44. In rolling a biased die, the probability of getting 6 is twice that of not getting it. If such a die will be rolled 4 times then the probability of getting 6 at least once is ________ .

Ans. In rolling a biased die, the probability of getting 6 is twice that of not getting it. If such a die will be rolled 4 times then the probability of getting 6 at least once is $1-\binom{4}{0}\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^4=1-\frac{1}{81}=\frac{80}{81}$

45. The probability of failure at each trial is $\frac{2}{3}$ and the number of trials is 15 then the binomial distribution is ——— modal.

Ans. The probability of failure at each trial is $\frac{2}{3}$ and the number of trials is 15 then the binomial distribution is uni- modal.

46. The relationship between mean, median and mode of a standard normal distribution is ________

Ans. The relationship between mean, median and mode of a standard normal distribution is they are equal.

47. In terms of kurtosis, Poisson distribution is always ________

Ans. In terms of kurtosis, Poisson distribution is always Leptokurtic

48. If the mean of a Poisson distribution is 5 then it is ________ skewed.

Ans. If the mean of a Poisson distribution is 5 then it is positively skewed

49. If p = 0.3 then binomial distribution is ________ skewed.

Ans. If p = 0.3 then binomial distribution is positively skewed.

50. If the standard deviation of a Poisson distribution is 2.5 then the distribution has ________ number of modes.

Ans. If the standard deviation of a Poisson distribution is 2.5 then the distribution has only one number of modes.

(2/3 MARKS) CHAPTER WISE SHORT-TYPE QUESTIONS WITH ANSWERS

CORRELATION & REGRESSION

1. **Define correlation coefficient and mention any two of its properties.**

Ans. Correlation coefficient may be defined as the degree of linear relationship between two variables.

It is computed by the formula: $r_{xy}=\frac{Cov(x,y)}{\sigma_x\sigma_y}$

Properties:

(i) Correlation coefficient is independent of change of origin and scale.

(ii) The limits of the correlation coefficient are ±1 i.e. $-1 \le r \le 1$ or $|r| \le 1$.

2. **Explain the method of interpretation of the value of correlation coefficient.**

Ans. The interpretation of the value of correlation coefficient can be done by using the probable error.

Probable error of r = PE (r) = 0.6745 × Standard error of r

Standard error of r = SE (r) = $\frac{1-r^2}{\sqrt{n}}$

Interpretation:

(i) If $|r| > 6 \times PE(r)$ then r is highly significant i.e. there is a strong association between the variables.

(ii) If $|r| < PE(r)$ then r is insignificant i.e. there is a weak association between the variables.

(iii) If $PE(r) \le |r| \le 6 \times PE(r)$ then the variables are moderately correlated.

3. **Express the regression coefficients by using correlation coefficient.**

Ans. The normal equations for fitting the line of regression of Y on X of the form $y = a + bx$ are:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad .... \text{(i)}$$

and $\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$ .... (ii)

On solving these normal equations (i) and (ii) the value of b comes out to be :

$$b = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i x_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}$$

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

$$= \frac{Cov(x,y)}{\sigma_x^2} \Rightarrow b_{yx} = \frac{Cov(x,y)}{\sigma_x^2}$$

and similarly $b_{xy} = \frac{Cov(x,y)}{\sigma_y^2}$

Further it is known that

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow Cov(x,y) = r_{xy}\sigma_x\sigma_y$$

Hence $b_{yx} = \frac{r_{xy}\sigma_x\sigma_y}{\sigma_x^2} = r_{xy}\frac{\sigma_y}{\sigma_x}$ and

$$b_{xy} = \frac{r_{xy}\sigma_x\sigma_y}{\sigma_y^2} = r_{xy}\frac{\sigma_x}{\sigma_y}$$

4. Describe the conditions of the method of least squares for fitting the line of regression of Y on X.

Ans. According to method of least squares, the line of regression of y on x of the form y = a + bx is that straight line passing through the scatter diagram satisfying the conditions:

(i) The sum of deviations parallel to the Y-axis should be Zero.

(ii) The sum of squares of these deviations is the minimum.

5. Discuss the effect of change of origin and scale on correlation coefficient.

Ans. Let there be n pairs of observation (x_i, y_i) having correlation coefficient $r_{xy} = \frac{Cov(x,y)}{\sigma_x\sigma_y}$

Applyling change of origin and scale let the new variables be $u = \frac{x-A}{h}$ and $v = \frac{y-B}{k}$ where A, B, h and k are constants such that $h, k \neq 0$.

