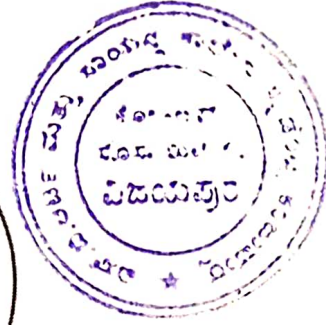


**S. B. Arts & K.C.P. Science College,
VIJAYAPUR- 586 103.**



ASSIGNMENT

For B.A./ B.Sc.^V Semester
2022 - 2023

Name of the Student Bhagyashri A T

Roll No. 236 R.C.U. Seat No. 52028524

Subject Statistics - I

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	06/02/23	03	03	J D Hosabekki Bhale	
2					
3					
4					

1. Mention various ^{components} types of time series. Explain any one

1. Secular Trend
2. Seasonal Variation
3. Cyclical Variation
4. Irregular Variation

→ Irregular Variation :- It is the irregular movements of the data over a period of time. These variations are not regular. These are sudden unexpected variations. These are caused by war, floods, strikes, lockouts, earthquakes, etc. These movements are very irregular & unpredictable ex: Sudden increase in death rate of a locality due to earth quake, tsunami etc

2. Describe moving average method. State any two merits. The method of moving averages is also based on arithmetic mean. In this method, depending on the specified period the average value for a no. of periods is taken into account and placing it at the centre of the time span, i.e. period of moving average. It is calculated for successive overlapping groups of time series data. The obtained averages are moving averages. Usually the period of moving averages will be 3 or 4 or 5 yearly / monthly.

Merits :- It is simple to understand

It is more flexible than other methods

3. Explain briefly quadratic method in time series

The quadratic equation is, $y = a + bx + cx^2$ — (1)
where, y denotes actual value, x denotes time points,
 a, b & c are constants

The constants a, b & c are obtained by solving the normal equations $\sum y = na + b\sum x + c\sum x^2$ — (2)

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \quad \text{--- (3)}$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad \text{--- (4)}$$

$$\text{if } \sum x = 0,$$

From eq ②. $\sum xy = b \sum x^2 \Rightarrow b = \frac{\sum xy}{\sum x^2}$

From eq ① $\sum y = na + c \sum x^2$ — ③

From eq ④ $\sum xy = a \sum x + c \sum x^2$ — ④

Solving ③ + ④, we obtain the values of a & c .
By substituting the values of x in the trend eqn the corresponding trend values can be obtained.

Explain briefly the straight line trend in time series.

A straight line is represented by eqn $y = a + bx$ where y - actual value, x - time. a & b are 2 constants.

The constants a & b are obtained by solving the normal eqn $\sum y = na + b \sum x$ — ①

$$\sum xy = a \sum x + b \sum x^2 \text{ — ②}$$

where n = the no. of years for which data are given.

If $\sum x = 0$, From ①

$$\sum y = na \Rightarrow a = \sum y / n$$

From eqn ②. $\sum xy = b \sum x^2$

$$\therefore b = \frac{\sum xy}{\sum x^2}$$

Thus, a is the mean of y & b is the rate of increase of y .

Explain briefly control charts.

Control. Mainly there are two types of control charts.

i) Control charts for variables

ii) Control charts for attributes

Control limits for $\bar{x}$ and R charts: Suppose X is variable quality characteristic, $\bar{x}$ -chart is used for the control of average quality level & R -chart is used for the control variation.

np or d-chart for no. of defectives: In some cases we merely want to assess whether a certain item is defective or not. Suppose k samples of size n drawn at regular intervals and d is the no. of

defectives in each of these samples. Then d no. of defectives follows a Binomial distn with mean np and standard deviation $\sqrt{npq}$, where p is proportion of defective or fraction defective.

6 Distinguish between chance & assignable causes.

Chance Causes	Assignable Causes
- It consists of many individual causes. - It results in only a small amount of time. - Chance cause of variation cannot be eliminated. - It is random & not in human hand. - Some causes of variations are ex: Slight variation in machine, lack of perfection.	- Consist of few individual causes. - It results in large amount of time. - Assignable cause of variation can be eliminated. - It is non-random & in human hand. - Assignable causes of variation are: defective raw material, quality equipment.

7 Distinguish between process and product control.

Process Control	Product Control
- Controlling the quality of the goods, during the manufacturing process itself is called process control. - This can be achieved by adopting continuous inspection methods using control charts.	- The process of inspection of manufactured lot for acceptability is called product control. - The inspection of finished or unfinished item for their acceptability is necessary. Only after inspection & confirmation quality standards the manufactured lot is accepted.

8 Explain briefly construction of $\bar{c}$ -chart

In this case instead of plotting c , the statistic $\bar{c} = c/n$ is plotted, n being the sample size which is varying. If n_i is the sample size & c_i the total no. of defects observed in the i th sample then

$$\bar{c}_i = c_i/n_i \quad (i=1, 2, \dots, k)$$

gives the average no. of defects per unit for the i th sample.

In this case an estimate of λ , the mean no. of defects per unit in the lot, based on all the k -samples is given by $\bar{\lambda} = \bar{\bar{c}} = 1/k \sum \bar{c}_i$

We know that if $\bar{x}$ is the mean of random sample of size n then $SE(\bar{x}) = \sigma/\sqrt{n}$

Hence the standard error of the average no. of defects per unit is given by

$$SE(\bar{c}) = \sqrt{\lambda/n} = \sqrt{\bar{c}/n} \quad \text{where } \bar{c} = \bar{\lambda}$$

Hence 3 σ -control limits for $\bar{c}$ -chart

$$UCL_{\bar{c}} = \bar{c} + 3\sqrt{\bar{c}/n}$$

$$LCL_{\bar{c}} = \bar{c} - 3\sqrt{\bar{c}/n}$$

$$CL_{\bar{c}} = \bar{c}$$

9 Write note on Control Chart

Control Chart is a geographical device which is used to study and verify whether the production process is in statistical control or not.

A control chart consists of three horizontal lines

1. Central line (CL)
2. Lower control limit (LCL)
3. Upper control limit (UCL)

1. The central line indicates the average quality level (CL)

2. The lower control limit is the minimum acceptable variation from the average quality for a process

3. The upper control limit is the maximum acceptable variation from the mean for a process

10 Explain briefly grouping Control Chart

Control Chart is a geographical device, which is used to verify whether a production process is in statistical control or not. Control chart is useful for the detection of existence of variation due to assignable causes.

A typical control chart consists of the following three horizontal lines: 1) A central line (CL) - Indicating the desired standard or the level of the process. 2) Upper control limit (UCL) - Indicating the upper limits. 3) Lower control limit (LCL) - Indicating the lower limits.

11 Explain briefly median test

This test is used for testing the hypothesis of identical populations. Let $x_1, x_2, \dots, x_m$ & $y_1, y_2, \dots, y_n$ be the two independent random sample for continuous population distribution $F(x)$ & $F(y)$.

To test $H_0: F_X(x) = F_Y(y)$ with corresponding one or two sided alternatives, the test statistic U is the no. of observation less than Median in combined samples. To find the null dist. of U arrange the obs. in both the samples in increasing or decreasing order & determine median.

To find $m = \left(\frac{m+n}{2} + 1 \right)$

If $m+n$ is even: $m = \frac{m+n}{2}$ total of k obs. $< m$

If $m+n$ is odd: $(m+n)/2$

Under H_0 , among $m+n$ observations every one of the $m+n$ C's set of k obs. is equally likely to comprise less than m . The no. of sets have exactly U

from x observations is ${}^m C_U {}^{n-k+U} C_{k-U}$. Thus, null distribution of U obs. among $m+n$ median is $P_{H_0}(U=u) = \frac{{}^m C_u {}^{n-k+u} C_{k-u}}{{}^{m+n} C_k}$ $u=0, 1, 2, \dots, m+n-k$

which is hypergeometric dist.

The appropriate critical regions under H_0 the alternative hypothesis are

$M_x \neq M_y$	$U \leq U_{\alpha/2}$ or $U \geq U_{\alpha/2}$
$M_x > M_y$	$U \geq U_{\alpha}$
$M_x < M_y$	$U \leq U_{\alpha}$

12 Explain briefly Mann-Whitney Wilcoxon test

This test is used to testing the hypothesis of identical pop's. Like the Run test, Mann-Whitney U test is based on the idea that the population distributed when $x_1, \dots, x_m$ random variables $y_1, \dots, y_n$ random variable are arranged together increasing order of magnitude provides information about their relationship between their population

Let $x_1, x_2, \dots, x_m$ & $y_1, y_2, \dots, y_n$ be two indep random samples from continuous distribution $F(x)$ & $F(y)$ respectively

To test $H_0: F(x) = F(y)$ with corresponding one two sided alternatives

The test statistic U then U is no. of times y precedes on x in combined order arrangement of $(m+n)$ observations.

So that U is given by $\sum_{i=1}^m \sum_{j=1}^n D_{ij}$ where,

$$D_{ij} = \begin{cases} 1 & \text{if } y_j > x_i \\ 0 & \text{if } y_j \leq x_i \end{cases} \quad i=1, 2, \dots, m \quad j=1, 2, \dots, n$$

To find Null distribution of test statistic U under H_0 , each of $m \times n$ random arrangement of combined sequence occurs with equal probability & $\gamma_{m,n}(U)$ is no. of distinguishable arrangement of m x 's values & n y 's values

Proceeds on x is exactly U so the distn func of U is $P_{H_0}(U=u) = P_{H_0}(U \leq u) = \frac{\gamma_{m,n}(u)}{m \times n}$

$$U = 0, 1, 2, \dots, m \times n$$

The appropriate critical region & alternative hypothesis

Alternatives:

$$F_y(x) < F_x(x)$$

$$F_y(x) > F_x(x)$$

$$F_y(x) \neq F_x(x)$$

Critical Region

$$U \geq U_{\alpha}$$

$$U \leq U_{\alpha}$$

$$U \leq U_{\alpha/2} \quad \text{or} \quad U \geq U_{\alpha/2}$$

10 Mark

1 Explain briefly components of time series and semi-average method

The observations of time series are affected by so many factors. The effect of these factors can be classified into four main categories, which are called as the components of time series

- 1) Secular Trend
- 2) Seasonal Variation
- 3) Cyclical Variation
- 4) Irregular Variations

Secular Trend: The general tendency of the time series data to increase or to decrease or to remain constant during a long period of time is called Secular trend. The long term or long period of time in trend samples depends upon the nature of data. Hence, long period doesn't denote a fixed period of time. In one case even few hours may be considered as long period & in another case several years may be considered as long period. It should be clearly understood that the trend is general, smooth, long term, average tendency

Ex: Increase in the price of gold in the past many years is an upward trend

Seasonal Variations: The regular, periodic & short term variation in a time series is called Seasonal variation. Generally, the period of seasonal variation would be within one year. Seasonal variation is predictable. The factors causing seasonal variations are

- 1) Climate & weather conditions

- 2] Customs, traditions & habits of people
 ex: Umbrella are sold more in rainy season
 3] Most of the students buy books & stationery in the first few months of the reopening of colleges

Cyclical Variation: Most of economic & business time series are influenced by the wave like changes of prosperity & depression. This is periodic up & down movements. This movement is known as cyclical variation or oscillatory variation. Cyclical variation is an oscillatory variation which occurs in 4 stages 1] Prosperity 2] decline 3] depression 4] Improvement

Irregular Variations

It is the irregular movements of the data over a period of time. These variations are not regular. There are sudden unexpected variations. These are caused by war, floods, strikes, earthquakes. Ex: Sudden increase in the price of vegetables due to ban on export given by an organization

Semi Average Method

In this method the whole data is divided into two parts with respect to time

For ex: If we are given U_t for t from 1971 to 1982 that is over a period of 12 years. The equal parts will be from 1971 and 1977 to 1982

In case of odd no. of years, the two parts are obtained by connecting the values corresponding to middle year

Merits: It is readily compressible as compared to method of least square

Demerits: This method assumes linear relationship between the plotted points which may not exist. It moves the limitations of arithmetic mean as average also stand in 2 ways

- 2 Explain briefly about a] Moving average method
 b] Straight line c] Linear method

Moving average Method: - The method of moving average is also based on arithmetic mean. In this method depending on the spread period the average value for a no. of periods is taken into account & plotting it at the centre of the time span i.e. period of moving average

Merits: 1] It is simple & easy to understand

2] It is more flexible than other methods

Demerits: The method does not provide the trend values at the beginning & at the end of the time series

Straight line

This straight line is represented by the eqn

$$U_t = a + bt \quad \text{--- (1) where } U_t = \text{actual time}$$

$t = \text{time}$, $a + b$ are constants

The constants a & b are obtained by solving the normal equations

$$\Sigma U_t = \Sigma (a + bt) \quad \text{--- (2)}$$

$$\frac{\partial F}{\partial a} = -2 \Sigma (U_t - a - bt) = 0$$

$$\frac{\partial F}{\partial b} = -2 t \Sigma (U_t - a - bt) = 0 \quad \left. \begin{array}{l} \frac{\partial F}{\partial a} \\ \frac{\partial F}{\partial b} \end{array} \right\} = 0$$

Take summation to eqn (1) we get

$$\Sigma U_t = \Sigma a + \Sigma bt$$

$$+ \Sigma U_t = na + b \Sigma t \quad \text{--- (3)}$$

Multiplying Σt of eqn (1) on both sides

$$\Sigma t U_t = \Sigma ta + \Sigma tbt$$

$$\Sigma t U_t = a \Sigma t + b \Sigma t^2 \quad \text{--- (4)}$$

When n is no. of year for which data are given

1] If put $\Sigma t = 0$ in eqn (3)

$$\Sigma U_t = na \Rightarrow a = \frac{\Sigma U_t}{n}$$

2] Put $\Sigma t = 0$ in eqn (4)

$$\Sigma t U_t = b \Sigma t^2 \Rightarrow b = \frac{\Sigma t U_t}{\Sigma t^2}$$

Explain construction of $\bar{x}$ & R chart

Central line and Control limits :- Control charts are plotted on a rectangular co-ordinate axis - vertical scale (ordinate) representing the statistical measures $\bar{x}$ and R, and horizontal scale (abscissa) representing the sample number. Hours, dates or lot numbers may also be represented in the horizontal scale. Sample points mean or range are indicated on the chart by points, which may or may not be joined.

For $\bar{x}$ -chart, the control line is drawn as a solid horizontal line at $\bar{\bar{x}}$ and UCL $\bar{x}$ and LCL $\bar{x}$ drawn at the computed values as dotted horizontal lines.

For R-chart, the control line is drawn as a solid horizontal line at $\bar{R}$ and UCL R is drawn at the computed values as a dotted horizontal line. If the sample size is seven or more ($n \geq 7$) LCL R is drawn as dotted horizontal line at the computed value otherwise ($n < 7$) LCL R is taken as zero.

