+ $\frac{(N/2)-cf}{f}$ *h **Note: cf will upper than the current cf and f will of current cf**

1. **Quartiles -** Divides the whole series in 4 equal parts.
2. **Deciles -** Divides the whole series in 10 equal parts.
3. **Percentiles -** Divide the whole series in 100 equal parts.

	Quartiles(i=1,2,3)	**Deciles(j=1,2... 9)**	**Percentiles(k= 1,2.....99)**
Ungrouped Data	i*(n+1)th/4	j*(n+1)th/10	k*(n+1)th/100
Grouped Data(frequency distribution)	i*(N+1)th/4	j*(N+1)th/10	k*(N+1)th/100
Continuous Series	(i*N/4)th, then L+ $\frac{(i*N/4)-cf}{f}$ *h **Note: cf will upper than the current cf and f will of current cf**	(j*N/10)th, then L+ $\frac{(j*N/10)-cf}{f}$ *h **Note: cf will upper than the current cf and f will of current cf**	(k*N/100)th, then L+ $\frac{(k*N/100)-cf}{f}$ *h **Note: cf will upper than the current cf and f will of current cf**

e. The mode

Ungrouped Data	**For two or more highest frequency or if continuous gives infinity value**	**Grouped Data(frequency distribution)**	**Continuous Series**
Most repeated value of X.	$M_o = 3M_d - 2\varpi$	Highest frequency value of X.	Highest frequency value of X, then L+ $\frac{fm-f1}{2fm-f1-f2}$ *h **Note: f1 and f2 are up and bottom frequency of highest frequency fm**

f. Relation between mean, median and mode

Symmetrical	**Positive/right skewed**	**Negative/left skewed**

Mean=Median=Mode	Mean>Median>Mode	Mean<Median<Mode

● Measures of dispersion

a. Absolute and relative measures

a. **Absolute measure:** When the measure of dispersion is expressed in original units of a series.

b. **Relative measure**: When the dispersion is expressed in terms of ratio or percentage and is pure numbers independent of the units of the variable.

b. The range

Absolute measure of Range	Relative measure of Range
Largest variable - Smallest variable	Coefficient of range=(L-S)/(L+S)
Highest class variable - Lower class variable	For percentage, Coefficient of range=((L-S)/(L+S))*100%

c. Interquartile range

It is the difference between the extreme quartiles Q_3 and Q_1

$$interquartiles\ Range(IQR) = Q_3 - Q_1$$

d. Quartile deviation

Half of the interquartiles is known as Quartile Deviation.

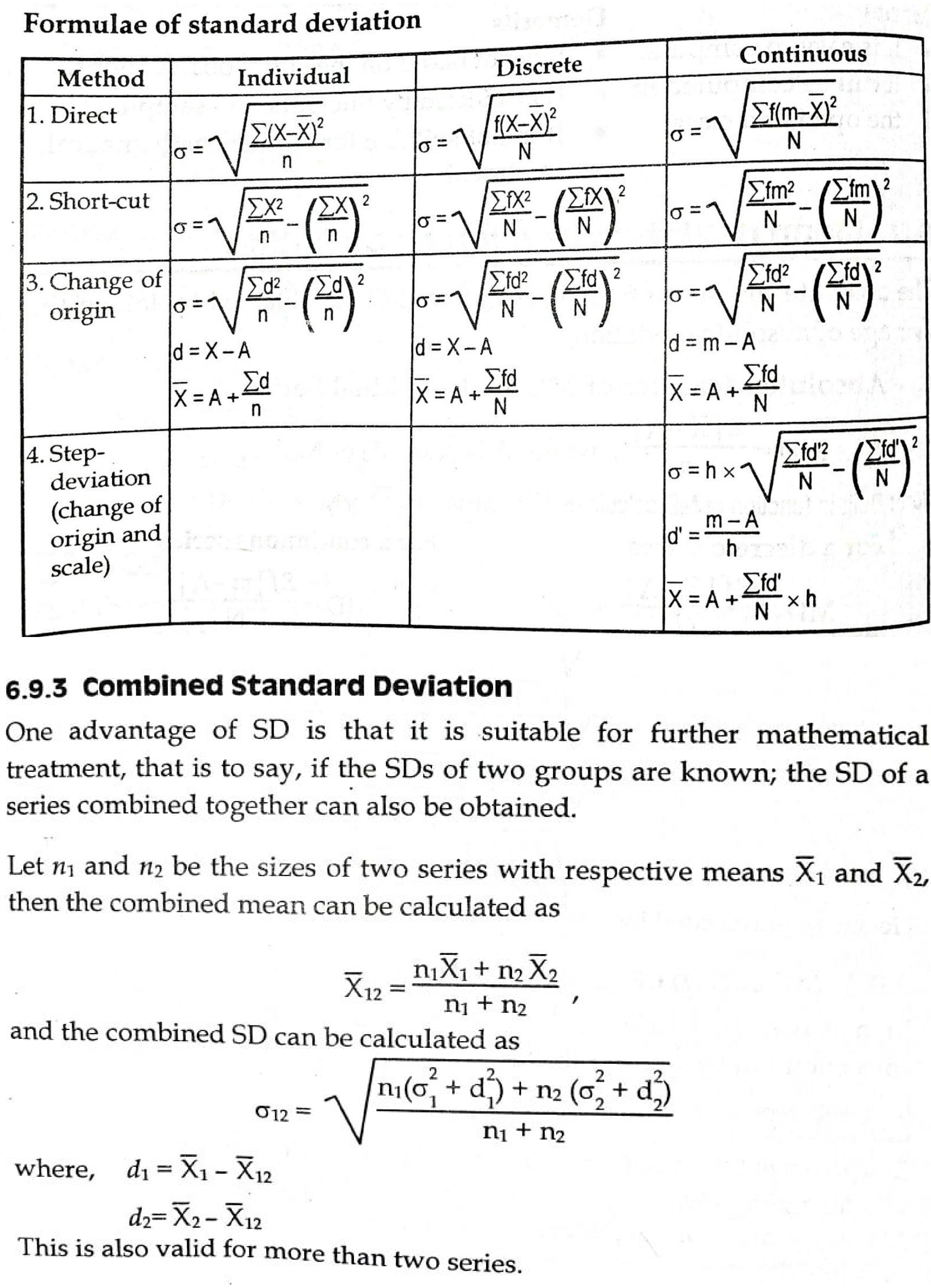
Absolute measure of QD	Relative measure of QD
$Quartiles\ Deviation(QD) = (Q_3 - Q_1)/2$	Coefficient of $QD = (Q_3 - Q_1)/(Q_3 + Q_1)$

Percentile Range: P_{90}-P_{10} Decile Range: D_9-D_1	For percentage, Coefficient of $QD = ((Q_3 - Q_1)/(Q_3 + Q_1)) * 100\%$

e. Mean deviation

Absolute measure of MD		**Relative measure of MD**	
For discrete series	**For continuous series**	**For discrete series**	**For continuous series**
MD= $(\Sigma f\ \vert X - A\vert\ )/N$ Where, X can be mean, median or mode	MD= $(\Sigma f\ \vert m - A\vert\ )/N$ Where, X can be mean, median or mode	Coefficient of MD= $(((1/n)\Sigma\ \vert X - A\vert\ )/A)) * 100\%$	Coefficient of MD= $(((1/N)\Sigma f\ \vert X - A\vert\ )/A)) * 100\%$

f. Standard deviation



Formulae of standard deviation

Method	Individual	Discrete	Continuous
1. Direct	$\sigma = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$	$\sigma = \sqrt{\frac{f(X-\overline{X})^2}{N}}$	$\sigma = \sqrt{\frac{\Sigma f(m-\overline{X})^2}{N}}$
2. Short-cut	$\sigma = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$	$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2}$	$\sigma = \sqrt{\frac{\Sigma fm^2}{N} - \left(\frac{\Sigma fm}{N}\right)^2}$
3. Change of origin	$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$ $d = X - A$ $\overline{X} = A + \frac{\Sigma d}{n}$	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$ $d = X - A$ $\overline{X} = A + \frac{\Sigma fd}{N}$	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$ $d = m - A$ $\overline{X} = A + \frac{\Sigma fd}{N}$
4. Step-deviation (change of origin and scale)			$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$ $d' = \frac{m-A}{h}$ $\overline{X} = A + \frac{\Sigma fd'}{N} \times h$

6.9.3 Combined Standard Deviation

One advantage of SD is that it is suitable for further mathematical treatment, that is to say, if the SDs of two groups are known; the SD of a series combined together can also be obtained.