Then $x_i = A + hu_i$, $y_i = B + kv_i$, $\bar{x} = A + h\bar{u}$,

$\bar{y} = B + k\bar{v}$, $\sigma_x^2 = h^2\sigma_u^2, \sigma_y^2 = k^2\sigma_v^2$

$$Cov(x,y) = E[(x-\bar{x})(y-\bar{y})]$$

$$= E[(A + hu - A - h\bar{u})(B + kv - B - K\bar{v})]$$

$$= E[hk(u-\bar{u})(v-\bar{v})] = hk\, E[(u-\bar{u})(v-\bar{v})]$$

$$= hk\, Cov(u,v)$$

$$\text{So } r_{xy} = \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{hkCov(u,v)}{h\sigma_u k\sigma_v} = r_{uv}$$

Thus it is clear that coefficient of correlation is independent of change of origin and scale.

6. By the help of a numerical example, show that r = 0 does not necessarily mean independence of the variables.

Ans. Let us consider the following bi-variate data:

x =	− 3	− 2	− 1	1	2	3
y =	9	4	1	1	4	9

It can be clearly observed that $y = x^2$ i.e. x and y are not independent.

$n = 6$, $\Sigma x = 0$, $\Sigma y = 28$, $\Sigma x^2 = 28$, $\Sigma y^2 = 196$, and $\Sigma xy = 0$

$$\text{So } r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{6\times 0 - 0\times 28}{\sqrt{[6\times 28 - (0)^2][6\times 196 - (28)^2]}} = 0$$

Thus it is clear that $r = 0$ does not necessarily mean independence of the variables.

7. Define a regression coefficient and state any two properties of the regression coefficients.

Ans. A regression coefficient may be defined as the increment in the value of the dependent variable corresponding to unit increment in the value of the independent variable. Thus every bi-variate data with variables x and y has two regression coefficients denoted by b_{yx} and b_{xy} called the regression coefficient of y on x and the regression coefficient of x on y respectively.

Thus b_{yx} = increment in the value of y corresponding to unit increment in the value of x and b_{xy} = increment in the value of x corresponding to unit increment in the value of y.

Properties of regression coefficients:

(i) The sign of both the regression coefficients is always equal to the sign of the correlation coefficient.

(ii) The geometric mean of the regression coefficients is the correlation coefficient.

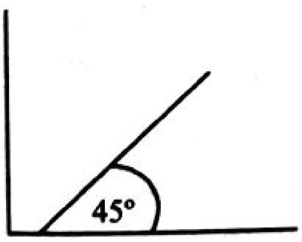
8. Explain perfect correlation in terms of scatter diagram.

Ans. There are two types of perfect correlation namely Perfect positive correlation and Perfect negative correlation.

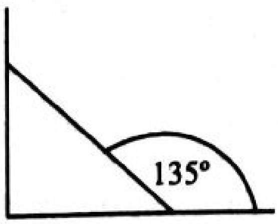
While inspecting the scatter diagram,

(i) if the points in the scatter diagram lie on a straight line making an angle of 45° with the positive x-axis, the variables have perfect positive correlation where the increase in the value of one variable results at an equal amount of increase in the value of the other variable.

(ii) On the other hand, if the points in the scatter diagram lie on a straight line making an angle of 135° with the positive x-axis, the variables have perfect negative correlation where the increase in the value of one variable results at an equal amount of decrease in the value of the other variable.



Perfect Positive Correlation



Perfect Negative Correlation

9. Prove that sum of the positive regression coefficients is always more than twice of the correlation coefficient.

Ans. The regression coefficients are given by

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Sum of the regression coefficients =

$$b_{yx} + b_{xy} = r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} = r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right)$$

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

Hence Sum of regression coefficients – 2 × Correlation coefficient

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

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

Since the regression coefficients are positive, r is positive. σ_y, σ_x being standard deviations are always positive and $(\sigma_y - \sigma_x)^2$ being the square of a real number is always positive. So $b_{yx} + b_{xy} - 2r > 0 \Rightarrow b_{yx} + b_{xy} - 2r$.