3 Explain construction of c and U chart

C-chart : In many manufacturing or inspection situations, the sample size n i.e. the area of opportunity is very large and the probability p of the occurrence of a defect in any one spot is very small such that np is finite. In such situations from statistical theory we know that the pattern of variations in data can be represented by Poisson dist. and consequently 3 σ control limits based on Poisson distribution are used. Since for a Poisson distribution mean & variance are equal, if we assume that c is Poisson variate with parameter λ , we have

$$E(c) = \lambda \quad \text{var}(c) = \lambda$$

Thus 3- σ control limits for c-chart are given by

$$UCL_c = E(c) + 3\sqrt{var(c)} = \lambda + 3\sqrt{\lambda}$$

$$LCL_c = E(c) - 3\sqrt{var(c)} = \lambda - 3\sqrt{\lambda}$$

$$CL_c = \lambda$$

case i) Standards given : If λ' is the specified value of λ , then $UCL_c = \lambda' + 3\sqrt{\lambda'}$

$$LCL_c = \lambda' - 3\sqrt{\lambda'}$$

$$CL_c = \lambda'$$

case ii) Standards are not given : If the value of λ is not known, it is estimated by the mean no. of defects per unit. Thus if c_i is the no. of defects observed on the i th ($i=1, 2, \dots, k$) inspected unit, then an estimate of λ is given by

$$\bar{\lambda} = \bar{c} = \sum c_i / k$$

It can be easily seen that $\bar{c}$ is an unbiased estimate of λ . The control limits, in this case are given by

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}}$$

$$CL_c = \bar{c}$$

Since, c can't be negative, if LCL given by above formulae comes out to be negative, it is regarded as zero.

U-chart : In this case instead of plotting c , the statistic $u = c/n_i$ is plotted, n_i being the sample size which is varying. If n_i is the sample size and c_i the total no. of defects observed in the i th sample then, $u_i = c_i/n_i$ ($i=1, 2, \dots, k$), gives the average no. of defects per unit for the i th sample. In this case an estimate of λ , the mean no. of defects per unit in the lot, based on all the k -samples is given by

$$\bar{\lambda} = \bar{u} = \frac{1}{k} \sum_{i=1}^k u_i$$

We know that if $\bar{x}$ is mean of a random sample of size n , then $SE(\bar{x}) = \sigma/\sqrt{n}$. Hence the standard

error of the average no. of defects per unit is given by

$$St(u) = \sqrt{\lambda/n} = \sqrt{\bar{u}/n}$$

Hence, 3σ control limits for u -chart

$$UCL_u = \bar{u} + 3\sqrt{\bar{u}/n}$$

$$LCL_u = \bar{u} - 3\sqrt{\bar{u}/n}$$

$$CL_u = \bar{u}$$

4 Explain construction of np and c -chart
np Chart: While dealing with attributes, a process will be adjusted in statistical control if all the samples are ascertained to have the same popn proportion P . If d is the no. of defectives in a sample of size n then the sample proportion defectives is $p = d/n$. Hence d is a binomial variate with parameters n, P

$$E(d) = np \quad \text{var}(d) = npQ \quad Q = 1-P$$

$$E(p) = E(d/n) = \frac{1}{n} E(d) = P$$

$$\text{var}(p) = \text{var}(d/n) = \frac{1}{n^2} \text{var}(d) = PQ/n$$

Thus, the $3-\sigma$ control limits for p -chart are given by

$$E(p) \pm 3SE(p) \quad (p) = P \pm 3\sqrt{PQ/n} = P \pm A\sqrt{PQ}$$

where $A = 3/\sqrt{n}$ has been tabulated for diff values of n

case i] Standards given: If p' is the given or known value of p then

$$UCL_p = p' + A\sqrt{p'(1-p')}$$

$$LCL_p = p' - A\sqrt{p'(1-p')}$$

$$CL_p = p'$$

case ii] Standards are given: Let d_i be the no. of defectives & p_i the fraction defectives for i th sample ($i=1, 2, \dots, k$) of size n ; then the popn proportion is estimated by the statistic $\bar{p}$ is given by

$$\bar{p} = \frac{\sum d_i}{\sum n_i} = \frac{\sum n_i p_i}{\sum n_i}$$

It may be remarked here that $\bar{p}$ is an unbiased estimate of p , since

$$E(\bar{p}) = \sum E(d_i) / \sum n_i = \sum (n_i p) / \sum n_i = p$$

In this case,

$$UCL_p = \bar{p} + A\sqrt{\bar{p}(1-\bar{p})}$$

$$LCL_p = \bar{p} - A\sqrt{\bar{p}(1-\bar{p})}$$

$$CL_p = \bar{p}$$

C-Chart: In many manufacturing or inspection situations the sample size, n i.e. the area of opportunity is very large & the probability p of the occurrence of a defect in any one spot is very small such that np is finite. In such situations from statistical theory we know that the pattern of variations in data can be represented by Poisson dist, and consequently $3-\sigma$ control limits based on Poisson dist are used

$$E(c) = \lambda \quad \text{var}(c) = \lambda$$

Thus $3-\sigma$ control limits for c -chart

$$UCL_c = E(c) + 3\sqrt{\text{var}(c)} = \lambda + 3\sqrt{\lambda}$$

$$LCL_c = E(c) - 3\sqrt{\text{var}(c)} = \lambda - 3\sqrt{\lambda}$$

$$CL_c = \lambda$$

case i] Standards given: If λ' is the specified value of λ , then

$$UCL_c = \lambda' + 3\sqrt{\lambda'}$$

$$LCL_c = \lambda' - 3\sqrt{\lambda'}$$

$$CL_c = \lambda'$$

case ii] Standards are not given: If the value of λ is not known it is estimated by the mean no. of defects per unit. Thus, if c_i is the no. of defects observed on the i th ($i=1, 2, \dots, k$) inspected unit, then an estimate of λ is given by

$$\hat{\lambda} = \bar{c} = \sum c_i / k$$

It can be easily seen that $\bar{c}$ is an unbiased estimate of λ . The control limits in this case are given by

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}}$$

$$LCL_c = \bar{c} - 3\sqrt{\bar{c}}$$

$$CL_c = \bar{c}$$

Since c can't be negative, if LCL given by above formula comes out be -ve, it is regarded as zero

5 Write note on a) 3 σ limits b) Warning limits

c) Probability limits also product control

a) 3 σ limits : 3 σ limits were proposed by Dr. Shewhart for his control charts from various considerations, the main being probabilistic considerations. Consider the statistic $t = t(x_1, x_2, \dots, x_n)$ a function, the sample observations $x_1, x_2, \dots, x_n$

$$\text{let } E(t) = \mu_t$$

$$V(t) = \sigma_t^2$$

If the statistic t is normally distributed, then from the fundamental area property of the normal distribution we have

$$P[\mu_t - 3\sigma_t < t < \mu_t + 3\sigma_t] = 0.9993$$

$$P[\mu_t - \mu_t < 3\sigma_t] = 0.9993$$

$$P[t - \mu_t > 3\sigma_t] = 0.0027$$

In other words the probability that the random value t goes outside the 3 σ limits $\mu_t \pm 3\sigma_t$ is 0.0027 which is small. Hence if t is normally distributed the limits of variation should be between $\mu_t + 3\sigma_t$ and $\mu_t - 3\sigma_t$ which are termed as UCL and LCL respectively.

b) Warning limits :

The central line indicates the average quality level (C.L.) i.e. μ_t

The lower control limit is the minimum acceptable variation from the average quality level for a process

$$\text{i.e. } \mu_t - 3\sigma_t$$

The upper control limit is the maximum acceptable variation from mean for a process

$$\text{i.e. } \mu_t + 3\sigma_t$$

c) Product Control :- By the product control we mean controlling the quality of the product by critical examination of strategic points and this is achieved through sampling inspection plans. pioneered by Dodge and Roming. Product control aims at guaranteeing a certain quality level to the consumer regardless of what quality level is being maintained by the producer. In other words, it attempts to ensure that the product marketed by sale department does not contain a large no. of defective items.

- 18 Describe the plotting control chart, normal and specification limits for drawing
- The control charts we need the population parameters of the quality characteristics (i.e. specifications). Sometimes parameters are available from past production process / experience and if they are well accepted we say that they are known standards
- If machines / manufacturing processes are set afresh, the parameters are to be estimated from the data (production process). Then we say that standards are not known
1. For construction of $\bar{x}$, R charts we need mean, range (standard deviation)
 - The standards known for $\bar{x}$, R charts are denoted by $\bar{x}'$, σ' and if standards are not known are denoted by $\bar{\bar{x}}$ average of sample mean and $\bar{R}$ - mean of sample ranges
 2. For construction of np or d-chart we need fraction defective. The fraction defective when standard is known is denoted by P and standard is not known is denoted by $\bar{p}$

3. For construction of $\bar{c}$ -chart we need average no. of defects. If the standard is known it is denoted by λ' and if not known then $\bar{c}$ - average number of defects in the samples

As we know that a normal variate $X \sim N(\mu, \sigma^2)$ then from the area property of Normal distribution.

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9974$
i.e. 99.74% of the observations lie within 3-limits and only 0.26% of observations lie outside these limits

While constructing control charts, control limits are nothing but 3 σ -limits. If the manufacturing process is said to be in control, then 99.74% of observations fall within 3 σ limits

Done
06/02/23

B. L. D. E. Association's

**S. B. Arts & K.C. P. Science College,
VIJAYAPUR-586 103**



ASSIGNMENT

For B.A. / B.Sc. V Semester

2022 -2023

Name of the Student : Nilkanta V. Angadi

Roll No. : 137 R.C.U. Seat No. : 2028654

Subject : Statistics - I

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	06/02/23	03	03	J. D. Horabek: <i>[Signature]</i>	
2					
3					
4					

2 Marks

1. Define run test?

Ans: - This test is used to test the randomness of the sample and a hypothesis of identical population. Since the validity of the test is based on the number of runs in the sequence of the sample, hence it is called Run-test.

2. Define index number?

Ans: - Index numbers are statistical devices designed to measure the relative change in the level of phenomena w.r.t. time, geographical location or other characteristics such as income, profession etc.

3. Define Consumer price I. No.

Ans: Consumer price Index number is the index number of the cost met by the specified class of consumer in buying a basket of goods & services. Here, basket of goods & services means goods & services needed in day-to-day life of the specified class of consumers.

4) Write the types of I. No.

Ans: type of I. No.

1) Price Index number

2) Value Index number

3) Quantity Index number

5) Define Quality in SQC.

Ans: - The most important word in the Statistical Quality Control is quality. By quality, we mean an attribute of the product that determines its fitness for use.

6) State the type of causes of variation in SQC.

Ans: type of causes of variation in SQC

1) chance causes

2) Assignable causes

7) Mention the stage of quality.

Ans: Stage of quality
 1) Quality of materials
 2) Quality of manpower
 3) Quality of machines
 4) Quality of management.

8) State the uses of SGC

Ans: SGC helps in detection and correction of many production troubles.
 * SGC brings a substantial improvement in the product quality and reduce spoilage and rework.
 * A process is in control is predictable by SGC.

9) What do you mean by 3σ-limit.

Ans: Consider a statistic $t = t(x_1, x_2, \dots, x_n)$ a function of sample observations $x_1, x_2, \dots, x_n$. let $E(t) = \mu_1$ and $Var(t) = \sigma_1^2$. If the statistic t is normally distributed, then from the fundamental area property of the normal distribution we have
 $P(\mu_1 - 3\sigma_1 < t < \mu_1 + 3\sigma_1) = 0.9973$
 $P(|t - \mu_1| < 3\sigma_1) = 0.9973$
 $P(|t - \mu_1| > 3\sigma_1) = 0.0027$.

10) Define product control.

Ans: Product control means controlling the quality of product by critical examination at strategic points and this is achieved through "Sampling Inspection plans".

12) If $\bar{c} = 34$ then calculate Control limits for c-chart.

$$CL = \bar{c} = 34 \quad CL = 34$$

$$UCL = \bar{c} + 3\sqrt{\bar{c}} \quad UCL = 34 + 3\sqrt{34}$$

$$LCL = \bar{c} - 3\sqrt{\bar{c}} \quad LCL = 34 - 3\sqrt{34}$$

13) If $\lambda' = 3.2$ calculate Control limits for c-chart

$$\lambda' = 3.2 = CL$$

$$UCL = \lambda' + 3\sqrt{\lambda'}$$

11) Define time series.

Ans: A time series is a set of statistical observations arranged in chronological order.

15) State any two general examples for time series data.

Ans: * Weather data.
 * Temperature reading.

16) State which components of time series of below

1) The sales of stationaries, notebooks are more in the month of June & July.

Ans: This is type of Seasonal variation.

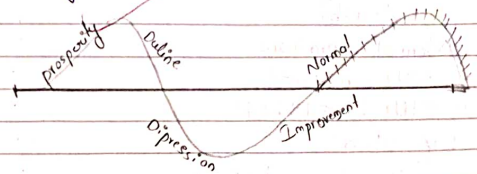
2) The no. of deaths are more due to flu & pneumonia in Feb 2022.

Ans: It is component type of irregular variation.

17) Draw the business cycle with marking in time series

Ans: Business Cycle consisting of 4 stages occurs in a period of few years. The period is not definite. Generally the period is 3-10 years.

Business Cycle.



18) State the various ^(components) measures of time series.

- 1) Secular trend
- 2) Seasonal variation
- 3) Cyclical variation
- 4) Irregular variation

19) State the various measurement of secular trend.

- Ans: 1) Graphical method
2) Semiaverages method
3) Moving averages method
4) Least squares method.

20) Write the normal eqn of $y = a + bx + cx^2$ in least square method of time series.

Ans: $U_x = a + bx + cx^2$
Normal equations
 $\sum U_x = na + b\sum t + c\sum t^2$
 $\sum tU_x = a\sum t + b\sum t^2 + c\sum t^3$
 $\sum t^2U_x = a\sum t^2 + b\sum t^3 + c\sum t^4$

21) State the results of seasonal indices ^{variation}.

22) Write down the normal equations for $y = a + bx$ in least square method.

$U = a + bt$
Normal equations
 $\sum U = na + b\sum t$
 $\sum tU = a\sum t + b\sum t^2$
if $\sum t = 0$
 $a = \frac{\sum U}{n}$ $b = \frac{\sum tU}{\sum t^2}$

5 Mark

23) Mention the various components of time series. Explain any one components.

Ans: The observations of time series are affected by so many factors. The effect of the these factors can be classified into four main categories which are called as component of time series. They are.

- 1) Secular trend
- 2) Seasonal variation
- 3) Cyclical variation
- 4) Irregular variation.

Cyclical variation: - (Business cycle)

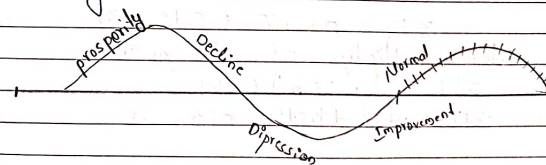
as Most of economic & business time series are influenced by the wave like changes of prosperity & depression. There is a periodic up & down movements. These movements are known as cyclical variation.

Cyclical variation is an oscillatory variation which occurs in 4 stage. Namely

- 1) Prosperity
- 2) Decline
- 3) Depression
- 4) Improvement.

Our cycle consisting of 4 stages occurs in a period of few years. The period is not definite. Generally the period is 3-10 years.

Business Cycle



24) Describe moving average method. State any two merits.

Ans: - The method of moving average is also based on arithmetic mean. In this method depending on the specified period of average value for a number of periods is taken into account & placing it at the center of the time span.

i.e. period of moving average. It is calculated for successive overlapping groups of time series the obtained average are moving average usually the period of moving average will be 3 or 5 yearly per monthly.

Merits.

- 1) it is simple & easy to understand
- 2) it is more flexible than other methods.
- 3) It is not only used for measurement of trend but also for the measurement of seasonal, cyclical and irregular variations.

25) Explain briefly quadratic method in time series.