Let n_1 and n_2 be the sizes of two series with respective means $\overline{X}_1$ and $\overline{X}_2$, then the combined mean can be calculated as

$$\overline{X}_{12} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} ,$$

and the combined SD can be calculated as

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where, $d_1 = \overline{X}_1 - \overline{X}_{12}$

$d_2 = \overline{X}_2 - \overline{X}_{12}$

This is also valid for more than two series.

Solution,

Before correction,

$$\overline{X} = \frac{\Sigma X}{n}$$

or $$20 = \frac{\Sigma X}{20}$$

$\therefore$ $\Sigma X = 400$

Correction in ΣX is made as

Corrected $\Sigma X = 400 + 19 - 9 = 410.$

$$\therefore \quad \text{Corrected Mean} = \frac{\text{Corrected } \sum X}{n}$$

$$= \frac{410}{20} = 20.5$$

Now, $$\sigma^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

Uncorrected ΣX^2 is computed as

$$\sigma^2 = \frac{\Sigma X^2}{20} - \left(\frac{400}{20}\right)^2 \quad \text{or } 25 = \frac{\Sigma X^2}{20} - 400$$

$\therefore$ $\Sigma X^2 = 425 \times 20 = 8500$ (Uncorrected)

Corrected $\Sigma X^2 = 8500 - 9^2 + 19^2$

$= 8500 - 81 + 361 = 8780$

$\therefore$ Corrected $\Sigma X = 410$ and Correct $\Sigma X^2 = 8780$

$$\therefore \quad \text{Corrected } \sigma^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

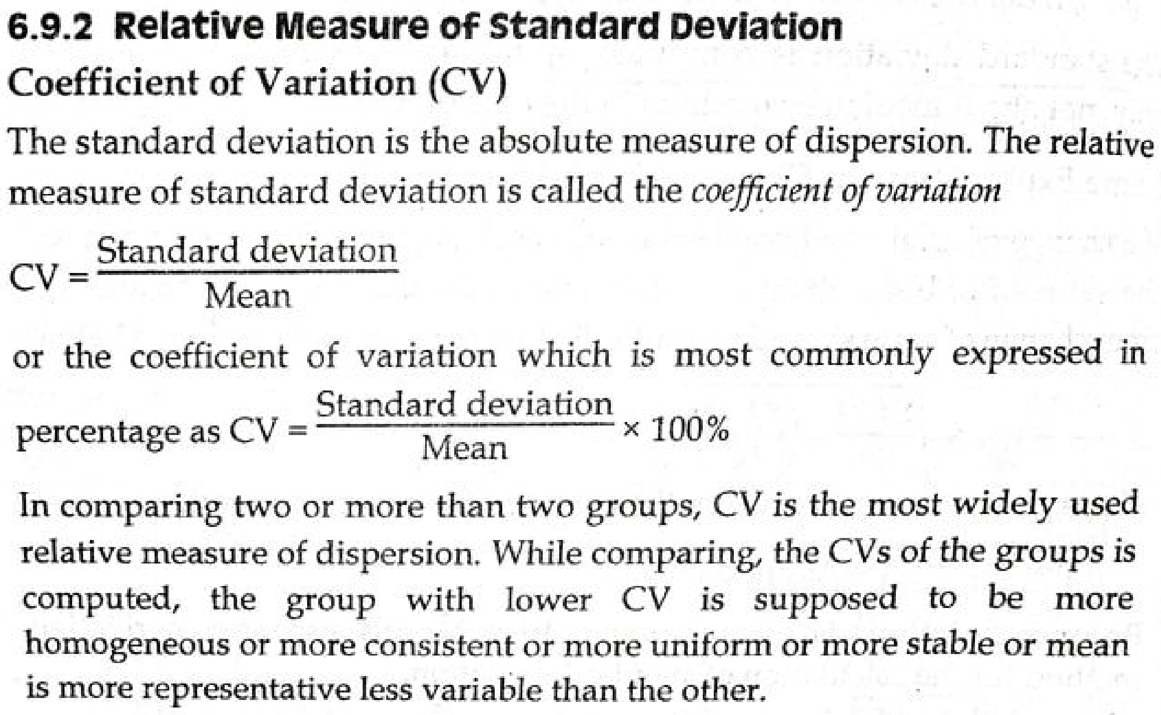
$$= \frac{8780}{20} - \left(\frac{410}{20}\right)^2$$

$$= 439 - 420.25 = 18.75$$

$\therefore$ Corrected $\sigma = \sqrt{18.75} = 4.33$

Hence, the corrected $\overline{X} = 20.5$ and corrected $\sigma = 4.33$

g. Coefficient of variation



6.9.2 Relative Measure of Standard Deviation

Coefficient of Variation (CV)

The standard deviation is the absolute measure of dispersion. The relative measure of standard deviation is called the *coefficient of variation*

$$CV = \frac{\text{Standard deviation}}{\text{Mean}}$$

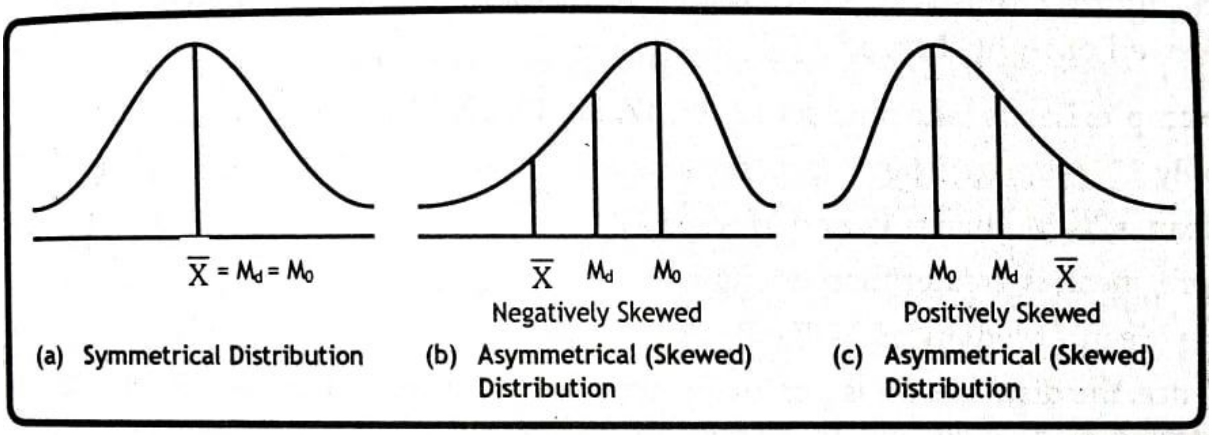
or the coefficient of variation which is most commonly expressed in percentage as $CV = \frac{\text{Standard deviation}}{\text{Mean}} \times 100\%$

In comparing two or more than two groups, CV is the most widely used relative measure of dispersion. While comparing, the CVs of the groups is computed, the group with lower CV is supposed to be more homogeneous or more consistent or more uniform or more stable or mean is more representative less variable than the other.

h. Skewness and kurtosis

Measure of Skewness

Figure 7.1: Different Nature of frequency distributions


$\overline{X} = M_d = M_o$
$\overline{X}$ M_d M_o
Negatively Skewed
M_o M_d $\overline{X}$
Positively Skewed

(a) Symmetrical Distribution (b) Asymmetrical (Skewed) Distribution (c) Asymmetrical (Skewed) Distribution

To find the S_k(Skewness) = Mean - Median

If S_k= 0 OR (Mean=Median) , the distribution is symmetrical.
If S_k< 0 OR (Mean < Median) , the distribution is negatively skewed.