TIME SERIES ANALYSIS

10. What are the popular mathematical models assumed for the analysis of a time series?

Ans. The popular mathematical models assumed for the analysis of time series are:

(i) Additive model: $y_t = T_t + S_t + C_t + I_t$

(ii) Multiplicative model: $y_t = T_t \times S_t \times C_t \times I_t$

Where y_t = Value of the time series at time t

T_t = Effect due to trend at time t

S_t = Effect due to seasonal variation at time t

C_t = Effect due to cyclical fluctuation at time t

I_t = Effect due to irregular fluctuation at time t

11. Write any two demerits of the free hand curve method for measurement of trend.

Ans. Free hand curve method is an eye-inspection method having no mathematical basis.

The trend curve obtained by this method is subject to the understanding and experience of the person drawing the trend. So, different persons are likely to get different trend curves while adopting this method.

12. Write any two merits of moving average method in comparison to the semi-average method.

Ans. The semi-average method pre-supposes the presence of linear trend in every time series and thus fails to provide non-linear trend. But moving average method can provide linear as well as non-linear trends.

Trend by the semi-average method has the drawback of being unduly influenced by higher values in the series but moving average method has no such demerit.

13. Write a short note on random component of time series.

Ans. Random component of time series also called the irregular fluctuation is the result of factors that occur unexpectedly. In other words, effects of the factors no specific date of whose occurrences can be predicted are the cause of the random component. Such factors can arise due to two type of causes: (i) Natural causes and (ii) Socio-economic issues.

Examples of natural causes that give rise to irregular fluctuation are: Earthquake, Cyclone, Flood, Epidemic etc. However if there is a regularity of occurrence of any of such events, that cannot be considered as irregular fluctuation. For example the coastal areas of Odisha are regularly affected by floods during every rainy season. Such an event should be considered as a cause for seasonal variation instead of being included in the random component.

Examples of Socio-economic issues that may give rise to irregular fluctuation are: War, Strike, Lock-out, Terrorist attack, Communal riot etc.

14. What is the general equation of a second degree trend and what are the normal equations required to be solved for fitting it to a time series.

Ans. The general equation of a second degree trend is given by : $y = a + bt + ct^2$ where a, b, c are constants and $c \neq 0$.

Here y = trend value, n = number of observations and t = variable representing time

The calculations can be simplified by suitably choosing the values of t such that $\sum_{i=1}^{n} t_i$ becomes 0.

If n is odd i.e $n = 2k + 1$ then $t_i = i^{th}$ point of time$-$ $(k+1)^{th}$ point of time.

If n is even i.e $n = 2k$ then

$$t_i = 2\times\left[i^{th}\text{point of time} - \left(k+\frac{1}{2}\right)^{th}\text{ point of time}\right]$$

The normal equations required to be solved for fitting this second degree trend to a time series are:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} t_i + c\sum_{i=1}^{n} t_i^2 \quad \ldots(i)$$

$$\sum_{i=1}^{n} y_i t_i = a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^3 \quad \ldots(ii)$$

$$\sum_{i=1}^{n} y_i t_i^2 = a\sum_{i=1}^{n} t_i^2 + b\sum_{i=1}^{n} t_i^3 + c\sum_{i=1}^{n} t_i^4 \quad \ldots(iii)$$

15. Describe the steps in the method of semi averages for measurement of trend.

Ans. Assumption : It is assumed that the time series has a linear trend.

Various steps in this method are :

Step – 1 :

The observations are divided into two disjoint groups each containing equal number of consecutive.

Observations : However if the total number of observations is odd then the middle observation is ignored and the two groups are formed by dividing the remaining observations.

Step – 2 :

The average of each group is computed by using arithmetic mean. These two averages are called the semi averages. A semi average indicates the trend value for the mid point of time of the corresponding group.

Step – 3 :

The two semi averages are plotted against the mid points of time of the respective groups and the two points so plotted on the graph are joined by straight line. This straight line extended in both directions represents the trend for the given time series.

Step – 4 :

The y-coordinate of each point against respective points of time on the trend gives the trend value for corresponding point of time.

16. Describe the relative merits and demerits of moving average method compared with the least square method.

Ans. (i) Moving average method is more flexible in comparison to least square method.