Ans: Proceeding similarly as in the case of straight line the normal equations for estimating unknown constants a, b, c

The parabolic equation is

$$U_t = a + bt + ct^2 \quad \text{--- (1)}$$

where U_t - actual value with respect to time

a, b, c - constant

The constants a, b & c are obtained by solving the normal equation

from (1) multiplying by Σ

$$\Sigma U_t = \Sigma a + b \Sigma t + c \Sigma t^2$$

$$\Sigma U_t = na + b \Sigma t + c \Sigma t^2 \quad \text{--- (2)}$$

Now multiply Σt to eqn (1)

$$\Sigma t U_t = \Sigma t a + \Sigma t b t + \Sigma t c t^2$$

$$\Sigma t U_t = a \Sigma t + b \Sigma t^2 + c \Sigma t^3 \quad \text{--- (3)}$$

Now multiplying Σt^2 to eqn (1)

$$\Sigma t^2 U_t = \Sigma t^2 a + \Sigma t^2 b t + \Sigma t^2 c t^2$$

$$= a \Sigma t^2 + b \Sigma t^3 + c \Sigma t^4 \quad \text{--- (4)}$$

If $\Sigma t = 0$ then eqn (2) & eqn (3) & eqn (4)

$$\Sigma t U_t = a + b \Sigma t^2 + 0$$

$$b = \frac{\Sigma t U_t}{\Sigma t^2}$$

$$\Sigma U_t = na + c \Sigma t^2 \quad \text{--- (5)}$$

$$\Sigma t^2 U_t = a \Sigma t^2 + c \Sigma t^4 \quad \text{--- (6)}$$

Solving eqn (5) & (6) we obtain the values of a & c

By substituting the values of t in trend eqn the corresponding trend value can be obtained.

26) Explain briefly the straight line trend in time series.

Ans: The straight line is represented by the equation

$$U_t = a + bt \quad \text{--- (1)}$$

where U_t - actual value
 t - time

a & b are constants.

The constants a & b are obtained by solving the normal equations.

$$E = \sum (U_t - a - bt)^2 \quad \text{--- (2)}$$

$$\frac{\partial E}{\partial a} = -2 \sum (U_t - a - bt) = 0$$

$$\frac{\partial E}{\partial b} = -2 \sum t (U_t - a - bt) = 0 \quad \text{--- (3)}$$

Taking summation to eqn (2) we get

$$\Sigma U_t = \Sigma a + \Sigma b t$$

$$\Sigma U_t = na + b \Sigma t \quad \text{--- (4)}$$

multiplying Σt of eqn (1) on b.s

$$\Sigma t U_t = \Sigma t a + \Sigma t b t$$

$$\Sigma t U_t = a \Sigma t + b \Sigma t^2 \quad \text{--- (5)}$$

where n is number of years for which data are given

i) If put $\Sigma t = 0$ in eqn (4)
 $\Sigma U_1 = na \rightarrow \begin{bmatrix} a \Sigma U_1 \\ n \end{bmatrix}$

ii) put $\Sigma t = 0$ in eqn (6)
 $\Sigma U_1 = b \Sigma t^2 = b \cdot \frac{\Sigma t^2}{2}$

28) Explain briefly Control charts

Ans: defn: Control charts is a graphical device which is used to verify whether a production process is in statistical control or not. Control charts are useful for the detection of existence of variation due to assignable causes.

The typical control chart consists of the following three horizontal lines.

- 1) A Central line (CL) indicating the desired standard or the level of the process.
- 2) Upper Control line (UCL) indicating the upper limits
- 3) Lower Control line (LCL) indicating the lower limits

* Major parts of Control chart:- A control chart generally includes following four major parts:

- ① Quality Scale
- ② Sample number
- ③ plotted Samples.
- ④ The Horizontal lines.

29) Distinguish between chance & assignable causes.

Ans: Chance Causes: Stable pattern of variation is inherent in any particular scheme of the production and inspection this pattern results temperature many minor causes that behave in a random manner the variation due to these causes is beyond the control of the human hand and cannot be prevented or eliminated under any circumstances. one has to go along for variation within these

assignable Causes:- The second type of causes attributed to any production process is due to non random or assignable causes and is termed as preventable causes. The assignable causes may creep in at any stage of the process right from arrival of raw materials to the final delivery of goods.

30) Distinguish between process and product control.

Ans: As already stated, the main objective in any production process is to control and maintain a satisfactory quality level of the manufactured product so that it conforms to specified quality standards.

In other words, we want to ensure that the proportion of defective items in the manufactured product is not too large. This is termed as process control and is achieved through the technique of 'control charts'.

product control:- On the other words, by product control we mean controlling the quality of product by critical examination at strategic points and this is approx achieved through 'sampling inspection plan'.

31) Explain briefly Construction of U-chart.

Ans: ① Calculate U-value of each lot.
 i.e. $u = \frac{c}{n}$ (no defects in each lot)
 (lot size)

② Compute Central line (CL)

$$\text{Central line} = \frac{\Sigma C_i}{\Sigma n_i} \quad \left(\frac{\text{Total no of defect}}{\text{Total no of samples}} \right)$$

③ Calculate UCL & LCL of Sample Size are to unequal to control limits varies from the sample interval to sample.

④ Plot the graph with no of defect per unit on y-axis lots on the x-axis

⑤ draw Central line, UCL and LCL

⑥ finally interpret the data to determine whether the process is in control.

Alternatives

$$f_y(x) < f_x(x)$$

$$f_y(x) > f_x(x)$$

$$f_y(x) \neq f_x(x)$$

Critical region

$$U > U_{\alpha}$$

$$U \leq U_{\alpha}$$

$$U < U_{\alpha/2} \quad U > U_{\alpha/2}$$

10 Marks

1. Explain briefly components of time series and semi-average method

The observations of time series are affected by so many factors. The effect of these factors can be classified into four main categories which are called as the components of time series.

- (i) Secular trend
- (ii) Seasonal variation
- (iii) Cyclical variation
- (iv) Irregular variation

* **Secular Trend** :- The general tendency of the time series data to increase or to decrease or to remain constant and during a long period of time. is called secular trend. The long term or long period of time is trend. Samples depends upon the nature of data. Hence long period doesn't denotes a fixed period of time. In one case even few hours may be considered as long period & in another case several years may be considered as the trend is general. Smooth, long term, upward or downward tendency.

Ex: Increase the price of oil in the past many years is an upward trend.

* **Seasonal Variations** :- The regular periodic & short term variation. In a time series is called seasonal variation. Generally the period of seasonal variation is predictable. The factors causing seasonal variations are

- 1) Climate & weather conditions
- 2) Customs, traditions & habits of people
- 3) Most of the students buy books & stationery in first few months of the reopening of colleges.

Cyclical variation :- most of economic & business here series are influenced by the wave like changes of prosperity & depression. This is periodic up & down movements. This movement is known as Cyclical Variation or Oscillatory Variation. Cyclical variation is an oscillatory variation which occurs in 4 stages 1) Prosperity 2) decline 3) depression 4) Improvement.

Irregular Variation :- It is the irregular movements of the data over a period of time. These variations are not regular. These are sudden unexpected variations.

Semi Average Method

In this method the whole data is divided into two parts with respect to time.
for ex: If we are given U for 4 from 1971 to 1982. That is over a period of 12 years. The equal parts will be from 1971 and 1977 to 1982. In case of odd no of years the two parts are obtained by ignoring the value corresponding to middle year.

Merits :- It is readily compressible as compared to method of least square.

Demerits :- This method assumes linear relationship between the plotted points which may not exist. Moreover the limitations of arithmetic mean as an average also stand in its way.

2. Explain briefly about a) Moving average method b) Straight line c) Linear method

Moving average method :- The method of moving average is also based on arithmetic mean. In this method depending on the specified period the average scales for a no of periods is taken into account & plotting it to the centre of the time span. i.e period of moving average. Merits 1) It is simple & easy to understand 2) It is more flexible than other methods.

Demerits :- The method does not provide the trend values at the end of the time series.

Straight line :-

This straight line is represented by the eqn
 $U_t = a + bt$ — (1) where U_t = actual time. t = time a & b are constants. The constants a & b are obtained by

Solving the normal equations.

$$F = \sum (U_t - a - bt)^2 \text{ — (2)}$$

$$\frac{\partial F}{\partial a} = -2 \sum (U_t - a - bt) = 0. \quad \left. \begin{array}{l} \\ \end{array} \right\} \text{ (3)}$$

$$\frac{\partial F}{\partial b} = -2 \sum t(U_t - a - bt) = 0$$

Take summation to eqn (1) we get

$$\sum U_t = \sum a + \sum bt$$

$$\sum U_t = na + b \sum t \text{ — (4)}$$

multiplying $\sum t$ of eqn (1) on b.s

$$\sum t U_t = \sum ta + \sum bt^2$$

$$\sum t U_t = a \sum t + b \sum t^2 \text{ — (5)}$$

When n is no. of years for which data are given

1) If put $\sum t = 0$ in eqn (4)

$$\sum U_t = na \Rightarrow a = \frac{\sum U_t}{n}$$

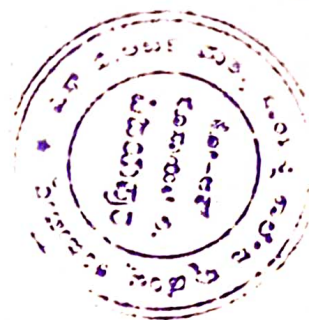
2) put $\sum t = 0$ in eqn (5)

$$\sum t U_t = b \sum t^2 \Rightarrow b = \frac{\sum t U_t}{\sum t^2}$$

Done
 06/02/23

B.L.D.E. Association's

S. B. Arts & K.C.P. Science College, VIJAYAPUR - 586 103.



ASSIGNMENT

For B.A. / B.Sc. ^{Vth} Semester
2022 - 2023

Name of the Student Aishwarya Rajput

Roll No. 100 R.C.U. Seat No. 52028411

Subject STATISTICS - II

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	08/02/23	03	03	Miss S. Y. Nagole	
2				<u>Shruti</u>	
3					
4					

STATISTICS - II — 2021

Sampling Technique & Demography

SECTION - A

I

① Differentiate blw population & Sample:

Population	Sample
① The group of individuals under study is called population or universality. Ex = Group of people.	① Sample is a subset of the population. Ex = the taking a sample of food before purchasing

② What do you mean by non-Sampling errors. Non Sampling error is the error that arises in the data collection process as a result of factors other than taking a sample.

③ Name the some sets of random number table commonly used in practice

- ① Tippett's random Number table
- ② Fisher & Yates Random Number table
- ③ Ranks Corporation Random Number table.
- ④ Kendall & Babington Smiths R. no. table.

④ What do you mean by Sampling Fraction of finite population correction.

Sampling fraction:

= The Ratio of sample size (n) to the population size (N) is called Sampling fraction (f).

$(1-f) = \left(1 - \frac{n}{N}\right)$ is called finite pop'n correction.

⑤ Define Stratified Random Sampling:-

=> The Sample termed as stratified random sample of size 'n' and the technique of drawing such a sample is called Stratified Random Sampling.

⑥ Write the expression for gain in precision due to Stratified Random Sampling over SRS.

$$= E - 1 = \frac{\text{var}(\bar{y}_n)_{\text{SRS}}}{\text{var}(\bar{y}_{st})} - 1 = \frac{\text{var}(\bar{y}_n)_{\text{SRS}} - \text{var}(\bar{y}_{st})}{\text{var}(\bar{y}_{st})}$$

- (7) Define Systematic Sampling.
Systematic Sampling Consists in selecting the samples only the 1st unit at random, the rest being automatically selected according to some pre-determined pattern involving regular spacing of units.

- (8) Write any two demerits of systematic sampling.

Demerits

- (i) The main disadvantage of systematic sampling is that systematic samples are not in general random samples since the requirement in merit.
- (ii) If N is not multiple of n , then
 - (i) The actual sample size is different from that required.
 - (ii) Sample mean is not an unbiased estimate of the population mean.
- (iii) Not able to obtain an unbiased estimate of this variance.

- (9) Name types of Mortality Rates:-

- (i) Crude Death Rate
- (ii) ~~Sp~~ ~~Sp~~ Specific Death Rate
- (iii) Age Specific Death Rate
- (iv) Standardization Death Rate
- (v) Infant Mortality Rate (I.M.R)
- (vi) Neonatal Mortality Rate
- (vii) Maternal Mortality Rate.

- (10) Define GRR:-

⇒ Gross Reproduction Rate:- The gross reproduction rate (G.R.R) is a step in this direction & it defined as the sum of age specific fertility rates calculated from female births for each year of reproductive period.

SECTION - B

II (B)

- (13) Explain the principles of sample survey

- (1) Objectives of the survey.
- (2) Defining the population to the sample
- (3) Data to be collected
- (4) The Questionnaire / schedule.
- (5) The Frame and Sampling units
- (6) Method of Collecting Information.
- (7) Non Respondents
- (8) Selection of Proper Sampling design.
- (9) Organization of field work
- (10) The Pre-test
- (11) Summary & analysis of data.
- (12) Information gained for future survey.

- (14) Name the different methods of selecting a SRS. Explain any one of them.

→ There are two types of method of selecting SRS:-

- (i) Lottery Method / Lottery System
- (ii) Mechanical Randomization / Random number method.

(a) # Lottery Method:-

⇒ The Simplest method of selecting a random sample is the lottery system which is as follows.

→ Suppose we want to select our candidates out of 'n' we assign the numbers 1, 2, ..., n one number to each candidate & write these numbers on 'n' slips which are made as homogeneous as possible in shape, size, colour, etc. These slips are then put in a bag and thoroughly shuffled & then our slips are drawn one by one.

→ The Candidates corresponding to numbers on the slips drawn, will constitute a random sample.

→ This method of selection is quite independent of Properties of population.

→ Generally in place of cheats / slips we can use cards.

(15) With usual notations prove that

$$\text{Var}(\bar{y})_{\text{prop}} \leq \text{Var}(\bar{y})_{\text{SR}} \\ \text{W.K.T, } \text{Var}(\bar{y})_{\text{WOR}} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \quad \text{--- (1)}$$

$$\text{Var}(\bar{y})_{\text{prop}} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^k p_i s_i^2 \quad \text{--- (2)}$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_N)^2$$

In order to compare eq^s (1) & (2) we shall first express S^2 in term of

$$S^2 = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i} + \bar{y}_{N_i} - \bar{y}_N)^2$$

$$(N-1)S^2 = \sum_{i=1}^k \left[\sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2 + (\bar{y}_{N_i} - \bar{y}_N)^2 + 2 \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})(\bar{y}_{N_i} - \bar{y}_N) \right]$$

$$= \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2 + \sum_{i=1}^k (\bar{y}_{N_i} - \bar{y}_N)^2 + 2 \sum_{i=1}^k \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})(\bar{y}_{N_i} - \bar{y}_N)$$

$$S_i^2 = \frac{1}{N_i-1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2$$

$$(N-1)S^2 = (N-1)S_i^2 + \sum_{i=1}^k (\bar{y}_{N_i} - \bar{y}_N)^2 + 0$$

$$= (N-1)S_i^2 + \sum_{i=1}^k (\bar{y}_{N_i} - \bar{y}_N)^2$$

If we assume that N_i consequently N are sufficiently large so that we can take $N_i - 1 \approx N_i$ & $N - 1 \approx N$ then we get

$$NS^2 \approx N S_i^2 + \sum_{i=1}^k N_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

$$S^2 \approx p_i s_i^2 + \sum p_i (\bar{y}_{N_i} - \bar{y}_N)^2 \quad \text{--- (3)}$$

Substituting eq⁽³⁾ in eq⁽¹⁾ we get

$$NS^2 \approx \sum N_i s_i^2 + \sum N_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

$$S^2 \approx \sum p_i s_i^2 + \sum p_i (\bar{y}_{N_i} - \bar{y}_N)^2 \quad \text{--- (3)}$$

$$\text{Var}(\bar{y})_{\text{prop}} = \text{Var}(\bar{y})_{\text{prop}} + \left(\frac{1}{n} - \frac{1}{N}\right) \sum p_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

$$\Rightarrow \text{Var}(\bar{y})_{\text{prop}} \geq \text{Var}(\bar{y})_{\text{prop}} \quad \text{--- (4)}$$

eq⁽⁴⁾ leads to the following important conclusion.