If $S_k > 0$ OR (Mean > Median) , the distribution is positively skewed.

- **Karl Pearson's Coefficient Of Skewness**

  $S_k(P)$ = (Mean-Mode)/SD

  If the mode is ill-defined meaning if two variable's frequencies are high and same. Then, the following formulas are used.

  $S_k(P)$=(3(Mean-Median))/SD

- **Bowley's Coefficient of Skewness**

  $S_{k,}(B)= ((Q_3-Q_2)-(Q_2-Q_1))/((Q_3-Q_2)+(Q_2-Q_1))$

  OR,

  $S_{k,}(B)= (Q_3+Q_1-2Q_2)/(Q_3-Q_1)$

- **Kelly's Coefficient of Skewness**
  $S_{k,}(K)= (P_{90}+P_{10}-2P_{50})/(P_{90}-P_{10})$ --Percentile

  $S_{k,}(K)= (D_9+D_1-2D_5)/(D_9-D_1)$ --Deciles

Measure of Kurtosis

$k=(Q_3-Q_1) / 2*(P_{90}-P_{10})$

If k =0.263, the distribution is normal or mesokurtic.
If k < 0.263, the distribution is platykurtic.
If k > 0.263, the distribution is leptokurtic.

- # Probability
  a. **Preliminaries ???**
  b. **Classicial, empirical, axiomatic approaches of probability theory**

- Classical approaches - The probability is calculated before the experiment is really conducted and the events have really occurred.

  P(A)=m/n , n≥m ; n = possible outcome and m= favourable outcomes to A.

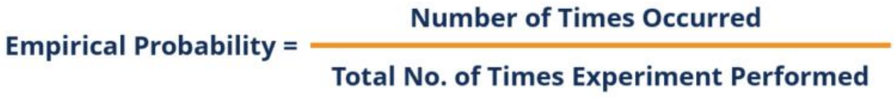
  q=P(A’)=1-(m/n)=1-p ; probability of not happening.

  P(A)+P(A’)=p+q=1;

  Limitations of Classical Approach:

  - If the various outcomes of the trial are not equally likely.
  - If the exhaustic (it includes all possible outcomes of a random experiment) number of cases in a trial is infinite.

- Empirical approaches - empirical probabilities are based upon how likely an event has proven in the past. Thus, they are always estimates. This approach of probability is based upon statistical data. If an experiment is repeated a great number of times under essentially the same condition, then the ratio of the number of repetition of an event to the total number of trials is calculated. This ratio is called relative frequency.

  $$\text{Empirical Probability} = \frac{\text{Number of Times Occurred}}{\text{Total No. of Times Experiment Performed}}$$

  P(E)=lim n->infinite(m/n)

- Axiomatic approaches - probability is obtained on the basis of some properties and axioms.

  Given sample space S of a random experiment, the probability of the occurrence of any event E is defined as a set of function P(E) satisfying the following axioms.

  1. P(E) is real and non-negative.
  2. P(S) = 1 (Certainty axiom)
  3. If A_1 A_2 …… A_n is any finite or infinite sequence of disjoints event of S then,

$\mathbf{P}\left(\bigcup_{i=1}^{n} Ai\right) = \sum_{i=1} P(Ai)$ **(Addition axiom)**

Properties of probability of a event

- **$0 \le$ P(E) ≤ 1 -** should lie between 0 and 1
- **P(E) = 0,** then E cannot occur or impossible event
- **P(E) =1,** then E must occur and is called a certain or sure event.
- **P(E)+P(E')=1,** succession plus failure of probability must be 1.

c. Conditional probability

Conditional probability is the probability of one event occurring with some relationship to one or more other events.

Let A and B are two (dependent or independent) events, then the happening of A under the condition that B has already happened is denoted as.

Formulae: P(A|B) = P(A$\cap$B) / P(B), given P(B)$\neq$ 0

Similarly, the happening of B under the condition that A has already happened is denoted as.

Formulae: P(B|A) = P(A$\cap$B) / P(A), given P(A)$\neq$ 0

P(B|A) is read as P(B) given P(A).

d. Inverse probability

//Inverse probability is the probability of things that are unobserved; or, more technically, the probability distribution of an unobserved variable. It's generally considered **an obsolete term.**

Bayes' theorem is an example of the application of conditional probability. It shows how the probability of past events can be reused to compute the probability of the same events. Therefore, it is also called *revised probability* or *inverse probability*.

Let E_1, E_2, … E_n be *n* mutually exclusive and exhaustive events, with non-zero probability of a random experiment. If A be any arbitrary event of the given sample space of above experiment with *P(A)* > 0 then the probability that it was preceded by the particular event E_i (i= 1, 2, … n) is given by

$$\mathbf{P(E_i|A) = P(A \cap E_i) \;/\; \sum_{i=1}^{n} P(E_i)P(A|E_i) = P(E_i)\;(A|E_i) \;/\; \sum_{i=1}^{n} P(E_i)P(A|E_i)}$$

The **Bayes Rule** formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

e. Probability distribution

A probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume. In other words, the values of the variable vary based on the underlying probability distribution.

f. Mathematical expectation

Random Variable

A variable whose value is determined by the outcome of a random experiment is called a random variable. A random variable is also known as chance or stochastic variable. For example, in a throw of dice, X denotes the number obtained then X is a random variable which can take any one of the values 1,2,3,4,5 or 6, each with equal probability ⅙.

A random variable may be discrete or continuous if the random variable takes on integer values such as, 1,2,3,4… then it is called a discrete random variable.

If the variable takes all possible values within the certain interval then the random variable is called a continuous random variable.

Expected value of discrete random variable

The mean $\bar{x}$ or μ of a probability distribution is the expected value of its random variable. To obtain the expected value of discrete random variable, we multiply each value of the random variable with the corresponding probability of occurrence of that value, and then sum of the product.

Formula: $\mathbf{E(X)= x_1p_1 + x_2p_2 + x_3p_3 + ... + x_np_n = \Sigma xp}$
Where, $x_1,x_2,x_3,...,x_n$ = random variable
$p_1,p_2,p_3,...,p_n$ = probability of random variable
X = expected value

If the probabilities are replaced by relative frequency f/N where, N = Σf then,
$\mathbf{E(X) = \Sigma fx / N = \bar{x} = mean}$

g. Variance of random variable

Variance of X

$$\begin{aligned} Var(X) = \sigma_x^2 &= E[X - E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \\ &= \Sigma X^2P(x) - [\Sigma XP(X)]^2 = \Sigma (X-\bar{x})^2.P(X) \end{aligned}$$

S.D. of X = $\sqrt{Var(X)}$

● Theoretical distribution

a. Introduction

Having started with pre-assumption, if it is possible to mathematically reduce the frequency distribution of certain populations then such distributions are said to be theoretically distributed.

Theoretical distribution provides data on the basis of which the result of actual observation can be accessed. Theoretical distribution provides a sound basis for the decision maker to make rational decisions. The 3 types of theoretical distribution are:

- Binomial distribution
- Poisson distribution
- Normal distribution

b. Binomial distribution and its chief features (without proofs)

Binomial distribution is a widely used probability distribution of a discrete random variable.