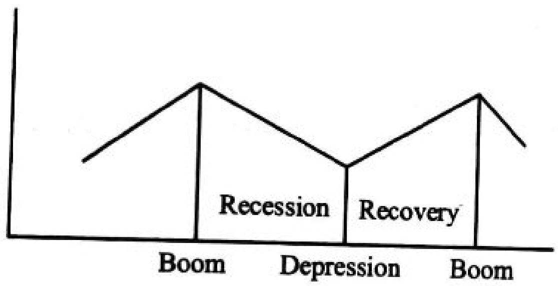
(ii) Least square method gives trend values for every point of time in the series but moving average method does not give all the trend values.

(iii) The trend obtained by the least square method is expressed as a function of time. As a result forecasting is better in case of least square method as compared to that by moving average method.

17. Explain a business cycle using the suitable diagram.

Ans. Every business passes through 4 stages Boom – Recession – Depression – Recovery – Boom which is called a business cycle. Boom refers to the period in which the business experiences the maximum profit level and Depression is the period having the minimum level of profit. The period from Boom to Depression in which the profit level goes on decreasing is called the stage of Recession and the period from Depression to the next Boom where the profit level again moves on increasing to reach a highest level is termed as the Recovery stage. The time interval between two consecutive booms is called the period of the cycle provided it is more than one year.

The diagram given below presents a business cycle.



18. Fit a linear trend to the time series given below by the method of least squares and estimate the trend value for 2017.

Year:	2011	2012	2013	2014	2015	2016
Sales in Crores	8	5	6	10	8	11

Ans. Let the equation of the linear trend be $y = a + bt$ where a and b are constants to be determined by solving the following normal equations:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} t_i \quad ...(i)$$

$$\sum_{i=1}^{n} y_i t_i = a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 \quad ... (ii)$$

Here y_i = Observed value of the time series at time t_i.

$t_i = 2\times [i^{th}$ Year – 2013.5], $n = 6$

Computation of trend values of a linear trend fitted by least square method

Year	Sales in Crores (y)	t	t^2	yt	Trend value in crores
2011	8	– 5	25	– 40	7.875
2012	5	– 3	9	– 15	7.925
2013	6	– 1	1	– 6	7.975
2014	10	1	1	10	8.025
2015	8	3	9	24	8.075
2016	11	5	25	55	8.125

Total: $\Sigma y = 48$ $\Sigma t = 0$ $\Sigma t^2 = 70$ $\Sigma yt = 28$

Using the normal equations (i) and (ii) we have: $48 = 6a \Rightarrow a = 8$ and $28 = 70b \Rightarrow b = 0.025$

Hence the trend equation becomes: $y = 8 + 0.025$ t. Putting different values of t the corresponding values of the trend have been obtained and are presented in the table above.

It is evident from the table that the yearly increment in the trend value is 0.050. Hence the trend value for 2017 will be: 8.125 + 0.050 = 8.175 crores

19. Explain with suitable examples, seasonal variation as a component of time series.

Ans: Seasonal variation in a time series indicates the changes in the value of the time series observed at uniform and regular interval of less than one year. It mainly occurs due to two types of causes such as:

(i) natural causes and

(ii) social customs and traditions.

Some of the examples of seasonal variations due to natural causes are:

Increase in the sale of refrigerators, air-coolers, ice-cream, cold drinks etc during the summer season.

Increase in the sale of woolen garments, blankets etc during winter season.

Decrease in the price of agricultural commodities during their respective harvesting season.

Some of the examples of social customs and traditions due to which seasonal variation may occur are:

Increase in the sale of fire crackers during Diwali.

Increase in the sale of readymade garments and other clothes during Dushera.

Increase in the sale of uniforms, books, copies etc during the admission season.

The study of seasonal variation is important for the producer, the seller as well as for the consumer. On the basis of seasonal variation, the producer plans to regulate the production. The seller makes plan to order and maintain stock on the basis of seasonal variation and the consumer plans to purchase goods at the lowest price which depends upon the seasonal variation. For example the consumer can purchase agricultural commodities during the harvesting season or can purchase items like refrigerator, air conditioner etc during winter season by availing off-season discount.

INDEX NUMBER

20. What is the simplest form of an index number and how is it computed ?

Ans. The simplest form of an index number is a price relative. It is a univariate index which indicates the relative change in the price of a commodity with respect to time or space or due to any other characteristic. It is computed by the formula:

Price relative of a commadity

$$= \frac{\text{Price of the commodity in the current period}}{\text{Price of the commodity in the base period}} \times 100$$

It is a pure number free from unit of measurement.