"Stratified Random Sampling with proportional allocation gives greater efficiency than Simple random Sampling."

(16) With usual notations P.T

$$\text{Var}(\bar{y})_{\text{SRSHOR}} = (N-1)S^2 + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

Proof = In case of SRSHOR we have

$$\text{Var}(\bar{y})_{\text{SRSHOR}} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \quad \text{or} \quad \left(\frac{N-n}{Nn}\right) S^2$$

$$\text{Var}(\bar{y}_{\text{sys}}) = \frac{N-1}{N} S^2 + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

where $N = nk$

Consider $\Rightarrow \text{Var}(\bar{y})_{\text{SRSHOR}} - \text{Var}(\bar{y}_{\text{sys}})$

$$\Rightarrow \left(\frac{N-n}{Nn}\right) S^2 - \frac{N-1}{N} S^2 + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

$$\Rightarrow \left[\frac{N-n}{n} - (N-1)\right] \frac{S^2}{N} + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

$$\Rightarrow \left[\frac{N-n-(N-1)n}{n}\right] \frac{S^2}{N} + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

$$\Rightarrow \left[\frac{N-n-nN+n}{n}\right] \frac{S^2}{N} + \frac{k(n-1)}{N} S^2_{\text{wsy}}$$

$$\Rightarrow \left[\frac{N(1-n)}{N} \right] \frac{s^2}{N} + k \frac{(n-1)}{N} s_{wsy}^2$$

$$\Rightarrow \frac{k(n-1)}{N} s_{wsy}^2 - \frac{(n-1)}{N} s^2$$

$$\frac{k(n-1)}{NnK} s_{wsy}^2 - \frac{(n-1)}{N} s^2$$

$$\left(\frac{n-1}{N} \right) (s_{wsy}^2 - s^2)$$

$$\frac{n-1}{N} (s_{wsy}^2 - s^2)$$

The systematic sampling gives more precise estimate of the popn mean as compared with SRSWOR iff $\text{Var}(\bar{Y}_n)_{\text{SRSWOR}} - \text{Var}(\bar{Y}_{\text{sys}}) > 0$

$$\Rightarrow s_{wsy}^2 > s^2$$

This leads to the following important conclusion.

Conclusion:-

"A systematic sample is more precise than an SRSWOR. If the mean square within the systematic sample is larger than the population mean square."

Hence Proved.

(B) 41 Explain NRR $\Rightarrow$

The Net Reproduction Rate (NRR) takes into account current fertility but not current mortality. To take into account the factor of mortality in measuring population growth we use Net reproduction rates.

Net reproduction rate is defined as the average number of daughters that would be born to a female if she passed through her life time conforming to the

-age-specific fertility & mortality rates of a given year

Net reproduction rate is the number of female children surviving till their reproductive ages or number of female children to one woman as she passes through child bearing age.

NRR cannot exceed GRR for the reason that in its calculation we also take into account mortality. Both the rates will be equal when all the newly born daughters reached the child bearing age and passed through it. If the NRR per woman is exactly one, it indicates that on the basis of current fertility and mortality rates, a group of newly born females will exactly replace itself in the new generation i.e. the population will be constant. If the NRR per woman is below one, it indicates a declining population and if it is more than one, the population has a tendency to increase.

The NRR is computed by the following formula

$$NRR = \sum_{i=1}^n w_i F_i \times S$$

i.e. Magnitude of age class

$$WSFR = \frac{\text{Female birth}}{\text{Female popn}} \times 1000$$

$$= \frac{P_{B1}}{P_{P1}} \times 1000$$

S = Survival Rates.

Merits:-

(i) It takes into account fertility & current mortality.

- (2) It will show a tendency of increase or decrease in popⁿ.

Demerits:-

- (1) It assumes constant rate of fertility & mortality over a generation. In actual life, both these rates go on changing.
- (2) The popⁿ of a country may become depleted more by migration than by declining birth rate.
- (3) It cannot be used for forecasting future Population Changes.

SECTION-C

- (19) Discuss the advantages and disadvantages of Sample survey over Censuses.

Advantages of Sampling

Sampling ensures convenience, collection of intensive and exhaustive data, suitability in limited resources and better rapport. In addition to this, sampling has the following advantages also

- (1) Low cost of Sampling
⇒ If data were to be collected for the entire population, the cost will be quite high.

- (2) Less time consuming in Sampling

⇒ Use of Sampling takes less time also.

It consumes less time than census technique. Tabulation, analysis etc., take much less time in the case of a sample than in the case of a population.

- (3) Scope of Sampling is high:-
⇒ The investigator is concerned with the generalization of data.

- (4) Accuracy of data is high:-
⇒ A sample represents the population from which it is drawn.

- (5) Organization of Convenience
⇒ Organization of problems involved in Sampling are very few.
* Study of samples involves less space & equipment.

- (6) Intensive and exhaustive data
* In Sample studies, measurements or observations are made of a limited number.
* So intensive & exhaustive data are collected

- (7) Suitability in limited resources:-
* The resources are limited.
* Studying the entire universe is not viable.
* The population can be satisfactorily covered through Sampling.

- (8) Better rapport
* A effective research study requires a good rapport between the researcher & the respondents.
* When the popⁿ of study is large, the problem of rapport arises.

Disadvantages:-

- * Chance of bias
- * Difficulties in selecting truly a representative sample
- * Need for subject specified knowledge
- * Impossible of Sampling.

(B) Chance of bias:-

The serious limitation of the sampling method is that is biased selection & thereby leads us to draw erroneous conclusion.

Bias arises when the method of selection of sample employed is faulty. Relative small samples employed & properly selected may be much more reliable than large samples poorly selected.

(3) Difficulties in selecting a truly representative sample produces reliable & accurate results only when they are representative of the whole group. Selection of a truly representative sample is difficult when the phenomena under study are of a complex nature. Selecting good sample is difficult.

(3) Inadequate knowledge in the subject. Use of a sampling method requires adequate subject specific knowledge in sampling technique. Sampling involves statistical analysis and calculation of probable error, when the researcher lacks specialized knowledge in sampling, he may commit serious mistakes. Consequently, the results of the study will be misleading.

(4) Changeability of units:-

When the units of the population are not homogeneous, the sampling technique will be unscientific. In sampling, though the no. of cases is small, it is not always easy to stick to the selected cases. The units of sample may be widely dispersed.

Some of the samples of cases may not cooperate with the researcher & some others may be inaccessible.

Because of these problems, all the cases may not have to be replaced by other cases. Changeability of units stands in the way of results of the study.

(5) Impossible of Sampling.

= Deriving a representative sample is difficult, when the universe is too small or too heterogeneous. In this case, Census study is the only alternative. Moreover, in studies requiring a very high standard of accuracy, the sampling method may be unsuitable. There will be chances of errors even if samples are drawn most carefully.

(2) With usual notations Prove that
$$\text{Var}(\bar{Y}_n)_{\text{SRV}} = \frac{N-n}{N} S^2$$

Let the population observations are $y_1, y_2, \dots, y_n$

Population mean $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^N Y_i$

Sample mean $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$

Population mean square $S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y}_N)^2$

Sample mean square $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$

also and also $(n-1)S^2 = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$

Now variance $(\bar{Y}_n) = E(\bar{Y}_n)^2 - (E(\bar{Y}_n))^2$

$[E(\bar{Y}_n)]^2 = \bar{Y}_N^2 - \text{---}$

Consider, $E(\bar{Y}_n)^2 = E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right]^2$

$$= E\left(\frac{1}{n^2} \left(\sum_i y_i^2 + \sum_{i \neq j} y_i y_j \right)\right)$$

$$= \frac{1}{n^2} [E(\sum y_i^2) + E(\sum y_i y_j)] \quad \text{--- (4)}$$

$$E(\sum y_i^2) = n \sum_{i=1}^N y_i^2 \quad \text{--- (2)}$$

$$S^2 = \frac{1}{N-1} \sum (y_i - \bar{y}_N)^2$$

$$(N-1)S^2 = \sum y_i^2 - N\bar{y}_N^2$$

$$\sum y_i^2 = (N-1)S^2 + N\bar{y}_N^2$$

Substituting this eq (2) in eq (4) we get

$$E(\sum y_i y_j) = \frac{n}{N} [(N-1)S^2 + N\bar{y}_N^2]$$

$$= n \left[\frac{N-1}{N} S^2 + \bar{y}_N^2 \right] \quad \text{--- (3)}$$

Now

$$E(\sum y_i y_j) = \frac{n(n-1)}{N(N-1)} \sum y_i y_j$$

$$= \frac{n(n-1)}{N(N-1)} [\sum y_i^2 - \sum y_i^2]$$

$$= \frac{n(n-1)}{N(N-1)} [\sum y_i^2 - (N-1)S^2 - N\bar{y}_N^2]$$

$$= \frac{n(n-1)}{N(N-1)} [N\bar{y}_N^2 - (N-1)S^2]$$

$$= n(n-1) \left[\bar{y}_N^2 - \frac{S^2}{N} \right] \quad \text{--- (4)}$$

Substituting (3) & (4) in eq (2) we get

$$E(\bar{y}_N)^2 = \frac{1}{n^2} \left[n \left(\frac{N-1}{N} S^2 + \bar{y}_N^2 \right) + n(n-1) \left(\bar{y}_N^2 - \frac{S^2}{N} \right) \right]$$

$$= \frac{1}{n} \left[\frac{N-1}{N} S^2 + \bar{y}_N^2 + (n-1) \bar{y}_N^2 - (n-1) \frac{S^2}{N} \right]$$

$$= \frac{1}{n} \left[\frac{S^2}{N} (N-1-n+1) + \bar{y}_N^2 (1+n-1) \right]$$

$$= \frac{1}{n} \left[\frac{S^2}{N} (N-n) + n\bar{y}_N^2 \right]$$

$$E(\bar{y}_N)^2 = \frac{N-n}{n} \frac{S^2}{N} + \bar{y}_N^2$$

Substituting the above eq (1) in eq (1)

$$V(\bar{y}_N) = \frac{N-n}{nN} S^2 + \bar{y}_N^2 - \bar{y}_N^2$$

$$= \frac{N-n}{nN} S^2$$

Hence proved.

(2) In Stratified Sampling, P.T. $\text{Var}(\bar{y}_N)$ is minimum for fixed total sample size n , if $n_i \propto N_i - S_i$ & obtain $\text{Var}(\bar{y}_N)$ by

Hence the problem is to minimise

$$\text{Var}(\bar{y}_{st}) = \frac{1}{n^2} \sum_{i=1}^K n_i (N_i - n_i) \frac{S_i^2}{n_i}$$

Subject to the condition

$$\sum n_i = n \text{ (fixed)}$$

this is equivalent to minimizing ϕ

$$\therefore \phi = \text{Var}(\bar{y}_{st}) + \lambda \left(\sum_{i=1}^K n_i - n \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^K n_i (N_i - n_i) \frac{S_i^2}{n_i} + \lambda \left(\sum_{i=1}^K n_i - n \right)$$

unconditionally for variation in n_i , λ being unknown lagrange multiplier.

for an extremum, we should have

$$\frac{\partial \phi}{\partial n_i} = 0$$

$$\frac{1}{n^2} \left[N_i S_i^2 - \frac{1}{n} \sum_{i=1}^K n_i \cdot n_i \cdot \frac{S_i^2}{n_i} \right] + \lambda \left(\sum_{i=1}^K n_i - n \right)$$

$$- \frac{N_i^2 S_i^2}{n^2 n_i} + \lambda = 0$$

$$\lambda = \frac{N_i^2 S_i^2}{n^2 n_i}$$

$$n_i = \frac{N_i S_i}{n \lambda}$$

$$\text{Also } \frac{\partial^2 \phi}{\partial n_i^2} > 0$$

which implies that n_i is given by eq (1) provided minimum of ϕ .
To determine λ , we have to sum eq (1) over i 's from 1 to k & use the condition $\sum n_i = n$ to get

$$\sum \lambda = \frac{\sum n_i s_i}{n}$$

$$= \sum_{i=1}^k \left(\frac{n_i s_i}{n} \right)$$

$$= \sum_{i=1}^k \frac{n_i s_i}{n}$$

Substituting the value of λ in eq (2) we get $n_i = \frac{n_i s_i}{\sum \frac{n_i s_i}{n}}$

$$= \frac{n_i s_i}{\sum \frac{n_i s_i}{n}}$$

This is known as Neyman's formula for optimum allocation.

(23) With usual notations P.T.

$$\text{var}(\bar{y}_s) = \frac{(nk-1)}{nk} \frac{s^2}{n} [1+(n-1)f]$$

we have

$$\text{var}(\bar{y}_{sys}) = \frac{1}{k} \sum_{i=1}^k (\bar{y}_i - \bar{y}_{..})^2$$

$$= \frac{1}{k} \sum_{i=1}^k \left[\frac{1}{n} \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \right]$$

$$nk \text{var}(\bar{y}_{sys}) = \sum_{i=1}^k \left[\sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \right] \quad \text{--- (1)}$$

$$= \sum_{i=1}^k \left[\sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 + \sum_{j \neq i}^n (y_{ij} - \bar{y}_{..})(y_{ij} - \bar{y}_{..}) \right]$$

$$= (nk-1)s^2 + (n-1)(nk-1)s^2 f$$

$$= (nk-1)s^2 [1+(n-1)f]$$

$$nk \text{var}(\bar{y}_{sys}) = (nk-1)s^2 [1+(n-1)f]$$

$$\text{var}(\bar{y}_{sys}) = \frac{(nk-1)s^2}{nk} [1+(n-1)f]$$

$$= \frac{(nk-1)}{nk} \frac{s^2}{n} [1+(n-1)f]$$

Hence proved.

(24) Mention the source of vital statistics. Explain any one of them.

(a) Explain standardised death rate.

$\Rightarrow$ (a) Sources of Vital Statistics

(i) Two methods

(1) Registration method (Vital statistics)

(2) Census method (Census Report)

(i) Registration method

$\Rightarrow$ In almost all countries the occurrence of every vital event is registered to proper authorities under with laudable arguments.

Birth and death are registered in town municipality office, Panchayat offices, Marriages & Migrations are registered in the respective registration offices.

Birth Death registers provides information about date & place of death, age, sex, occupation, marital status, religion causes of death.

The information regarding marriages and migrations can be obtained from the respective registers.

(b) Standardised Death Rate:-

For comparing death rates of different populations CDR and ASDR cannot be effectively used. Because age composition in these populations are likely to be different.

- Therefore when a comparison of death rates is needed, the standardised death rate are made use of.

Here one particular population is taken as standard and then based on the age structure of this standard population the death rates for the given population are compared. It is assumed that all the given populations have common age structure population.