The following condition are required for binomial distribution:-

1. Random experiment should be conducted “n” number of times where,

   “n” is infinite and fixed.
2. The outcome of the trial results in each trial having only two mutually exclusive possible outcomes. For example, head or tail, yes or no, success or failure, etc.
3. The occurrence of the event is called success while the non-occurrence of the event is call failure.

   p = success
   n = failure
   p+q = 1
4. The probability of success in each trial remains constant and does not change from trial to trial.
5. The trials are independent i.e. the outcomes of one toss of coin does not affect the outcome of any other tosses.

In any experiment when there are any two outcomes i.e, dual outcomes with fixed probability p and q throughout the whole process, its theoretical distribution is called binomial probability distribution.

The probability of r successes and n-r failures in n independent trials is given by

$P(X = r) = {}^nC_r p^r q^{n-r}$

Where, r = random variable

X = number of successes in end trial

p = success

q = failure

n = number of trials

A random variable X is said to follow binomial distribution if it assumes only non negative values and its probability mass function is given by

$P(X = r) = P(r) = {}^nC_r p^r q^{n-r}$ where, r= 0,1,2,3…..,n, and q = 1-p

Properties/Chief features

- It is completely determined if n and p are known.
- Since random values take only integer values, it is a discrete probability distribution.

Mean	Variance	Standard Deviation
np	npq	$\sqrt{npq}$

- Since q is the probability failure, we will always have 0<q<1 npq<np. Hence, variance is less than mean.

p=q=1/2	Symmetrical
p<1/2	Positive skewed
p>1/2	Negative skewed

- If np is a whole number then its binomial distribution is unimodal and mean = mode is being np.

c. Fitting a binomial distribution

d. Poisson distribution and its chief features (without proofs)

Poisson distribution explains the behavior of those discrete random variable for which the probability of occurrence of an event is small and the total number of possible cases is very large.

Major assumptions of Poisson distribution are as follows:-

- The occurrence of subsequent is not at all influenced by the occurrence of previous event i.e. the events occurred independently.
- The probability of occurrence of a single event within a specified time period is generally proportional to the length of the time period or time interval.
- For the very small proportion of time period under consideration, the probability of occurrence of two or more events is negligible.

Poisson distribution can be observed as an approximation or the limiting case of binomial under certain condition. Therefore, the following conditions are:-

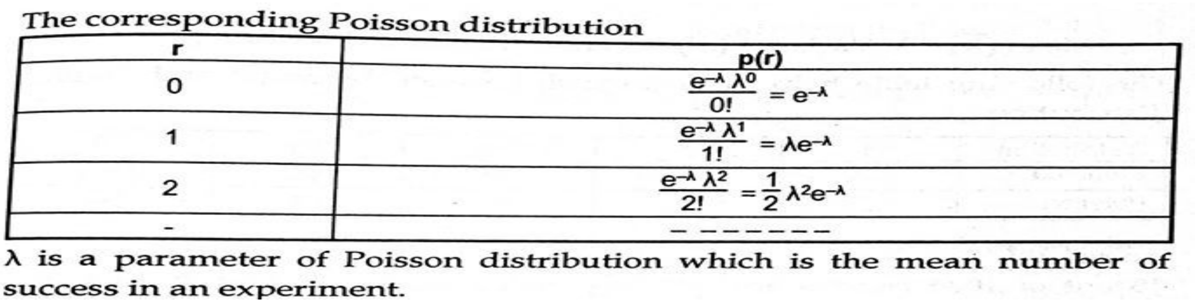
1. n, the number of trials be indefinitely very large i.e $n \rightarrow \infty$

2. Probability of occurrence p is very small i.e p → 0
   p= λ/n

   $P(r) = P(X = r) = (e^{-\lambda}\lambda^{r})/r!$, r = 0,1,2,3,....., ∞

Where,

X is the occurrence of the event
λ is the mean of the occurrence
e=2.718



The corresponding Poisson distribution

r	p(r)
0	$\frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$
1	$\frac{e^{-\lambda}\lambda^1}{1!} = \lambda e^{-\lambda}$
2	$\frac{e^{-\lambda}\lambda^2}{2!} = \frac{1}{2}\lambda^2 e^{-\lambda}$
-	– – – – – – –

λ is a parameter of Poisson distribution which is the mean number of success in an experiment.

Properties/Features

- It is a discrete probability distribution since the variable x only takes integer value.
- If we know lambda (λ), then we can probabilities of poisson distribution.
- The probability of success is very small and the probability of failure is large almost equal to one and n is insufficiently large.
- It is an independent number of trials.

Distribution	n	p	Mean	Variance	Standard Deviation
Binomial	<30	≥0.05	np	npq	$\sqrt{npq}$
Poisson	≥30	<0.05	np	np	$\sqrt{np}$

e. Fitting Poisson distribution

f. Normal Distribution and its chief features

Normal Distribution is a very important continuous probability distribution (continuous random variable).

Let a continuous variable x is said to follow normal distribution with mean μ and standard deviation σ of each probability density function (p.d.f) is given by

$f(x)=(1/\sigma\sqrt{2\pi})\,e^{-½\,[(x-\mu)/\sigma]2}$; $-\infty<X<\infty$

Standard Normal Variate

$Z=(X-\mu)/\sigma$ where,
X = random variable
μ = mean
σ = standard deviation (S.D.)

Probability density of S.N.V. is
$f(Z)=(1/\sigma\sqrt{2\pi})\,e^{-½\,Z\hat{}2}$; $-\infty<Z<\infty$

Relation between binomial and normal distribution

$n \rightarrow \infty$ and p/q not very small
$Z=(X-np)/\sqrt{np}$

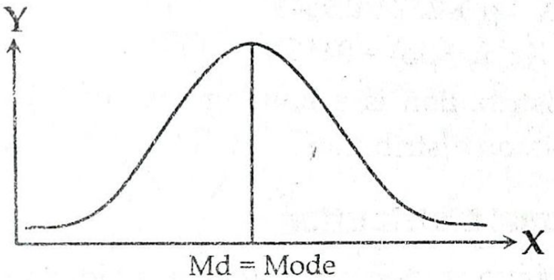
Relation between binomial and poisson distribution

$\lambda \rightarrow \infty$ and where $\lambda = np$
$Z=[X-E(X)]/\sigma = (X-\lambda)/\sqrt{\lambda}$

Properties/Features

- It is symmetrical in nature so that mean = mode = median.
- Total area under the curve is one, area under left and right is ½ which means 50%.
- The moment coefficient of skewness is 0 i.e, $\beta_1 = 0 \Rightarrow \Upsilon_1 = 0$ and the coefficient of kurtosis is $\beta_2 = 3 \Rightarrow \Upsilon_2 = 0$.
- P(X) is probability and can never be negative so no portion of the curve lies below the X-axis.