21. Examine TRT for simple GM of relatives method.

Ans. The formula for index number by the simple GM of relatives method is given by:

$$P_{01} = \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_1}{p_0}\right)\right]$$

$$\text{So } P_{10} = \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_0}{p_1}\right)\right]$$

$$\text{Hence } P_{01} \times P_{10} = \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_1}{p_0}\right)\right] \times \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_0}{p_1}\right)\right]$$

$$= \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_1}{p_0}\right) + \frac{1}{n}\Sigma \log\left(\frac{p_0}{p_1}\right)\right]$$

$$= \text{Antilog}\left[\frac{1}{n}\Sigma\left\{\log\left(\frac{p_1}{p_0}\right) + \log\left(\frac{p_0}{p_1}\right)\right\}\right]$$

$$= \text{Antilog}\left[\frac{1}{n}\Sigma \log\left(\frac{p_1}{p_0} \times \frac{p_0}{p_1}\right)\right]$$

$$= \text{Antilog}\left[\frac{1}{n}\Sigma 0\right] = \text{Antilog}(0) = 1$$

Since $P_{01} \times P_{10} = 1$,

the index number formula based on simple geometric mean of relatives satisfies Time Reversal Test.

22. Justify the statement, "Index numbers are economic barometers".

Ans. A barometer is the instrument used for measuring the atmospheric pressure. Index numbers are used to determine the relative changes in the price level, changes in the cost of living of the people, to decide the taxation policies, for decisions regarding minimum wage etc. In other words they are mainly used as economic indicators to measure the pressure of the economic policies of the state on the life of the people. So index numbers are called economic barometers.

23. Name the different tests for the adequacy of an index number formula.

Ans. Various tests for the adequacy of an index number formula are:

(i) Unit test, (ii) Time Reversal Test (TRT), (iii) Factor Reversal Test (FRT) and (iv) Circular Test.

24. Mention at least 4 uses of index number.

Ans. The uses of index numbers are:

(i) It is used by the government and other organizations for adjustment of DA and other allowances to be paid to employees of various categories.

(ii) Used by the trade unions for settlement of wages of industrial workers.

(iii) Used by the government for preparation of policies relating to taxes to be levied.

(iv) Used by the manufacturers to decide the pricing policies.

25. Explain the criteria for selection of commodities.

Ans. For the purpose of constructing an index number all the commodities in use should be taken into consideration. But practically it does not become possible to take into account all the commodities in use. Hence some representative commodities are selected for constructing the index number keeping in view the following points:

(i) The commodities should be representatives of the necessity, taste, habit, customs and traditions of the people chosen for the purpose of construction of the index number.

(ii) They should be stable in quality i.e the price level of the selected commodities should not be fluctuating too frequently.

(iii) All qualities of a commodity that are in common use should be included.

(iv) There is no specific rule regarding the number of commodities to be selected. However it should be neither very large nor very small. Number of commodities to be selected should be decided keeping in view the availability of data and resources.

26. What is the necessity for assigning weights? Also describe the different methods for assigning weights to various commodities.

Ans. It is clear that all commodities in use do not have the same importance. The importance of different commodities varies with necessity, age, sex, place, taste, habit, customs etc. Hence during the process of constructing an index number it becomes highly essential to assign a proper and rational number to each commodity so as to represent its relative importance in the group of commodities selected for the purpose. The rational number so associated with a commodity is called the weight of the commodity and this process of assigning rational numbers to all the selected commodities in accordance with their relative importance in the group is called the method of weighting.

The popularly used types of weights are mentioned below:

(i) Quantity weights: Where the quantity of a commodity consumed is taken as the weight of the commodity.

(ii) Value weights: Where the value of a commodity i.e. the total amount of money spent in purchasing the commodity is taken as the weight of the commodity.

Selection of the method of weighting depends on the availability of data.

27. Why Fisher's index is called the ideal index number?

Ans. The reasons for calling Fisher's index number as the Ideal index number are given below:

(i) It uses the geometric mean which is considered the best average for measurement of relative change.

(ii) It is the average of Laspeyre's and Paasche's index numbers. So the overestimation due to Laspeyre's method and the underestimation due to Paasche's method are likely to cancel out each other giving the true level of relative change.