Consider ASRR of A & B are as follows
 $m^a = \frac{D^a}{p^a}$ $m^b = \frac{D^b}{p^b}$ (for CDR)

$m_x^a = \frac{D_x^a}{P_x^a}$ $m_x^b = \frac{D_x^b}{P_x^b}$ (for ASOR)

The CDR's of A & B in terms of ASDR's are as follows

$$m^a = \frac{D^a}{p^a} = \frac{\sum M_x^a P_x^a}{\sum P_x^a} \quad \text{--- (1)}$$

$$m^b = \frac{\sum M_x^b P_x^b}{\sum P_x^b} \quad \text{--- (2)}$$

§ for STDR

$$\text{i.e. (STDR) A} = \frac{\sum m_x^a P_x^S}{\sum P_x^S}$$

$$\text{(STDR) B} = \frac{\sum m_x^b P_x^S}{\sum P_x^S}$$

∴ P_x^S = standard population in different age groups.

~~08/02/23~~
 08/02/23

B.L.D.E. Association's

**S. B. Arts & K.C.P. Science College,
VIJAYAPUR- 586 103.**



ASSIGNMENT

For ~~B.A.~~/ B.Sc.^I..... Semester
2022 - 2023

Name of the Student Alfiya. Choudhari.
Roll No. 81 R.C.U. Seat No. U15KM22S0124
Subject Statistics

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	11/01/23	05	05	Mrs S. Y. Nagole	
2				<i>[Signature]</i>	
3					
4					

PART-A

I. 2 marks questions.

1. Define quantitative data with an example.

The data which is numerically expressed is called Quantitative Data.

eg:- Height and Weight of students.

2. Define statistical table.

The systematic arrangement of the classified data in rows and columns in the form of table. This table is known as Statistical table.

3. Define Deciles and Quartiles.

Deciles: Measures which divides the data into ten equal parts.

Quartiles: Measures which divides the data into four equal parts.

for grouped data

$$D_i = L + \left(\frac{iN - CF}{\frac{N}{10}} \right) \frac{h}{f}$$

for ungrouped data

$$D_i = \left(\frac{i(n+1)}{10} \right)^{th} \text{ term}$$

$$Q_i = L + \left(\frac{iN - CF}{\frac{N}{4}} \right) \frac{h}{f}$$

$$Q_i = \left(\frac{i(n+1)}{4} \right)^{th} \text{ term}$$

4. What is kurtosis?

Kurtosis is the measurement of flatness or peakedness of the distribution. It is measured by the co-efficient β_2 or its derivation γ_2 are given by

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

$$\gamma_2 = \beta_2 - 3$$

where, μ_4 is 4th moment about mean

μ_2 is 2nd moment about mean.

5. Define Standard Deviation.

The standard deviation is defined as positive root of arithmetic mean of squared deviations, when deviations are taken from mean.

Standard deviation is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \text{ for ungrouped data.}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N}} \text{ for grouped data.}$$

6. Given $b_{xy} = 0.64$ & $b_{yx} = 1.21$, then find correlation co-efficient.

Correlation co-efficient is given by

$$\begin{aligned} r_{xy} &= \sqrt{b_{xy} \cdot b_{yx}} \\ &= \sqrt{0.64 \times 1.21} \\ &= \sqrt{0.7744} \end{aligned}$$

$$r_{xy} = 0.88$$

7. Write the equations for two regression lines.

The equation for regression line of x on y is

$$x - \bar{x} = b_{xy} (y - \bar{y}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

The equation for regression line of y on x is

$$y - \bar{y} = b_{yx} (x - \bar{x}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

where b_{xy} and b_{yx} are regression co-efficient.

σ_x and σ_y are standard deviations of x and y .

$\bar{x}$ and $\bar{y}$ are arithmetic means of x and y .

8. Define multiple correlation coefficient.

In trivariate distribution, in which variables x_1, x_2 & x_3 has n observations, the multiple correlation coefficient between x_1 and joint effect of x_2 and x_3 is called Multiple Correlation Coefficient.

$$R_{1.23} = \sqrt{\frac{x_{12}^2 + x_{13}^2 - 2x_{12}x_{23}x_{13}}{1 - x_{23}^2}}$$

The multiple correlation co-efficient $R_{1.23}$ is the

correlation coefficient between x_1 and estimated value of x_1 ($\hat{x}_1$).

PART-B

11 5 marks questions [Any 4]

9. Explain the scales of measurement.

Scales of measurement:

The process of measuring the object in specific units is called Measurement of Scale or Scales of Measurement.

Types of Scales:

i. Nominal Scale:-

A scale used to identify an object (names) by numbers is called Nominal Scale.

eg:- Roll number of students.

ii. Ordinal Scale:-

A scale that identifies the object in the form of relations ($<$, $>$, $=$)

eg:- Rank of students.

iii. Interval Scale:-

A scale that identifies an object with equal intervals is called Interval Scale.

eg:- 10-20, 20-30 and 30-40 are the marks obtained by students in a particular subject.

iv. Ratio Scale:-

A scale that identifies the object in terms of multiple way is called Ratio Scale.

eg:- The price of commodities in 2014 is 2 times than its price in 2000.

10. Explain the types of Graphs used for presentation of frequency distribution.

5. Define Standard Deviation.

The standard deviation is defined as positive root of arithmetic mean of squared deviations, when deviations are taken from mean.

Standard deviation is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \text{ for ungrouped data.}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N}} \text{ for grouped data.}$$

6. Given $b_{xy} = 0.64$ & $b_{yx} = 1.21$, then find correlation co-efficient.

Correlation co-efficient is given by

$$\begin{aligned} r_{xy} &= \sqrt{b_{xy} \cdot b_{yx}} \\ &= \sqrt{0.64 \times 1.21} \\ &= \sqrt{0.7744} \end{aligned}$$

$$r_{xy} = 0.88$$

7. Write the equations for two regression lines.

The equation for regression line of x on y is

$$x - \bar{x} = b_{xy} (y - \bar{y}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

The equation for regression line of y on x is

$$y - \bar{y} = b_{yx} (x - \bar{x}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

where b_{xy} and b_{yx} are regression co-efficient.

σ_x and σ_y are standard deviations of x and y .

$\bar{x}$ and $\bar{y}$ are arithmetic means of x and y .

8. Define multiple correlation coefficient.

In trivariate distribution, in which variables x_1, x_2 & x_3 has n observations, the multiple correlation coefficient between x_1 and joint effect of x_2 and x_3 is called Multiple Correlation Coefficient.

$$R_{1.23} = \sqrt{\frac{x_{12}^2 + x_{13}^2 - 2x_{12}x_{23}x_{13}}{1 - x_{23}^2}}$$

The multiple correlation co-efficient $R_{1.23}$ is the

correlation coefficient between x_1 and estimated value of x_1 ($\hat{x}_1$).

PART-B

11 5 marks questions [Any 4]

9. Explain the scales of measurement.

Scales of measurement:

The process of measuring the object in specific units is called Measurement of Scale or Scales of Measurement.

Types of Scales:

i. Nominal Scale:-

A scale used to identify an object (names) by numbers is called Nominal Scale.

eg:- Roll number of students.

ii. Ordinal Scale:-

A scale that identifies the object in the form of relations ($<$, $>$, $=$)

eg:- Rank of students.

iii. Interval Scale:-

A scale that identifies an object with equal intervals is called Interval Scale.

eg:- 10-20, 20-30 and 30-40 are the marks obtained by students in a particular subject.

iv. Ratio Scale:-

A scale that identifies the object in terms of multiple way is called Ratio Scale.

eg:- The price of commodities in 2014 is 2 times than its price in 2000.

10. Explain the types of Graphs used for presentation of frequency distribution.

Graphs used for presentation of frequency distribution are:

- i) Histogram
- ii) Frequency Polygon
- iii) Frequency Curves
- iv) Ogives

i) Histogram:

Histogram is a set of adjacent rectangles, whose height and width are proportional to class frequency and the width of the class frequency, such that area of the histogram is proportional to the total frequency. It is constructed for continuous frequency distribution.

ii) Frequency Polygon:

It is used to represent either continuous or discrete frequency distribution. Here, the total area under the figure is proportional to the total frequency.

iii) Frequency Curve:

It is a smooth free hand curve drawn along the frequency polygon by eliminating random or erratic fluctuations to show the regular movement of frequencies.

iv) Ogives:

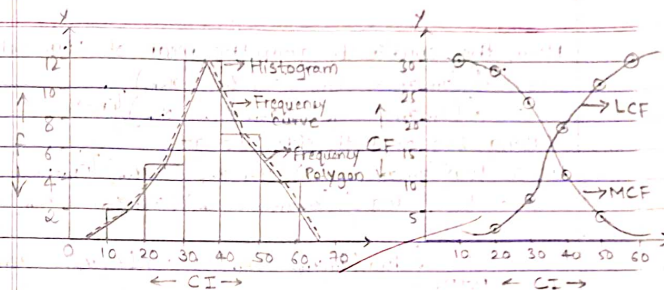
The curves drawn for cumulative frequencies against the class limits are called cumulative frequency curves or ogives.

They are of 2 types:

- a) Less than ogive.
- b) More than ogive.

eg:-

CI	f	Less than ogive		More than ogive	
		UL	LCF	LL	McF
10-20	2	20	2	10	30
20-30	5	30	7	20	28
30-40	12	40	19	30	23
40-50	7	50	26	40	11
50-60	4	60	30	50	1



11. With usual notations, show that

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Let $x_{11}, x_{12}, \dots, x_{1n_1}$ are n_1 observations of first set.

$$G_1 = \sqrt[n_1]{x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1}}$$

$$= (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1})^{1/n_1}$$

Taking log on both sides, we get

$$\log G_1 = \log (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1})^{1/n_1}$$

$$= \frac{1}{n_1} \log (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1})$$

$$= \frac{1}{n_1} (\log x_{11} + \log x_{12} + \dots + \log x_{1n_1})$$

$$\log G_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \log x_{1i} \rightarrow (a)$$

$$G_1 = \text{Antilog} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \log x_{1i} \right) \rightarrow (1)$$

Let $x_{21}, x_{22}, \dots, x_{2n_2}$ are n_2 observations of second set.

$$G_2 = \sqrt[n_2]{x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}}$$

$$= (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2})^{1/n_2}$$

Taking log on both sides, we get

$$\log G_2 = \log (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2})^{1/n_2}$$

$$= \frac{1}{n_2} (\log x_{21} + \log x_{22} + \dots + \log x_{2n_2})$$

$$\log G_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \log x_{2i} \rightarrow (b)$$

$$G_2 = \text{Antilog} \left(\frac{1}{n_2} \sum_{i=1}^{n_2} \log x_{2i} \right) \rightarrow (2)$$

Now, the combined geometric mean of two sets is

$$G = \sqrt[n_1+n_2]{(x_{11} \cdot x_{12} \cdots x_{1n_1})(x_{21} \cdot x_{22} \cdots x_{2n_2})}$$

$$= \sqrt[n_1+n_2]{(x_{11} \cdot x_{12} \cdots x_{1n_1})(x_{21} \cdot x_{22} \cdots x_{2n_2})}$$

Taking log on both sides, we get

$$\log G = \log \sqrt[n_1+n_2]{(x_{11} \cdot x_{12} \cdots x_{1n_1})(x_{21} \cdot x_{22} \cdots x_{2n_2})}$$

$$= \frac{1}{n_1+n_2} \log \{ (x_{11} \cdot x_{12} \cdots x_{1n_1})(x_{21} \cdot x_{22} \cdots x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \{ \log(x_{11} \cdot x_{12} \cdots x_{1n_1}) + \log(x_{21} \cdot x_{22} \cdots x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \{ (\log x_{11} + \log x_{12} + \cdots + \log x_{1n_1}) + (\log x_{21} + \log x_{22} + \cdots + \log x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \left(\sum_{i=1}^{n_1} \log x_{1i} + \sum_{i=1}^{n_2} \log x_{2i} \right)$$

From eqs (A) & (B), the above eqⁿ can be written as

$$\log G = \frac{1}{n_1+n_2} (n_1 \log G_1 + n_2 \log G_2)$$

12. Define i) Positive correlation

ii) Negative correlation

iii) Zero correlation.

i) Positive correlation:

If the two variables correlated are moving in the same direction then, it is called positive correlation. Here, either both the variables increase or decrease.
eg:- Demand and Supply.

ii) Negative correlation:

If two variables are moving in opposite directions then, the correlation is called negative correlation i.e., if one variable decreases, the other variable increases.
eg:- Price and Demand.

iii) Zero correlation:

If there is no correlation, then it is said to be zero correlation. It is also called as Spurious correlation.
eg:- Marks scored by students and the amount of rainfall.

III. 10 marks questions 3 [any 3]

15. Define Median and Mode. Write the formula for calculation of Merits and Demerits of median and mode.

Median:-

For ungrouped data:

Median is defined as middle most observation in the arranged (ascending or descending order) given set.

* Computation Process:

If 'n' denotes number of given observations then we can compute median in following cases:

a) When n is odd no.

Median = the size of $(\frac{n+1}{2})^{\text{th}}$ observation in the given arranged set.

b) When n is even no.

Median = A.M of size of $(\frac{n}{2})^{\text{th}}$ and $(\frac{n+1}{2})^{\text{th}}$ observations in the given arranged set.

* For grouped data:

Median is defined as the value of variables which divides total frequency into two equal parts such that half of the part lies below that value and remaining half of the part lies above that value.

* Computation Method:

For discrete data:

Step 1) Compute a cumulative frequency (less than type)

Step 2) Compute $N/2$ where N is total of frequency.

Step 3) Using cumulative frequencies, find out the variable value corresponding to $(N/2)^{\text{th}}$ value.

Using cumulative frequencies locate median value where $(N/2)^{\text{th}}$ value lies.

For continuous data:

Step 1) Calculate less than cumulative frequencies

Step 2) Calculate $N/2$ where N is sum of frequencies

Step 3) Using cumulative frequencies, locate median value where $(N/2)^{\text{th}}$ value lies, then to find exact value of median,

the following formula is given:

$$\text{Median} = L + \left(\frac{N - cf}{f} \right) \cdot h$$

where, L = lower limit of median class

h = width of the class.

f = frequency of median class.

Cf = cumulative frequency just above median class.

N = sum of frequencies.

* Merits of Median and Mode:

- Median is simple to understand & easy to calculate.
- Median is not affected by extreme values.
- Median can be calculated for open-end classes without any assumptions.

* Demerits of Median and Mode:

- Median is not rigidly defined.
- It is not based on all the observations.
- It is not suitable for further mathematical treatments.

ii) Mode

- For ungrouped data:

Mode is that value of the variable which occurs more frequently in the given set of values.

- For grouped data:

Mode is that value of the variable which has highest frequency $\rightarrow$ Discrete data.

- # For continuous data, to compute mode we have to use following formula:

$$\text{Mode} = L + \left(\frac{f_m - f_0}{2f_m - f_0 - f_1} \right) h$$

First we have to find modal class in which highest frequency lies.

where, L = lower limit of modal class.

h = width of class interval.

f_m = frequency of modal class.

f_0 = frequency of previous to modal class.

f_1 = the next frequency of modal class

17. i) Derive the angle between two regression lines.
ii) Write the properties of regression lines.

i) Equation of regression of x on y is
 $(x - \bar{x}) = r \cdot \frac{\sigma_x}{\sigma_y} (y - \bar{y})$

and regression equation of y on x is
 $(y - \bar{y}) = r \cdot \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

Slope of these lines are $r \cdot \frac{\sigma_x}{\sigma_y}$ divided by $r \cdot \frac{\sigma_y}{\sigma_x}$ respectively.