6. The mode occurs at $X = \mu$ hence the distribution is unimodal.



Md = Mode
$X = \mu(z = 0)$

7. The range of the distribution is from $-\infty$ to ∞ but practically, range $= \pm 6\sigma$

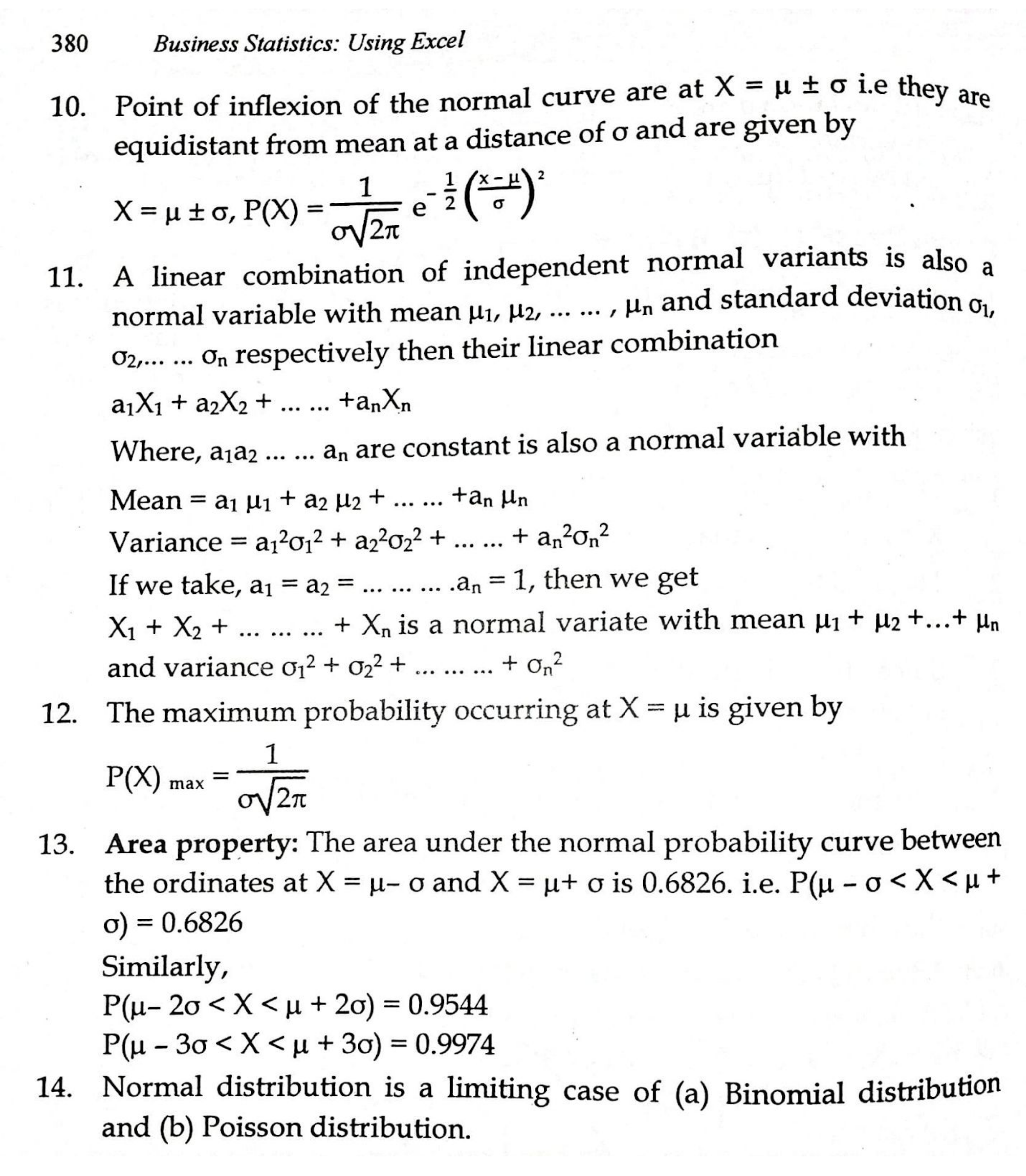
8. In normal distribution mean deviation from mean, MD $= \frac{4}{5}\sigma$

Quartile deviation, QD$=\frac{2}{3}\sigma$

Also, $QD = \frac{2}{3}\sigma = \frac{5}{6}MD$

or, $4\sigma = 5\ MD = 6QD$

9. In normal distribution, the quartiles are equidistant from median, that is $Q_3 - M_d = M_d - Q_1 \Rightarrow Q_1 + Q_3 = 2M_d = 2\mu$.



10. Point of inflexion of the normal curve are at $X = \mu \pm \sigma$ i.e they are equidistant from mean at a distance of σ and are given by

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

11. A linear combination of independent normal variants is also a normal variable with mean $\mu_1, \mu_2, \dots \dots, \mu_n$ and standard deviation $\sigma_1, \sigma_2, \dots \dots \sigma_n$ respectively then their linear combination

$a_1X_1 + a_2X_2 + \dots \dots + a_nX_n$

Where, $a_1a_2 \dots \dots a_n$ are constant is also a normal variable with

Mean $= a_1 \mu_1 + a_2 \mu_2 + \dots \dots + a_n \mu_n$

Variance $= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots \dots + a_n^2\sigma_n^2$

If we take, $a_1 = a_2 = \dots \dots \dots .a_n = 1$, then we get

$X_1 + X_2 + \dots \dots \dots + X_n$ is a normal variate with mean $\mu_1 + \mu_2 + \dots + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + \dots \dots \dots + \sigma_n^2$

12. The maximum probability occurring at $X = \mu$ is given by

$$P(X)_{\max} = \frac{1}{\sigma\sqrt{2\pi}}$$

13. **Area property:** The area under the normal probability curve between the ordinates at $X = \mu - \sigma$ and $X = \mu + \sigma$ is 0.6826. i.e. $P(\mu - \sigma < X < \mu + \sigma) = 0.6826$

Similarly,

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9974$

14. Normal distribution is a limiting case of (a) Binomial distribution and (b) Poisson distribution.

g. Areas under normal distribution

The continuous random variable Z defined as Z = (X-E(X))/SD = (X-Mean)/SD is called standard normal variate (SNV).

Symbolically, Z~N (0, 1)

The standard normal variate corresponding to X is $Z = \frac{X-\mu}{\sigma}$ which is relative measure, hence does not pose any units.

When $X = \mu + \sigma$, $Z = \frac{\mu + \sigma - \mu}{\sigma} = 1$

When $X = \mu - \sigma$, $Z = \frac{\mu - \sigma - \mu}{\sigma} = -1$

When $X = \mu + 2\sigma$, $Z = \frac{\mu + 2\sigma - \mu}{\sigma} = 2$

When $X = \mu - 2\sigma$, $Z = \frac{\mu - 2\sigma - \mu}{\sigma} = -2$

When $X = \mu + 3\sigma$, $Z = \frac{\mu + 3\sigma - \mu}{\sigma} = 3$

When $X = \mu - 3\sigma$, $Z = \frac{\mu - 3\sigma - \mu}{\sigma} = -3$

Hence the area under the standard normal probability curve

a. Between the ordinates at $z = \pm 1$ is 0.6826

b. Between the ordinates at $z = \pm 2$ is 0.9544

c. Between the ordinates at $z = \pm 3$ is 0.9974

a. When $X = \mu \pm \sigma$, the area under $\mu - \sigma$ to $\mu + \sigma$ can be computed

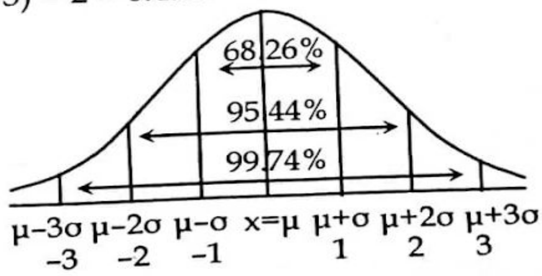
$$
\begin{aligned}
P(\mu - \sigma < X < \mu + \sigma) &= p(-1 < Z < 1) \\
&= P(-1 < Z < 0) + P(0 < Z < 1) \\
&= P(0 < Z < 1) + P(0 < Z < 1) \\
&= 2P(0 < Z < 1) \\
&= 2 \times 0.3413 = 0.6826
\end{aligned}
$$

b. Similarly, $P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2)$

$$
\begin{aligned}
&= P(-2 < Z < 0) + p(0 < Z < 2) \\
&= P(0 < Z < 2) + p(0 < Z < 2) \\
&= 2P(0 < Z < 2) = 2 \times 0.477 = 0.9544
\end{aligned}
$$

c. Similarly, $P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3)$

$$
\begin{aligned}
&= P(-3 < Z < 0) + P(0 < Z < 3) \\
&= P(0 < Z < 3) + P(0 < Z < 3) \\
&= 2P(0 < Z < 3) = 2 \times 0.4987 = 0.9974
\end{aligned}
$$



h. Hyper-geometric distribution ???

Definition

A discrete random variable X is said to follow the hyper-geometric distribution if it assumes only non-negative values and its probability mass function is given by

$$p(X=k) = h(k; N, M, n) = \frac{\binom{M}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \quad ; k = 0,1,2,...,\min(n,M)$$

$$= 0, \text{ otherwise}$$

● Estimation theory and testing of hypothesis

a. Idea of sample and population

Sample
The smallest part selected for study to generalize about the population is called the sample.

Some units selected from the population is known as sample and the process of selecting some units from the population in order to draw a conclusion about the population is known as sampling.

Population
The collection of all items and units under study is called the population or the universe.

In sample study, following to populations are to considered:

- Targeted population
  It is the population for which representative information is desired.

- The sampling population
  It is a population from which a sample will actually be taken as determined by the sample frame. The frame is merely a list of sampling units.

Census

Census is the study of each and every unit in the population. Hence, it is called the complete enumeration method.

Advantages of Census

- Complete information about the population is obtained.
- The findings can be more accurate and reliable.
- It is more useful if the area to be covered is not vast.

Disadvantages of Census

- This method is expensive in terms of time and effort.
- If the geographical area of study is large, the method might be highly expensive.
- Not in all cases can this method be adopted.

b. Point estimation and interval estimation

Point estimation

A particular value of sample statistics which is used to estimate the unknown population parameter is known as point estimation.
It is either right or wrong so it is insufficient and probability of being the right estimation is not assigned in case of point estimation. Point estimation doesn't provide a major degree of uncertainty in terms of probability attached to the estimate.

Interval estimation

A range of values which is used to estimate the population parameter is called interval estimation. The interval estimation indicates the accuracy of an estimate because the interval estimation has an advantage over the point estimation as it provides a major of degree of uncertainty in terms of probability attached to the interval.

Under it, probable range is specified within which the true value of the population parameter is expected to lie.

Should the value added or subtracted from point of estimation to convert into interval estimation?
The following factors are to be considered:

1. Standard error of estimate

The value of sample statistic and thereby, the value of point estimate may differ from sample to sample. So, the sample error of the estimate is considered to construct the interval.

2. The level of confidence
   It determines the size of interval i.e. (the difference between lower and upper limit). The greater the level of confidence requires, the larger will be the interval and vice versa. This probability indicates how confident we are that the constructed interval will include the population parameter. If we require a higher probability of falling population parameters within the interval, then more levels of confidence is assigned.

c. Characteristics of a good estimator

- Unbiasedness
  A statistic T is said to be unbiased estimator of the corresponding population parameter (θ) if the mean value of sample distribution of T statistic is equal to the population parameter i.e, E (t) = θ

- Consistency
  A statistic T is said to be a consistent estimator if it approaches the population parameter θ as sample size n increases to infinity.
  $t_n \rightarrow \theta$ as $n \rightarrow \infty$
  Alternatively, $\lim_{n\to\infty} p\{t_n \rightarrow \theta\} = 1$

- Efficiency
  An estimator is said to be an efficient estimator if its variance is smaller than others. Let t1 and t2 be the two consistent estimators than t1 is said to be a more efficient estimator than t2 if Var (t1) < Var (t2). It is because the sampling distribution of t1 is more closely clustered around θ than t2 and hence t1 is better estimator of θ than t2.

  An estimator with less variability is said to be more efficient and consequently more reliable than others.

- Sufficiency

An estimator T is said to be a sufficient estimator of population parameter θ if it contains all the information in the sample regarding the population parameter.

Properties of sufficient estimator

- It is always efficient.
- It is always consistent.
- It may or may not be unbiased.

d. Interval estimation of population parameters

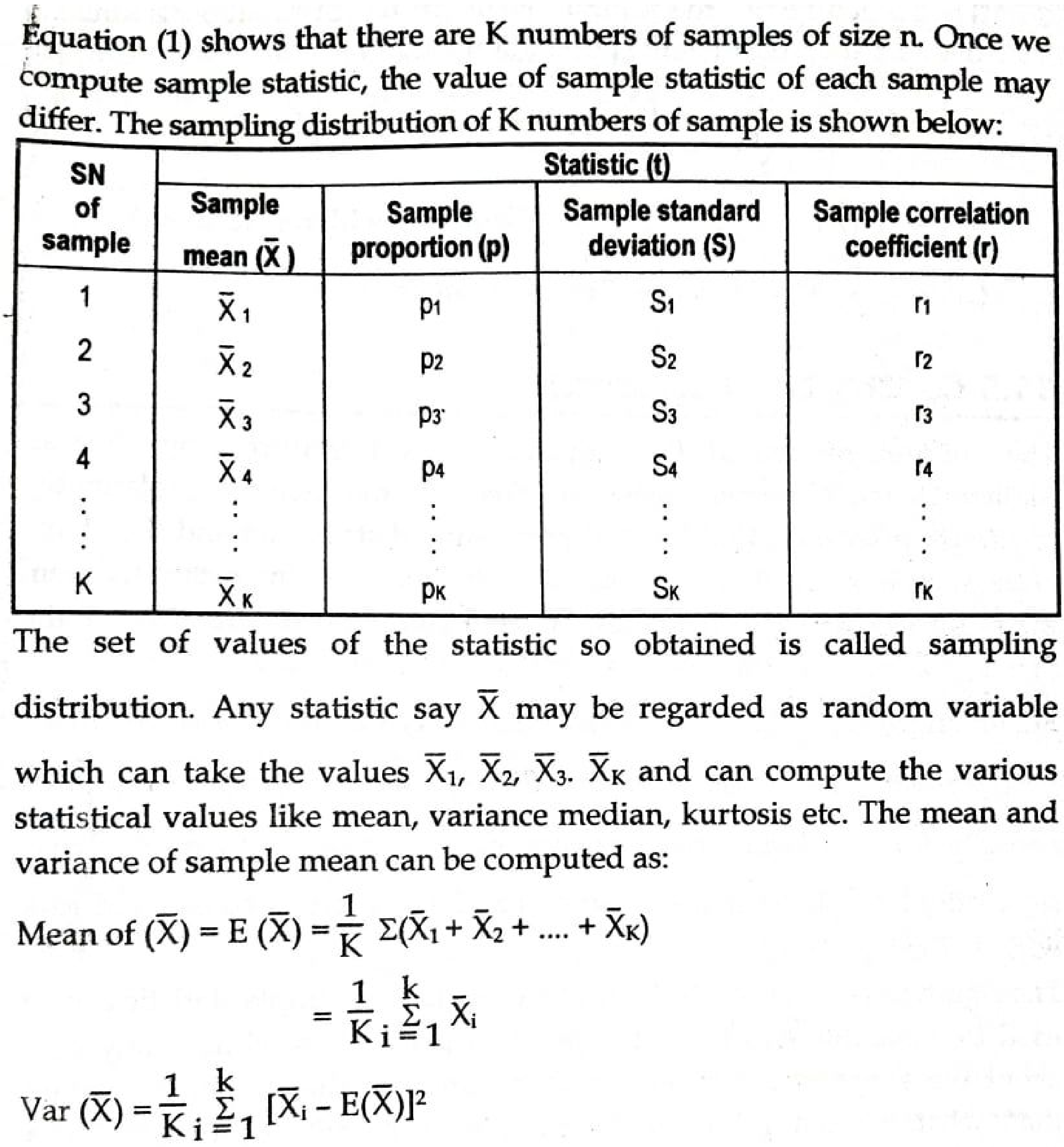
In an interval estimation of population parameter θ, if we find two quantities t1 and t2 based on sample observation drawn from the population such that the unknown parameter θ is included in the interval t1 t2 then this interval is called confidence interval for parameter θ. In other words, the interval between which unknown value of population parameter is to be expected is known as the confidence interval (Neyman) or fudicial interval (R.A. Fisher).