If θ is acute angle between the lines of regression

$$\tan \theta = \frac{m_1 - m_2}{1 + m_1 m_2}$$

$$= \frac{r \cdot \frac{\sigma_x}{\sigma_y} - \frac{1}{r \cdot \frac{\sigma_y}{\sigma_x}}}{1 + r \cdot \frac{\sigma_x}{\sigma_y} \cdot \frac{1}{r \cdot \frac{\sigma_y}{\sigma_x}}}$$

$$= \frac{\frac{\sigma_y}{\sigma_x} (r^2 - 1)}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$= \frac{\frac{\sigma_y}{\sigma_x} (r^2 - 1)}{\frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_x^2}} \div \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}$$

$$= \frac{\sigma_y \sigma_x (r^2 - 1)}{\sigma_x^2 + \sigma_y^2}$$

$$\therefore \tan \theta = \left(\frac{r^2 - 1}{r} \right) \cdot \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

$$\theta = \tan^{-1} \left(\frac{r^2 - 1}{r} \right) \cdot \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

If θ is acute angle then

$$\theta = \tan^{-1} \left(\frac{r^2 - 1}{r} \right) \cdot \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2} \quad (\because r^2 \leq 1)$$

ii) Properties of Regression lines:

a) If $r=0$, the two variables are un-correlated or there is no correlation, the lines of correlation become perpendicular to each other.

if $r=0$, $\tan \theta = \infty \Rightarrow \theta = \pi/2$

b) If $r=\pm 1$ in this case, the lines of regression either coincide or parallel to each other but the lines of regression passes through the point $(\bar{x}, \bar{y})$.

They cannot be parallel, they are coinciding.

i.e., $r=\pm 1$, $\tan \theta = 0 \Rightarrow \theta = 0$ or π

18. Derive the expression for rank correlation coefficient.

Let x_i & y_i be the rank of the i^{th} individual in two characteristics each variable x & y takes values

$1, 2, \dots, n$ i.e., $x=1, 2, \dots, n$ & $y=1, 2, \dots, n$

$$\Rightarrow \bar{x} = \frac{\sum x_i}{n} = \frac{1+2+\dots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\text{Similarly } \bar{y} = \frac{\sum y_i}{n} = \frac{n+1}{2}$$

$$\Rightarrow \bar{x} = \bar{y} = \frac{n+1}{2} \rightarrow \text{--- (1)}$$

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

$$= \frac{1^2 + 2^2 + \dots + n^2}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \frac{n^2 + 2n + 1}{4}$$

$$= \frac{2n^2 + 3n + 1}{6} - \frac{n^2 + 2n + 1}{4}$$

$$= \frac{4n^2 + 6n + 2 - 3n^2 - 6n - 3}{12}$$

$$\sigma_x^2 = \frac{n^2 - 1}{12}$$

$$\text{Similarly } \sigma_y^2 = \frac{1}{n} \sum y_i^2 - \bar{y}^2 = \frac{n^2 - 1}{12}$$

$$\Rightarrow \sigma_x^2 = \sigma_y^2 = \frac{n^2 - 1}{12}$$

But in general if $x_i \neq y_i \Rightarrow d_i = x_i - y_i$

$$\text{i.e., } d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \quad (\because \bar{x} = \bar{y})$$

$$\frac{1}{n} \sum d_i^2 = \frac{1}{n} \sum [(x_i - \bar{x}) - (y_i - \bar{y})]^2$$

$$= \frac{1}{n} \sum (x_i - \bar{x})^2 + \frac{1}{n} \sum (y_i - \bar{y})^2 - 2 \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sigma_x^2 + \sigma_y^2 - 2 \text{Cov}(x, y)$$

$$= 2\sigma^2 - 2r\sigma^2$$

$$\frac{1}{n} \sum d_i^2 = 2\sigma^2(1-r)$$

$$r = 1 - \frac{1}{n} \frac{\sum d_i^2}{2\sigma^2}$$

$$= 1 - \frac{1}{n} \frac{\sum d_i^2}{2 \left(\frac{n^2 - 1}{12} \right)}$$

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Thus, the expression for rank correlation coefficient

is $r = \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$ and $r = \pm 1$

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

~~11/01/23~~

**S. B. Arts & K.C.P. Science College,
VIJAYAPUR- 586 103.**



ASSIGNMENT

For B.A./ B.Sc.^I Semester
2022 - 2023

Name of the Student Mehboob Makandar
Roll No. 69 R.C.U. Seat No. U15KMI22S0107
Subject Statistics

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	11/01/23	05	05	Mess S. Y. Nagole <u>Teacher</u>	
2					
3					
4					

PART - A

I

- ① Define quantitative data with an example.
The data which is numerically expressed is called Quantitative Data.

Eg: Height and weight of students

- ② Define statistical table.

The systematic arrangement of the classified data in rows and columns in the form of table. This table is known as statistical table.

- ③ Define Deciles and Quartiles

Deciles: Measures which divide the data into ten equal parts.

Quartiles: Measures which divide the data into four equal parts.

- ④ What is Kurtosis

Kurtosis is the measurement of flatness or peakedness of the distribution. It is measured by the Co-efficient β_2 or its derivation γ_2 are given by $\beta_2 = \frac{\mu_4}{\mu_2^2}$

- ⑤ Define standard Deviation.

The standard deviation is defined as positive root of arithmetic mean of squared deviations, when deviation are taken from mean.

- ⑥ Given $b_{xy} = 0.64$ & $b_{yx} = 1.21$, then find correlation Co-efficient

Correlation Co-efficient is given by

$$r_{xy} = \frac{b_{xy} \cdot b_{yx}}{\sqrt{0.64 \times 1.21}}$$
$$= \frac{0.7744}{\sqrt{0.7744}}$$

$$r_{xy} = 0.88$$

⑦ Write the equation for two regression lines
The equation for regression line of x on y is
$$x - \bar{x} = b_{xy} (y - \bar{y}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

The equation for regression line of y on x is
$$y - \bar{y} = b_{yx} (x - \bar{x}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

⑧ Define multiple correlation coefficient.
In trivariate distribution in which variable x_1, x_2 & x_3 has n observations, the multiple correlation coefficient between x_1 and joint effect of x_2 and x_3 is called multiple correlation coefficient.

PART - B.

⑨ Explain the scales of measurement
Scale of measurement

The process of measuring the object in specific units is called measurement of scale of scales
Types of scales

① Nominal scale :- A scale used to identify an object by number is called Nominal scale.

② Ordinal scale :- A scale that identifies the object in the form of relations ($<, >, =$)

③ Interval scale :- A scale that identifies an object with equal intervals is called Interval scale.

④ Ratio scale :- A scale that identifies the object in terms of multiple way is called Ratio scale.

⑩ Explain the types of graphs used for presentation of frequency distribution.

① Histogram :- Histogram is a set of adjacent rectangles whose height and width are proportional to class frequency and the width of the class frequency such that area of the histogram is proportional to the total frequency. It is constructed for Continuous frequency distribution.

② Frequency polygon :- It is used to represent either continuous or discrete frequency distribution. Here the total area under the figure is proportional to the total frequency.

③ Frequency Curve :- It is a smooth free hand curve drawn along the frequency polygon by eliminating random or erratic fluctuations to show the regular movement of frequencies.

④ Ogives :- The curves drawn for cumulative frequencies against the class limits are called Cumulative frequency curves or ogives.

Types of ogives

⑤ Less than ogives

⑥ More than ogives

11) With usual notations, show that

$$\log G_1 = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

Let $x_1, x_2, \dots, x_{n_1}$ are n_1 observations of first set

$$G_1 = \sqrt[n_1]{x_1 \cdot x_2 \dots x_{n_1}}$$

$$= (x_1 \cdot x_2 \dots x_{n_1})^{1/n_1}$$

Taking log on both sides, we get

$$\log G_1 = \log (x_1 \cdot x_2 \dots x_{n_1})^{1/n_1}$$

$$= \frac{1}{n_1} \log (x_1 \cdot x_2 \dots x_{n_1})$$

$$= \frac{1}{n} (\log x_{11} + \log x_{12} + \dots + \log x_{1n_1})$$

$$\log G_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \log x_{1i} \rightarrow (a)$$

$$G_1 = \text{Antilog} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \log x_{1i} \right) \rightarrow (1)$$

Let $x_{21}, x_{22}, \dots, x_{2n_2}$ are n_2 observations of second set

$$G_2 = \sqrt[n_2]{x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}}$$

$$= (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2})^{1/n_2}$$

Taking log on both sides, we get

$$\log G_2 = \log (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2})^{1/n_2}$$

$$= \frac{1}{n_2} (\log x_{21} + \log x_{22} + \dots + \log x_{2n_2})$$

$$\log G_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \log x_{2i} \rightarrow (b)$$

$$G_2 = \text{Antilog} \left(\frac{1}{n_2} \sum_{i=1}^{n_2} \log x_{2i} \right) \rightarrow (2)$$

Now the combined geometric mean of two sets is

$$G = \sqrt[n_1+n_2]{x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1} \cdot x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}}$$

$$= \{ (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1}) (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}) \}^{1/(n_1+n_2)}$$

Taking log on both sides, we get

$$\log G = \log \{ (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1}) (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}) \}^{1/(n_1+n_2)}$$

$$= \frac{1}{n_1+n_2} \log \{ (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1}) (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \{ \log (x_{11} \cdot x_{12} \cdot \dots \cdot x_{1n_1}) + \log (x_{21} \cdot x_{22} \cdot \dots \cdot x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \{ (\log x_{11} + \log x_{12} + \dots + \log x_{1n_1}) + (\log x_{21} + \log x_{22} + \dots + \log x_{2n_2}) \}$$

$$= \frac{1}{n_1+n_2} \left(\sum_{i=1}^{n_1} \log x_{1i} + \sum_{i=1}^{n_2} \log x_{2i} \right)$$

From eqn (a) & (b) The above eqn can be written as

$$\log G = \frac{1}{n_1+n_2} (n_1 \log G_1 + n_2 \log G_2)$$

(12) Define i) Positive correlation

ii) Negative correlation

iii) Zero correlation

i) Positive correlation:- If the two variables correlated are moving in the same direction then, it is called positive correlation. Here, either both the variables increase or decrease

Eq: Demand and supply

ii) Negative correlation:- If two variables are moving in opposite directions then, the correlation is called negative correlation. i.e. if one variable decreases the other variable increases.

Eq: Price and Demand

iii) Zero correlation:- If there is no correlation, then it is said to be zero correlation. It is also called as spurious correlation

Eq: Marks scored by students and the amount of rainfall

PART - C

(15) Define Median and Mode write the formula for calculation & merits and Demerits of median and mode

i) Median:- For ungrouped data.

Median is defined as middle most observation in the arranged (ascending or descending order) given set

Computation process:-

If n denotes number of given observations, then we can compute median in following cases:

a) when n is odd no.

Median = the size of $(\frac{n+1}{2})^{th}$ observation in the given arranged set

b) when n is even no.

Median = A.M. of size of $(\frac{n}{2})^{th}$ and $(\frac{n+2}{2})^{th}$ observation in the given arranged set

For grouped data:-

Median is defined as the value of variable which divides total frequency into two equal parts such that half of the part lies below that value and remaining half of the part lies above that value.

Computation Method

For discrete data:

- Step 1) Compute a cumulative frequency
 - Step 2) Compute $N/2$ where N is total of Frequency
 - Step 3) Using cumulative frequencies find out the variable value corresponding to $(N/2)^{th}$ value.
- Using cumulative frequencies locate median value where $(N/2)^{th}$ value lies.

For continuous data

- Step 1) Calculate less than cumulative frequencies
- Step 2) Calculate $N/2$ where N is sum of frequencies
- Step 3) Using cumulative frequencies, locate median value where $(N/2)^{th}$ value lies. Then to find exact value of median, the following formula is given

$$\text{Median} = L + \left(\frac{N - cf}{f} \right) h$$

* Merits of Median and Mode:

- It is simple to understand & easy to calculate
- It is not affected by extreme values
- It can be calculated for open end classes without any assumptions.

* Demerits of Median and Mode:

- It is not rigidly defined
- It is not based on all the observations
- It is not suitable for further mathematical treatments

i) Mode

For ungrouped data:-

Mode is that value of the variable which occurs more frequently in the given set of values

For grouped data

Mode is that value of the variable which has highest frequency $\rightarrow$ Discrete data

For Continuous data, to compute mode we have to use following formula:

$$\text{Mode} = L + \left(\frac{f_m - f_0}{2f_m - f_0 - f_1} \right) h$$

First we have to find modal class in which highest frequency lies

- (17) i) Derive the angle between two regression lines
- ii) Write the properties of regression lines

1) Equation of regression of x on y is

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

and regression equation of y on x is

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Slope of these lines are $r \frac{\sigma_x}{\sigma_y}$ and $r \frac{\sigma_y}{\sigma_x}$ respectively

If θ is acute angle between the lines of regression $\tan \theta = \frac{m_1 - m_2}{1 + m_1 m_2}$

$$= \frac{r \frac{\sigma_x}{\sigma_y} - 1}{1 + r \frac{\sigma_x}{\sigma_y} \cdot r \frac{\sigma_y}{\sigma_x}}$$

$$= \frac{\sigma_y}{\sigma_x} \left(r - \frac{1}{r} \right)$$

$$= \frac{\sigma_y}{\sigma_x} \left(\frac{r^2 - 1}{r} \right) = \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}$$

$$\sigma_y \sigma_x \left(\frac{r^2 - 1}{r} \right) = \sigma_x^2 + \sigma_y^2$$

$$\therefore \tan \theta = \left(\frac{r^2 - 1}{r} \right) \cdot \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

$$\theta = \tan^{-1} \left(\frac{r^2 - 1}{r} \right) \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

If θ is acute angle then (because $r \leq 1$)

$$\theta = \tan^{-1} \left(\frac{r^2 - 1}{r} \right) \frac{\sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

11. Properties of Regression line

a) If $r=0$, the two variables are uncorrelated or there is no correlation the lines of correlation become perpendicular to each other.

If $r=0$, $\tan \theta = \infty \Rightarrow \theta = \pi/2$.

b) If $r = \pm 1$ in that case the lines of regression either coincide or parallel to each other but the lines of regression pass through the point $(\bar{x}, \bar{y})$.

$\therefore$ They cannot be parallel, they are coinciding.

h.e. $r = \pm 1$, $\tan \theta = 0 \Rightarrow \theta = 0$ or π .

18) Derive the expression for rank correlation coefficient. Let x_i & y_i be the ranks of the i^{th} individual in two characteristics each variable x & y takes values

1, 2, ..., n i.e. $x = 1, 2, \dots, n$ & $y = 1, 2, \dots, n$.

$$\Rightarrow \bar{x} = \frac{\sum x_i}{n} = \frac{1+2+\dots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\text{Similarly } \bar{y} = \frac{\sum y_i}{n} = \frac{n+1}{2}$$

$$\Rightarrow \bar{x} = \bar{y} = \frac{n+1}{2} \rightarrow (1)$$

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

$$= \frac{1^2 + 2^2 + \dots + n^2}{n} - \left(\frac{n+1}{2} \right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \frac{n^2 + 2n + 1}{4}$$

$$\frac{2n^2 + 3n + 1}{6} - \frac{n^2 + 2n + 1}{4}$$

$$\frac{4n^2 + 6n + 2 - 3n^2 - 6n - 3}{12}$$

$$\sigma_x^2 = \frac{n^2 - 1}{12}$$

$$\text{Similarly } \sigma_y^2 = \frac{1}{n} \sum y_i^2 - \bar{y}^2 = \frac{n^2 - 1}{12}$$

$$\sigma_x^2 = \sigma_y^2 = \frac{n^2 - 1}{12}$$

But in general if $x_i \neq y_i \Rightarrow d_i = x_i - y_i$

i.e. $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$ ($\because \bar{x} = \bar{y}$)

$$\frac{1}{n} \sum d_i^2 = \frac{1}{n} \sum [(x_i - \bar{x}) - (y_i - \bar{y})]^2$$

$$= \frac{1}{n} \sum (x_i - \bar{x})^2 + \frac{1}{n} \sum (y_i - \bar{y})^2 - \frac{2}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sigma_x^2 + \sigma_y^2 - 2 \operatorname{cov}(x, y)$$

$$\frac{1}{n} \sum d_i^2 = \frac{\sigma_x^2 - 2r\sigma_x^2}{\sigma_x^2 (1-r)}$$

$$r = 1 - \frac{\frac{1}{n} \sum d_i^2}{2\sigma_x^2}$$

$$= 1 - \frac{\frac{1}{n} \sum d_i^2}{2 \left(\frac{n^2-1}{12} \right)}$$

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Thus the expression for rank correlation co-efficient
 is $r = \rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$ and $r = \pm 1$

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

~~Ex 10~~
 11/01/23

**S. B. Arts & K.C.P. Science College,
VIJAYAPUR- 586 103.**



ASSIGNMENT

For B.A./ B.Sc.^{IIIrd} Semester
20²² - 20²³

Name of the Student Ravi. K. Nanaguli

Roll No. 290 R.C.U. Seat No. U15KH21S0069

Subject Statistics

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	09/02/23	05	05	J. D. Horakur + D. D. D. D.	
2					
3					
4					

2 marks

① Define Geometric distribution.

Random variable 'x' is said to have a geometric distribution if it assumes only non-negative values and its pmf is given by.

$$P(X=x) = \begin{cases} q^x p & , x=0, 1, 2, \dots ; 0 < p \leq 1, q=1-p. \\ 0 & , \text{otherwise} \end{cases}$$

② Define parameter & statistic

Parameter: A statistical constant of the population is called parameter.

Statistic: A function of sample value is called statistic.

③ Define Beta distribution of first kind.

A random variable 'x' as continuous beta distribution of first kind if $f(x) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}$; $0 < x < 1$ & $\alpha > 0, \beta > 0$

④ Write the pdf of beta distribution of 2nd kind.

$$\text{pdf: } \frac{x^{\alpha-1}}{B(\alpha, \beta) (1+x)^{\alpha+\beta}} ; x > 0, \alpha > 0, \beta > 0$$

⑤ Write the limiting form of the t distribution.

$$\lim_{n \rightarrow \infty} f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

⑥ Define Rectangular distribution (discrete & continuous).
Discrete:

A random variable 'x' is said to have a discrete random uniform distribution over the range [1, n] if its pmf is expressed as follows.

$$P(X=x) = \begin{cases} \frac{1}{n} & , x=1, 2, \dots, n \\ 0 & \text{on } \dots \end{cases}$$

Here 'n' is known as the parameter of the distribution. 'n' lies in set of all positive integers.

- (7) Define Cauchy's distribution.
Random variable x is said to have a standard Cauchy's distribution if its PDF is given by

$$f_x(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty$$

and x is termed as standard Cauchy's variate
more generally Cauchy's distribution with parameters 1 & μ as the PDF

$$g_y(y) = \frac{1}{\pi[1^2 + (y-\mu)^2]}, \quad \mu > 0, \quad -\infty < y < \infty$$

and we write $x \sim (1, \mu)$

- (8) Applications F distribution.

- (i) F-test for equality of two population variance
- (ii) F-test for testing the significance of an observed multiple correlation co-efficient
- (iii) F-test for testing the significance of observed simple correlation ratio.
- (iv) F-test for testing the linearity of regression

- (9) Define sequence.

A sequence is a function whose domain is the set of all natural number 'N' where as the range may be set 'A' that is $f: \mathbb{N} \rightarrow A$ is a sequence

- (10) Define Taylor's series expansion

If f is a function of continuously & n times differentiable in an interval $[x, x+h]$; then there exist some points in this interval denoted by $x+th$ for some $t \in [0, 1]$ such that

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \dots + \frac{h^{(n-1)}}{(n-1)!} f^{(n-1)}(x) + \frac{h^n}{n!} f^{(n)}(x+th)$$

and infinite order Taylor series expansion is given by

$$f(x+h) = \sum_{j=0}^{\infty} \frac{h^j}{j!} f^{(j)}(x).$$

5 marks:

① state & prove chebyshev's inequality:

If x is a random variable with mean μ and variance σ^2 then for any positive number k we have probability of

$$P\{|x-\mu| \geq k\sigma\} \leq \frac{1}{k^2} \quad \text{or} \quad P\{|x-\mu| < k\sigma\} \geq 1 - \frac{1}{k^2}.$$

proof: case I: x is a continuous random variable then by the definition

$$\sigma^2 = \sigma_x^2 = E(x - E(x))^2$$

$$\sigma_x^2 = E(x - \mu)^2$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

where $f(x)$ is pdf of x

$$\sigma_x^2 = \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx.$$

$$\sigma_x^2 \geq \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx \quad \text{--- (1)}$$

WKT

$$x < \mu - k\sigma \text{ or } x \geq \mu + k\sigma$$

$$|x - \mu| \geq k\sigma \quad \text{--- (2)}$$

Substituting this equation (2) we get.

$$\sigma_x^2 \geq \int_{-\infty}^{\mu - k\sigma} k^2 \sigma^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} k^2 \sigma^2 f(x) dx.$$

$$\sigma_x^2 \geq k^2 \sigma^2 \int P(x \leq \mu - k\sigma) + P(x \geq \mu + k\sigma)$$

$$P[|x - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

$$P[|x - \mu| \geq k\sigma] + P[|x - \mu| < k\sigma] = 1$$

$$P[|x - \mu| < k\sigma] = 1 - P[|x - \mu| \geq k\sigma]$$

$$P[|x - \mu| < k\sigma] = 1 - 1/k^2$$

case II: x is a discrete random variable in

case of discrete random variable, the proof follows exactly similarly on replacing integration by summation (Σ).

In particular if we take $k\sigma = c > 0$ then eqn

① & ② give respectively

$$P[|x - \mu| \geq c] \leq \frac{\sigma^2}{c^2}$$

$$P[|x - E(x)| \geq c] \leq \frac{\text{var } x}{c^2}$$

② Derive the χ^2 pdf distribution with mdf.

By method of MGF

If $x_1, x_2, \dots, x_n$ are n independent normal variates with mean $\mu_1, \mu_2, \dots, \mu_n$ and $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectively then

$$\chi^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 + \dots + \left(\frac{x_n - \mu_n}{\sigma_n}\right)^2$$

$$\chi^2 = \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

$$\chi^2 = \sum_{i=1}^n u_i^2 \quad \text{where } u_i = \frac{x_i - \mu_i}{\sigma_i}$$

MGF of χ^2 is

$$M_{\chi^2}(t) = M \sum_{i=1}^n u_i^2(t)$$

$$M_{\chi^2}(t) = M_{u_1^2}(t) + \dots + M_{u_n^2}(t)$$

$$M_{\chi^2}(t) = [M_{u_i^2}(t)]^n \quad \text{--- (1)}$$

Now

$$M_{u_i^2}(t) = E(e^{t u_i^2})$$

$$= \int_{-\infty}^{\infty} e^{t u_i^2} f(u_i) du_i$$

$$= \int_{-\infty}^{\infty} e^{t u_i^2} \frac{1}{\sqrt{2\pi} \sigma_i} e^{-\frac{1}{2} \left(\frac{u_i - \mu_i}{\sigma_i}\right)^2} du_i$$

$$\text{put } u_i = \frac{x_i - \mu_i}{\sigma_i}$$

diff w.r.t x_i

$$du_i = \frac{dx_i}{\sigma_i} \Rightarrow dx_i = \sigma_i du_i$$

$$= \int_{-\infty}^{\infty} e^{t u_i^2} \frac{1}{\sqrt{2\pi} \sigma_i} e^{-\frac{1}{2} u_i^2} \sigma_i du_i$$

$$= \int_{-\infty}^{\infty} e^{t u_i^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} u_i^2} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{(t - \frac{1}{2}) u_i^2} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u_i^2 \left(\frac{1 - 2t}{2}\right)} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u_i^2 \left(\frac{1 - 2t}{2}\right)} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{\pi}}{\sqrt{\frac{1 - 2t}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{2\pi}}{\sqrt{1 - 2t}}$$

$$= \frac{1}{(1 - 2t)^{1/2}}$$

$$M_{\chi^2}(t) = (1 - 2t)^{-n/2} \quad \text{--- (2)}$$

putting (2) in (1)

$$M_{\chi^2}(t) = [(1 - 2t)^{-1/2}]^n \quad \text{--- (3)}$$

$$M_{\chi^2}(t) = (1 - 2t)^{-n/2}$$

which is MGF of gamma distribution with parameter $\frac{1}{2}$ & $n/2$ hence by uniqueness theorem of MGF χ^2 is a gamma variate with parameter $\frac{1}{2}$ & $n/2$.

$$\therefore f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} (x/2)^{n/2 - 1}, \quad 0 < x < \infty$$

③ Derive the MGF of χ^2 distribution.

$$M_{\chi^2}(t) = E(e^{t \chi^2})$$

$$= \int_{-\infty}^{\infty} e^{t \chi^2} f(\chi^2) d\chi^2 \quad 0 < \chi^2 < \infty$$

$x = \chi^2$

$$M_{\chi^2}(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \int_0^{\infty} e^{tx} \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} x^{n/2 - 1} dx$$

$$= \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^{\infty} e^{-(\frac{1}{2} - t)x} x^{n/2 - 1} dx$$

$$= \frac{1}{2^{n/2} \Gamma(n/2)} \frac{\sqrt{\pi/2}}{\left[\frac{1 - 2t}{2}\right]^{n/2}}$$

$$= \frac{1}{2^{n/2} \sqrt{\pi/2}} \frac{\sqrt{\pi/2}}{\left[\frac{1 - 2t}{2}\right]^{n/2}} \left[\int_0^{\infty} e^{-ax} x^{n-1} dx = \frac{\Gamma(n)}{a^n} \right]$$

$$= \frac{1}{2^{n/2}} \frac{2^{n/2}}{(1 - 2t)^{n/2}}$$

$$M_{\chi^2}(t) = (1 - 2t)^{-n/2}$$

- ④ State and prove additive property of χ^2 distribution. The sum of independent χ^2 -variate is a χ^2 variate if $x_i (i=1, 2, \dots, k)$ are independent χ^2 variate with n_i degrees of freedom. The sum $\sum_{i=1}^k x_i$ is also a χ^2 variate with $\sum_{i=1}^k n_i$ degrees of freedom.

Proof: Since $\chi_1^2, \chi_2^2, \dots, \chi_k^2$ are χ^2 variates with $n_1, n_2, \dots, n_k$ degrees of freedom.

$$M_{\chi_1^2}(t) = (1-2t)^{-n_1/2}$$

$$M_{\chi_2^2}(t) = (1-2t)^{-n_2/2}$$

$$M_{\chi_k^2}(t) = (1-2t)^{-n_k/2}$$

MGF of χ^2 variates $\chi_1^2 + \chi_2^2 + \dots + \chi_k^2$ or $\sum_{i=1}^k \chi_i^2$

The sum of independent χ^2 variates

$$M_{\chi_1^2 + \chi_2^2 + \dots + \chi_k^2}(t) = M_{\chi_1^2}(t) \cdot M_{\chi_2^2}(t) \cdot \dots \cdot M_{\chi_k^2}(t)$$

$$= (1-2t)^{-(n_1+n_2+\dots+n_k)/2}$$

which is the MGF of χ^2 variate with theorem of MGF ($n_1+n_2+\dots+n_k$) df hence by uniqueness

Theorem of MGF $\sum_{i=1}^k \chi_i^2$ is χ^2 variate $\sum_{i=1}^k n_i$

$$\text{i.e. } \sum_{i=1}^k \chi_i^2 \sim \chi^2_{\sum_{i=1}^k n_i}$$

- ⑤ Weak law of large numbers.

Let $x_1, x_2, \dots, x_n$ be sequence of random variables and $\mu_1, \mu_2, \dots, \mu_n$ be their respective expectations or means and let

$$B_n = \sqrt{V(x_1 + x_2 + \dots + x_n)} < \infty \text{ then}$$

$$P \left[\left| \frac{x_1 + x_2 + \dots + x_n}{n} - \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n} \right| < \epsilon \right] \geq 1 - \eta$$

For all $n > n_0$ where n, η are arbitrary small positive numbers provided that

$$\lim_{n \rightarrow \infty} \frac{B_n}{n^2} \rightarrow 0$$

- ⑥ Obtain the mean and variance of negative binomial distribution.

$$\mu_1' = \left[\frac{d}{dt} M_X(t) \right]_{t=0}$$

$$\mu_1' = \frac{d}{dt} [(q-pe^t)^{-n}]_{t=0} \quad x^n = n x^{n-1}$$

$$= [-n (q-pe^t)^{-n-1} (pe^t)]_{t=0}$$

$$= np$$

$$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t) \right]_{t=0}$$

$$= [n (q-pe^t)^{-n-1} (pe^t)]_{t=0}$$

$$= n [pe^t (-n-1) (q-pe^t)^{-n-2} (pe^t) + (q-pe^t)^{-n-1} (pe^t)]_{t=0}$$

$$= n [-p^2 (-n-1) + p]$$

$$= np [np + p + 1]$$

$$\mu_2' = np + n(n+1)p^2$$

$$\text{variance } \mu_2 = \mu_2' - (\mu_1')^2$$

$$= np + n(n+1)p^2 - n^2p^2$$

$$= np + (n^2 + n)p^2 - n^2p^2$$

$$= np + p^2n^2 + p^2n - n^2p^2$$

$$= np + np^2$$

$$= np(1-p)$$

$$\text{variance} = npq$$

7

obtain MGF of exponential distribution.

$$M_X(t) = E(e^{tx})$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \int_0^{\infty} e^{tx} \cancel{\theta e^{-\theta x}} \theta e^{-\theta x} dx$$

$$= \theta \int_0^{\infty} e^{-(\theta-t)x} dx$$

$$= \theta \left[-\frac{e^{-(\theta-t)x}}{(\theta-t)} \right]_0^{\infty}$$

$$= \frac{\theta}{\theta-t} [-(-1)]$$

$$M_X(t) = \frac{\theta}{\theta-t} //$$

Done

B.L.D.E. Association's

**S. B. Arts & K.C.P. Science College,
VIJAYAPUR- 586 103.**



ASSIGNMENT

For B.A./ B.Sc. Semester
2022-2023

Name of the Student	Hariprasad Kulkarni		
Roll No.	15	R.C.U. Seat No.	UJ5KM2IS0221
Subject	Statistics Assignment.		

Assignment No.	Date	Marks Assigned	Marks Obtained	Name and Signature of Teacher	Remarks
1	09/02/23	05	05	J D Morakeni Dharbari	
2					
3					
4					

Assignment Questions

2 marks

1) Define Geometric distribution?

→ Random Variable 'x' is said to have geometric distribution if it assumes only non-negative values and its pmf is given by

$$P(X=x) = \begin{cases} q^x \cdot p, & x=0,1,2,\dots \quad 0 < p \leq 1, q=1-p \\ 0 & \text{Otherwise.} \end{cases}$$

2) Define Beta distribution of first kind?

→ A random variable 'x' has continuous beta distribution of first kind. If $f(x) = \frac{x^{\alpha-1} (1-x)^{\mu-1}}{\beta(\alpha, \mu)}$, $0 < x < 1$, $\mu > 0$, $\alpha > 0$.