The lower limit t1 and upper limit t2. Let t be the value of sample statistics from which we estimate the population parameter θ then we can find upper limit *a* and *b* such that:
$P(a \leq \theta \leq b) = 1-\alpha$
Where, α = level of significance

Methods of finding confident limits Page 430 small book ???

e. Sampling distribution and standard error

Sampling distribution
Sampling distribution of any sample statistic is the probability distribution of the sample statistic. Consider the population size of N and sample of size n then the sample of size n from population size N can be drawn randomly without replacement by
${}^{N}C_{n}$ ways = (N!/((N-n)! n!)) - K ways



Equation (1) shows that there are K numbers of samples of size n. Once we compute sample statistic, the value of sample statistic of each sample may differ. The sampling distribution of K numbers of sample is shown below:

SN of sample	Statistic (t)			
	Sample mean ($\bar{X}$)	Sample proportion (p)	Sample standard deviation (S)	Sample correlation coefficient (r)
1	$\bar{X}_1$	p_1	S_1	r_1
2	$\bar{X}_2$	p_2	S_2	r_2
3	$\bar{X}_3$	p_3	S_3	r_3
4	$\bar{X}_4$	p_4	S_4	r_4
:	:	:	:	:
:	:	:	:	:
K	$\bar{X}_K$	p_K	S_K	r_K

The set of values of the statistic so obtained is called sampling distribution. Any statistic say $\bar{X}$ may be regarded as random variable which can take the values $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$. $\bar{X}_K$ and can compute the various statistical values like mean, variance median, kurtosis etc. The mean and variance of sample mean can be computed as:

$$\text{Mean of } (\bar{X}) = E(\bar{X}) = \frac{1}{K}\,\Sigma(\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_K)$$

$$= \frac{1}{K}\sum_{i=1}^{k} \bar{X}_i$$

$$\text{Var}(\bar{X}) = \frac{1}{K}\sum_{i=1}^{k} [\bar{X}_i - E(\bar{X})]^2$$

Standard Error

Standard deviation of sampling distribution of sample statistics is called standard error.

Standard Deviation	Standard Error
It is the scatterness of the variate value around the average in a single sample.	It is the scattering of sample statistics from the average of statistics.

Standard Deviation = Standard Error in sample mean, sample proportion, sample median, sample variance.

Standard Deviation ↓ Standard Error ↓ , vice versa

Standard Error ↓ Greater the uniformity of the sampling distribution and hence the greater the reliability of sample and vice versa.

Sample size ↑ Standard Error ↓

1. Standard Error of Sample Mean: Let $\overline{X_1}, \overline{X_2}, \overline{X_3} \ldots \overline{X_k}$ be the sample statistics of K numbers of samples of size n which are withdrawn from the finitely large population of size N. If the random sample is drawn with replacement from large population then the standard error of sample mean for large population is given by

SE($\overline{X}$) = $\sqrt{Var\,(\overline{X})}$

= σ **/** $\sqrt{n}$

Where, σ = Population standard deviation, n = sample size

The standard error of sample mean $(\overline{X})$ in case of sampling without replacement is less than the SE of $\overline{X}$ with replacement. So, the discrepancy between them is corrected by finite population correction.

$\sqrt{(N-n)/(N-1)}$

Hence, if the random sample is drawn without replacement from a finite population of size N then standard error of sample mean for a finite population (SRSWOR) is

SE($\overline{X}$)= $\dfrac{\sigma}{\sqrt{n}} * \sqrt{\dfrac{N-n}{N-1}}$

Where, N = Population size

n = Sample size

Importance of Sampling Distributions

i. Essential for inferential statistics.

ii. Allow us to estimate population parameters.

iii. Allow us to determine if a sample statistic differs from a known population parameter just because of chance.

iv. Allow us to compare differences between sample statistics - due to chance or to experimental treatment?

v. Sampling distribution is the most fundamental concept underlying all statistical tests

f. Sampling of attribute ??

g. Test of significance for single proportion

Z-test is applicable to test the significance of single sample proportion. If p be the sample proportion of given attribute then Z-test is used whether there is significant difference between the sample proportion and the population proportion.

The standard normal variate Z in case of single proportion is defined as -

$$Z_{cal} = \frac{p - E(p)}{SE(p)} \sim N(0,1)$$

$$= \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

where, p = sample proportion

P = population proportion

Q = 1 - P

n = sample size

Procedure

Step 1: Setting hypotheses:

a. Null hypothesis (H_0): $P = P_0$ i.e. the population proportion has some specified value P_0. In other words, there is no significant difference between the sample proportion and population proportion.

b. Alternative hypothesis (H_1): $P \neq P_0$ (two tail test) i.e. the population proportion differs from some specified value P_0. In other words, there is significant difference between the sample proportion and population proportion

or, $P > P_0$ (Right tail test) i.e. population proportion is greater than some specified value.

or, $P < P_0$ (Left tail test) i.e. the population proportion is less than P_0.

Step 2: Level of significance: 5% level of significance is used until and unless it is stated

Step 3: Choice of test statistic:
Under H_0, we use Z test for large sample.

Step 4: Computation: Z = $\dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$

where P = Population proportion of success
Q = 1 P
p = sample proportion
n = sample size

Step 5: Critical value: Mention critical value of Z at pre-assumed level of significance

Step 6: Decision:

- If I Z_{cal} I is less than or equal to I Z_{tab} I then null hypothesis (H_0) is not rejected
- If I Z_{cal} I is greater than I Z_{tab} I then null hypothesis (H_0) is rejected.

h. Test of significance for difference between two proportions

Consider two independent large population of size N_1 and N_2 possessing certain attributes with respect to the proportion then we take two independent large sample size of an n_1 and n_2 from the two populations in which x_1 and x_2 denote the observed number sample units possessing the certain attributes. Then, observed sample proportion possessing the attributes from first population (p_1) = $\frac{x_1}{n_1}$, an observed sample proportion possessing the attributes from second population proportion (p_2) = $\frac{x_2}{n_2}$

Z statistic is used to test whether there is significant difference between two sample proportions or not. The standard normal variate Z corresponding to statistic t (i.e. p_1 - p_2) is given by

Z = $\dfrac{Difference\ (t)}{SE(t)}$

Z = $\dfrac{(p_1 - p_2) - E(p_1 - p_2)}{SE(P_1 - P_2)}$

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}}$$

Under H_0, it is assumed $P_1 = P_2$. So the standard variate Z becomes

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}} \sim N(0,1)$$

Procedure

Step 1: Setting hypotheses:

a. Null Hypothesis (H_0): $P_1 = P_2$ i.e. two population proportions are same. In other words, there is no significant difference between two population proportions.

b. Alternative Hypothesis (H_1) : $P_1 \neq P_2$ i.e. two population proportions are not same. In other words, there is significant difference between two population proportions.

or, H_1: $P_1 > P_2$ (Right tail test) i.e. population proportion of first group is greater than that of the second group.

or, H_1: $P_1 < P_2$ (Left tail test) i.e. population proportion of first group is less than that of the second group.

Step 2: Level of significance: Choose the appropriate level of significance. Generally 5% level of significance is considered until and unless it is stated.