Here, μ and α are parameters.

3) Write the pdf of Beta distribution of second kind?

→ $f(x) = \frac{x^{\alpha-1}}{\beta(\alpha, \mu) (1-x)^{\alpha+\mu}}$, $x > 0$, $\mu > 0$, $\alpha > 0$

Here μ & α are parameters.

$$\text{Since } \int_0^1 \frac{x^{\alpha-1}}{\beta(\alpha, \mu) (1-x)^{\alpha+\mu}} \cdot dx = 1$$

$$\int_0^1 \frac{x^{\alpha-1}}{(1-x)^{\alpha+\mu}} \cdot dx = \beta(\alpha, \mu) = \frac{\Gamma(\alpha) \Gamma(\mu)}{\Gamma(\alpha+\mu)}$$

4) Define Cauchy distribution?

→ Random variable 'x' is said to have standard Cauchy's distribution if its pdf is given by.

$$f_x(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty$$

and 'x' is termed as standard Cauchy variate more generally, Cauchy distribution with parameters λ & μ has the pdf

$$g_y(y) = \frac{\lambda}{\pi [\lambda^2 + (y-\mu)^2]}, \quad \lambda > 0, -\infty < y < \infty$$

and we write $x \sim C(\lambda, \mu)$.

5) Define parameter and statistic?

→ Parameter :- A statistical constant of the population is called parameter.

Statistic :- A function of a sample values is called statistic.

6) Write the limiting form for the t-distribution?

→
$$f(t) = \frac{1}{\sqrt{n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < t < \infty$$

$$\lim_{n \rightarrow \infty} f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

Which is pdf of standard normal variate.

Hence as degree of freedom 'n' becomes larger t-distribution tends to standard normal distribution is called limit form of t-distribution.

7) Write the application of F-distribution?

→ Applications of F-distribution.

- * F-test for equality of two population variances
- * F-test for testing the significance of an observed multiple correlation coefficient.
- * F-test for testing the significance of an observed sample correlation ratio.
- * F-test for testing the linearity of regression
- * F-test for equality of several means.

8) Define rectangular distribution?

→ Discrete Distribution

A random variable 'x' is said to have a discrete uniform distribution over the range [1, n] if its pmf is expressed as follows.

$$p(X=x) = \begin{cases} \frac{1}{n} & x=1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

Here 'n' is known as the parameter of the distribution and lies in the set of all positive integers. It is also known as discrete rectangular distribution.

9) Continuous distribution.

A random variable has continuous uniform distribution if $f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$ Here a & b are

Parameters.

9) What are the mean, variance of Gamma distribution?

$$f(x) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)} & \lambda > 0, 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Mean, $E(x) = \lambda$

Variance, $V(x) = \lambda$

10) Define t-distribution?

→ Let $x_1, x_2, x_3, \dots, x_n$ be a random sample size 'n' drawn from population with mean μ and variance σ^2 . Then Student's 't' defined by the statistic:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

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

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

11) Define R-code?

→ R is an open source programming language and software environment for statistical computing and graphics.

5 marks

1) State and prove Chebyshev's inequality?

→ If 'X' is a random variable with mean μ & variance σ^2 then for any positive number k we have

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or

$$P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

proof:- Case i) X is Continuous random Variable

Then by the definition $\sigma^2 = E(X - E(X))^2$

$$= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

where $f(x)$ is pdf of X

$$= \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx$$

$$\sigma^2 \geq \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx \quad (1)$$

WKT

$$x < \mu - k\sigma \quad \& \quad x > \mu + k\sigma$$

$$|x - \mu| \geq k\sigma \quad (2)$$

Substituting this in eqⁿ (1) we get

$$\sigma^2 \geq k^2 \sigma^2 \left[\int_{-\infty}^{\mu - k\sigma} f(x) dx + \int_{\mu + k\sigma}^{\infty} f(x) dx \right]$$

$$\geq k^2 \sigma^2 [P(X < \mu - k\sigma) + P(X > \mu + k\sigma)]$$

$$\sigma^2 \geq k^2 \sigma^2 P(|X - \mu| \geq k\sigma)$$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Also since

$$P(|X - \mu| \geq k\sigma) + P(|X - \mu| < k\sigma) = 1$$

$$P(|X - \mu| < k\sigma) = 1 - P(|X - \mu| \geq k\sigma)$$

$$P(|X - \mu| < k\sigma) = 1 - \frac{1}{k^2}$$

Case ii) 'X' is discrete random Variable

In case of discrete random Variable, the proof follows exactly similarly on replacing integration by summation.

In particular if we take $k\sigma = c > 0$

Then eq^s (1) & (2) give respectively

$$P(|X - \mu| \geq c) \geq \frac{\sigma^2}{c^2}$$

$$\& \quad P(|X - \mu| \geq c) \geq 1 - \frac{\sigma^2}{c^2}$$

$$\Rightarrow P(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

$$P(|X - E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

and

$$\Rightarrow P(|X - E(X)| < c) \geq 1 - \frac{\text{Var}(X)}{c^2}$$

2) Obtain MGF of Exponential distribution?

$$f(x) = 0 \cdot e^{-\lambda x}, \lambda > 0, x > 0$$

$$M_X(t) = E(e^{tx})$$

$$= \int_{-\infty}^{\infty} e^{tx} \cdot f(x) dx$$

$$= \int_0^{\infty} e^{tx} \cdot 0 \cdot e^{-\lambda x} dx$$

$$= 0 \int_0^{\infty} e^{(t-\lambda)x} dx$$

$$= 0 \left[\frac{e^{(t-\lambda)x}}{(t-\lambda)} \right]_0^{\infty}$$

$$M_X(t) = \frac{0}{0-t}$$

3) Derive the pdf of chi-square distribution with n.d.f?

3) Derive the moment generating function of χ^2 -distribution?

$$M_{\chi^2}(t) = E(e^{tx})$$

$$= \int_{-\infty}^{\infty} e^{tx} \cdot f(\chi^2) d\chi^2 \quad 0 < \chi^2 < \infty$$

$$x \sim \chi^2_{(n)}$$

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) dx$$

$$= \int_0^{\infty} e^{tx} \cdot \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} (x)^{\frac{n}{2}-1} dx$$

$$= \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \int_0^{\infty} e^{-(\frac{1-t}{2})x} x^{\frac{n}{2}-1} dx$$

$$= \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \left[\frac{e^{-(\frac{1-t}{2})x} x^{\frac{n}{2}-1}}{-(\frac{1-t}{2})} \right]_0^{\infty} = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \left[\frac{0 - (-1)^{\frac{n}{2}-1}}{(\frac{1-t}{2})} \right]$$

$$= \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \left[\frac{1}{(\frac{1-t}{2})^{\frac{n}{2}}} \right] \quad a = 1-t$$

$$M_X(t) = (1-t)^{-\frac{n}{2}}$$

4) State and prove additive property of χ^2 -distribution?

Statement:- The sum of independent χ^2 -variables is also a χ^2 variate more precisely if X_i ($i=1,2,3,\dots,k$)

are independent χ^2 variates with n_i degrees of freedom

then the sum $\sum_{i=1}^k X_i$ is also χ^2 variate with $\sum_{i=1}^k n_i$ degrees of freedom

Proof:- Since $\chi^2_1, \chi^2_2, \dots, \chi^2_k$ are χ^2 variates with $n_1, n_2, n_3, \dots, n_k$ degrees of freedom

$$M_{\chi^2_1}(t) = (1-t)^{-n_1/2}$$

$$M_{\chi^2_2}(t) = (1-t)^{-n_2/2}$$

$$M_{\chi^2_k}(t) = (1-t)^{-n_k/2}$$

$\chi^2_1 + \chi^2_2 + \dots + \chi^2_k$ or $(\sum_{i=1}^k \chi^2_i)$ is given by

$$M_{\chi^2_1}(t) + M_{\chi^2_2}(t) + \dots + M_{\chi^2_k}(t)$$

$$M_{\sum_{i=1}^k \chi^2_i}(t) = M_{\chi^2_1}(t) \cdot M_{\chi^2_2}(t) \cdot \dots \cdot M_{\chi^2_k}(t)$$

$$= (1-t)^{-\frac{n_1}{2}} \cdot (1-t)^{-\frac{n_2}{2}} \cdot \dots \cdot (1-t)^{-\frac{n_k}{2}}$$

$$= (1-t)^{-\frac{(n_1+n_2+\dots+n_k)}{2}}$$

which is the MGF of χ^2 variates with $(n_1+n_2+n_3+\dots+n_k)$ degrees of freedom

Hence by uniqueness theorem of MGF

$\sum_{i=1}^k \chi^2_i$ is χ^2 variate with $\sum_{i=1}^k n_i$ degrees of freedom

$$\text{i.e. } \sum_{i=1}^k \chi^2_i \sim \chi^2_{\sum_{i=1}^k n_i}$$

5) Obtain the mean & variance of negative Binomial distribution?

$$\text{Mean} = \mu'_1 = \left[\frac{d}{dt} M_X(t) \right]_{t=0}$$

$$= \frac{d}{dt} (q - p \cdot e^t)^{-r}$$

$$= [-r (q - p \cdot e^t)^{-r-1} (p \cdot e^t)]_{t=0}$$

$$\mu'_1 = rp$$

$$\mu'_2 = \left[\frac{d^2}{dt^2} M_X(t) \right]_{t=0}$$

$$= \frac{d^2}{dt^2} (q - p \cdot e^t)^{-r}$$

$$= \frac{d}{dt} [-r (q - p \cdot e^t)^{-r-1} (-p \cdot e^t)]_{t=0}$$

$$= r [p \cdot e^t (-r-1) (q - p \cdot e^t)^{-r-2} (-p \cdot e^t) + (q - p \cdot e^t)^{-r-1} (p \cdot e^t)]_{t=0}$$

$$= r \cdot (-p^2 (-r-1) + p)$$

$$\mu'_2 = rp(r+p)$$

$$= rp(r+p+1)$$

$$\mu_2 = rp + r(r+1)p^2$$

$$\text{Variance } (x) = \mu_2 = \mu_2' - (\mu_1')^2$$

$$= rp + r(r+1)p^2 - r^2p^2$$

$$= rp + (r^2+r)p^2 - r^2p^2$$

$$= rp + r^2p^2 + rp^2 - r^2p^2$$

$$= rp + rp^2$$

$$= rp(1+p)$$

$$\mu_2 = rpq \quad [\because 1+p=q]$$

6) Derive the pdf of Chi-Square distribution with n ?
 By the method of MGF if $x_1, x_2, \dots, x_n$ are n independent random variables or normal variables with mean $\mu_1, \mu_2, \mu_3, \dots, \mu_n$ & variances is $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectively then $Q^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 + \dots + \left(\frac{x_n - \mu_n}{\sigma_n}\right)^2$

$$Q^2 = \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

$$Q^2 = \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \quad \text{where } u = \frac{x_i - \mu_i}{\sigma_i}$$

MGF of Q^2 is

$$M_{Q^2}(t) = M_{\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}}(t)$$

$$= M_{\frac{x_1^2}{\sigma_1^2}}(t) + M_{\frac{x_2^2}{\sigma_2^2}}(t) + \dots + M_{\frac{x_n^2}{\sigma_n^2}}(t)$$

$$M_{Q^2}(t) = [M_{\frac{x_1^2}{\sigma_1^2}}(t)]^n \quad \text{--- (1)}$$

Now,

$$M_{\frac{x_1^2}{\sigma_1^2}}(t) = E(e^{t \frac{x_1^2}{\sigma_1^2}})$$

$$= \int_{-\infty}^{\infty} e^{t \frac{x_1^2}{\sigma_1^2}} f(x_1) dx_1$$

$$= \int_{-\infty}^{\infty} e^{t \frac{x_1^2}{\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2} dx_1$$

$$\text{put } u_i = \frac{x_i - \mu_i}{\sigma_i}$$

diff wrt x_i

$$du_i = \frac{dx_i}{\sigma_i} \rightarrow dx_i = \sigma_i du_i$$

$$= \int_{-\infty}^{\infty} e^{t \frac{x_1^2}{\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2} dx_1$$

$$= \int_{-\infty}^{\infty} e^{t \frac{x_1^2}{\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u_i^2} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t \frac{x_1^2}{\sigma_1^2}} e^{-\frac{1}{2}u_i^2} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{u_i(t - \frac{1}{2})} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u_i(\frac{1}{2} - t)} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\sqrt{1-2t}} = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\sqrt{(1-2t)}} \sqrt{1-2t}$$

$$= \frac{1}{(1-2t)^{1/2}} = (1-2t)^{-1/2}$$

$$\Rightarrow M_{Q^2}(t) = (1-2t)^{-n/2} \quad \text{--- (2)}$$

Putting eq (2) in (1)

$$M_{Q^2}(t) = [(1-2t)^{-1/2}]^n$$

$$M_{Q^2}(t) = (1-2t)^{-n/2}$$

$$\because a = \frac{1-2t}{2} \quad x^2 = u^2 \quad \int_0^{\infty} e^{-ax^2} = \frac{\sqrt{\pi}}{\sqrt{a}}$$

which is MGF of gamma distribution with parameters $\frac{1}{2}$ & $\frac{n}{2}$. Hence by uniqueness theorem of MGF Q^2 is given variate with parameters $\frac{1}{2}$ & $\frac{n}{2}$

$$\therefore f(x^2) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x^2/2} (x^2)^{n/2-1} \quad 0 < x^2 < \infty$$

7) prove that weak law of large numbers?

→ Statement: Let $x_1, x_2, \dots, x_n$ be a sequence of random variables and $\mu_1, \mu_2, \dots, \mu_n$ be their respective expectations and let

$$B_n = \text{Var}(x_1 + x_2 + \dots + x_n) < \infty$$

$$\text{Then } P\left|\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} - \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n}\right| < \epsilon \rightarrow 1 \text{ as } n \rightarrow \infty$$

if $n > n_0$, where ϵ and n are arbitrary small positive numbers provided

$$\lim_{n \rightarrow \infty} \frac{B_n}{n^2} \rightarrow 0$$

proof using Chebyshev's inequality to the r.v.

$(X_1 + X_2 + \dots + X_n)/n$, we get for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + X_2 + \dots + X_n}{n} - E\left[\frac{X_1 + X_2 + X_3 + \dots + X_n}{n}\right]\right| \geq \epsilon\right\} \geq 1 - \frac{B_n}{n^2 \epsilon^2}$$

$$\begin{aligned} & \left\{ \text{Since } \text{Var}\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \right. \\ & \quad \left. = \frac{1}{n^2} \text{Var}(X_1 + X_2 + \dots + X_n) = \frac{B_n}{n^2} \right\} \end{aligned}$$

$$\Rightarrow P\left\{\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n}\right| \geq \epsilon\right\} \geq 1 - \frac{B_n}{n^2 \epsilon^2}$$

So far, nothing is assumed about the behaviour of B_n for indefinitely increasing values of n .

Since ϵ is arbitrary, we assume $\frac{B_n}{n^2 \epsilon^2} \rightarrow 0$, as n becomes indefinitely large.

Thus having chosen two arbitrary small +ve numbers ϵ and n , number n_0 can be found so that the inequality $\frac{B_n}{n^2 \epsilon^2} < n$ will hold for $n > n_0$. Consequently,

we shall have

$$P\left\{\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n}\right| \geq \epsilon\right\} \geq 1 - n,$$

$$\forall n > n_0(\epsilon, n).$$

This conclusion leads to the following important result, known as the (weak) Law of Large Numbers

$$\frac{B_n}{n^2} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ is fulfilled.}$$

Discrete