Step 3: Choice of test statistic: Under H_0, we use Z- test for large sample.

Step 4: Computation:

$$Z = \frac{(p_1 - p_2)}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}}$$

Where,

p_1 = Sample proportion of first group

p_2 = Sample proportion of second group

P_1 = Population proportion of first group

P_2 = Population proportion of second group
$Q_1 = 1 - P_1$
Q2 = $1 - P_2$
When population proportion is unknown then it is estimated as

$$\widehat{P} = \frac{n_1p_1+n_2p_2}{n_1+n_2}$$

$$\widehat{Q} = 1 - \widehat{P}$$

So that Z is

$$Z_{cal} = \frac{(p_1-p_2)}{\sqrt{\widehat{P}\widehat{Q}(\frac{1}{n_1}+\frac{1}{n_2})}}$$

Step 5: Critical value:
The critical value of Z at the pre-specified level of significance is obtained from area under normal curve.

Step 6: Decision:
• If the I Z_{cal} I is less than or equal to I Z_{tab} I then, null hypothesis (H_0) is not rejected
• If the I Z_{cal} I is greater than I Z_{tab} I then null hypothesis (H_0) is rejected.

i. Sampling of variables

j. Large samples test

Z-test is one of the most important parametric test which is based on the assumption of normality sample size is 30 or more (i.e. n>30). In other words, Z test is based on the normal probability distribution. However, sometimes the population under the study may not be normally distributed, yet this test will be made applicable as we consider sample and distribution mainly approach to normal distribution.

The Z test statistics (standardized variable) is defined as:

$$Z = \frac{Difference}{Standard\ Error}$$

$$= \frac{t - E(t)}{SE(t)}$$

For every population, the static Z is like a standard normal variate with mean 0 and variance 1. I.e Z ~ N(0,1) provided that the sample n is sufficiently large.

The z-test is used under the following assumptions:

1. The sample size is greater than 30 i.e. n > 30.
2. The samples are independent and random.
3. The population standard deviation (a) is known.
4. The population from which the sample is drawn is normally distributed.

Z-test is used for:

a. test of significance of a single mean (testing the significance of difference between sample mean and population mean): A test for significance for sampling of variables.

b. test of significance of difference between two means (testing the significance of difference between two independent sample means): A test for significance for sampling of variables.

c. test of significance of a sample proportion (testing the significance of difference between sample proportion and population proportion): A test for significance for sampling of attributes.

d. test of significance between two sample proportions (testing the significance of difference between two sample proportions): A test for significance of sampling of attributes.

k. Test of significance for single mean

Let the samples have been drawn from a normal population with mean (μ) and variance (σ^2). Then test of significance of single mean is used whether the sample mean differs significantly from the specified or hypothetical value of population mean.

Procedures to Test of Significance of a Single Mean

Step 1: Setting hypotheses:

a. Null hypothesis (H_0): $\mu = \mu_0$ i.e. the population mean has specified value μ_0. In other words, there is no significant difference between sample mean and population mean.

b. Alternative hypothesis (H_1): $\mu \neq \mu_0$ (two tail test) i.e. population mean (μ) is not equal to specific value (μ_0). In other words, there is significant difference between sample mean and population mean.

or H_1 : $\mu > \mu_0$ (Right tail test) i.e. The population mean is greater than specific value $\mu_{0.}$

or H_1 : $\mu < \mu_0$ (Left tail test) i.e. population mean is less than specific value μ_0.

Step 2: Level of significance: Generally, 5% level of significance is used until and unless it is stated.

Step 3: Test statistic: Under H_0, we use z-test

Step 4: Computation of test statistic: We have,

$$Z_{cal} = \frac{Difference}{SE(\overline{X})}$$

$$Z_{cal} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Where, $\overline{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

If the sample standard deviation (s) is given then, we can use estimated population standard deviation i.e. $\widehat{\sigma} = S$

$$Z_{cal} = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$$

Step 5: Critical value: Obtain critical value or tabulated value of Z at pre-assigned level of significance according as whether alterative hypothesis is one tail test or two tail test.

Step 6: Decision: Compare calculated value of Z (Z_{cal}) with tabulated value of Z (Z_{tab}) to draw conclusion. If | Z_{cal} | > | Z_{tab} | then H_0 is rejected otherwise H_0 is not rejected.

I. Test of significance for difference between two means

Let two independent random samples of sizes n_1 and n_2 are drawn from two different normal populations with mean μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively.

Furthermore, let $\overline{X}_1$ and $\overline{X}_2$ be the mean of samples of sizes n_1 and n_2 respectively then the sample mean $\overline{X}_1$ and $\overline{X}_2$ are normally distributed with means μ_1 and μ_2 and variances $\frac{\sigma_1^2}{n_1}$ and $\frac{\sigma_2^2}{n_2}$ respectively.

$\overline{X}_1 \sim$ N(μ_1, $\frac{\sigma_1^2}{n_1}$) AND $\overline{X}_2 \sim$ N(μ_2, $\frac{\sigma_2^2}{n_2}$)

and hence the difference ($\overline{X}_1$ - $\overline{X}_2$) is also normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

So the standard normal variate z for statistic ($\overline{X}_1$ - $\overline{X}_2$) is

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - E(\overline{X}_1 - \overline{X}_2)}{SE(\overline{X}_1 - \overline{X}_2)}$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

As null hypothesis is difference hypothesis, we set, $\mu_1 = \mu_2$ so Z becomes

$$Z_{cal} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Procedure

Step 1: Setting of hypotheses:

a. Null hypothesis (H_0): $\mu_1 = \mu_2$ i.e. two population means are equal. In other words there is no significant difference between two population means.

b. Alternative hypothesis (H_1): $\mu_1 \neq \mu_2$ (two tail test) i.e. two population means are not equal. In other words, there is significant difference between two population means.

OR H_1: $\mu_1 > \mu_2$ (Right tail test) i.e. mean of first population is greater than the mean of second population

OR H_1: $\mu_1 < \mu_2$ (Left tail test) i.e. mean of first population is less than that of the second population.

Step 2: Level of significance: Choose the appropriate level of significance. Generally, 5% level of significance is considered until and unless it is stated.

Step 3: Test statistic: Under H_0, with large sample sizes the test statistic is Z-test.

Step 4: Computation:

$$Z = \frac{Difference}{SE(\overline{X}_1 - \overline{X}_2)}$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - E(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ (Under H_0, $\mu_1 = \mu_2$)

where, $\overline{X}_1$ and $\overline{X}_2$ = sample means
σ_1^2 and σ_2^2 = population variances
n_1 and n_2 are sample size of first and second sample respectively

If population variances σ_1^2 and σ_2^2 are unknown then their estimates are provided by the corresponding sample variances σ_1^2 and σ_2^2 respectively.

i.e. $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$ large samples

$$Z_{cal} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$Z_{cal} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\sigma_1^2 = \sigma_2^2 = \hat{\sigma}^2$ common then Z statistic becomes when the variances are unknown then we can estimate by combined sample variance $\hat{\sigma}^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

m. Small sample test ???

n. Student’s T-distribution and its applications ???

● Chi-Square distribution ???

a. Introduction

b. Application

c. Test of goodness of fit

d. Test of independence of attributes

● Correlation and regression analysis

a. Introduction

b. Correlation analysis

Correlation analysis is a statistical tool which studies the relationship between two or more than two variables. Correlation analysis involves various methods and techniques for studying and measuring the extend to be correlated if the change in one variable results in a corresponding change in the other variable. For example,

- Income and expenditure of a family
- Demanding supply of commodity
- Age of husbands and wife
- Marks of students in two subjects
- Interest rate and deposit in a bank

a. Various methods of calculating correlation coefficient

b. Regression analysis