(iii) Being a weighted index number, it satisfies three out of four tests of adequacy. Fisher's index number satisfies Unit test, Time Reversal Test and Factor Reversal Test.

SAMPLING TECHNIQUES

28. Mention the assumption under which simple random sampling is adopted.

Ans. Simple random sampling is adopted under the following assumptions:

(i) All the units in the population are homogeneous i.e. every unit of the population bears all the characteristics of the population.

(ii) Every unit has equal chance of being included in the sample.

29. Differentiate between random sampling and non-random sampling.

Ans. In non-random sampling, the units of the population are selected into the sample by the judgment of the person in charge of selecting the units. Hence the representativeness of the sample depends largely on the experience, understanding, honesty etc of the sampler. It can be used while dealing with population having considerably less number of units. It is mostly applicable to scientific studies or academic researches which involve small populations.

Random sampling method is that method of selecting the units where each population unit is assigned with a specific probability of being included in the sample. Statistical surveys involving large populations spread over wide area are generally conducted by this method. By selecting the sample by the random sampling method reduces the chance of personal bias due to the sampler and so it can be considered as a better representative of the population as compared to that obtained by non-random sampling method.

30. Mention at least four demerits of Census when compared to sampling.

Ans. The demerits of census as compared to sampling are as follows:

(i) In census method data are collected from every unit of the population but in sampling, data are collected from a subset of the population consisting of only some selected units. Hence in census data are to be collected from more units as compared to that in sampling.

(ii) Census requires more time as compared to that needed for sampling.

(iii) Census is more expensive in comparison to sampling.

(iv) Census provides a limited scope of study in comparison to sampling.

31. Describe the lottery method for selecting a random sample of n units from a population of N units without replacement.

Ans. The steps involved in conducting the lottery method for selection of units into the sample are as follows:

Step – I:

Each unit of the population is assigned with a unique serial number. Let the numbers be 1, 2, 3, …, N.

Step – II:

Each number should be written on a piece of paper in such a way that neither the number nor its impression should be visible from the reverse side of the paper. Further all the paper pieces should be identical in shape, size, colour, quality etc.

Step – III:

All the paper pieces should be identically folded in such a way that the number written on them should not be seen without unfolding them. These folded paper pieces are called lots.

Step – IV:

All the lots should be shuffled well and to be kept in a container.

Step – V:

If the number of units to be selected into the sample is n, then n number of units are to be selected from the container at random one after another either with replacement or without replacement and the serial number of the population unit mentioned in the lot is selected into the sample. If the lots are selected with replacement, the sample is called a simple random sample with replacement (SRSWR); but if the lots are selected without replacement, the sample is called a simple random sample without replacement (SRSWOR).

This method is easy to apply. But for large populations, this method is not suitable because it involves a cumbersome procedure. However it is a good method for dealing with small populations.

32. With the help of suitable example explain sampling unit.

Ans. At the time of dealing with large population spread over a wide area, it becomes difficult and involves more time and money for collecting data if units are selected at random directly from the population units. Hence in such cases, the population is divided into disjoint subsets on some basis such as geographical location, in terms of institutions etc. Each such subset of the population is termed as a sampling unit and the required sample is obtained by selecting some of these sampling units.

The concept of sampling units can be clear from the following example:

Suppose a total of 500, 000 students passed +2 examinations of CHSE, Odisha in 2018 and it is required to get a 1% sample i.e a sample of 5000 pass out students from about 1500 colleges spread over the 30 districts of odisha. If the 5000 students are selected at random directly out of the 5 lakh students, there is every possibility that a very small number of students may be selected from different colleges. As a result for collection of data, extensive touring is necessary which is likely to consume more money as well as more time and also a large manpower if time needs to be minimized.

On the other hand the entire population can be divided into 1500 colleges and 1% i.e 15 colleges may be selected at random and data are collected from all the students passed out from these colleges, then the process becomes more convenient as well as likely to be economical and time saving.

In such a situation each college is called a sampling unit.

33. With the help of a suitable example, explain the situation where complete enumeration cannot be done.

Ans. Suppose it is required to find the average life of an electric bulb produced in a batch of 1000 bulbs. Then the experiment consists of lighting all the bulbs under experimentation and to note down their individual lives and to calculate the average life.

In such a situation if census method is adopted, all the 1000 bulbs manufactured in the batch are to be lighted and at the end of the experiment all the bulbs will be damaged. Hence in such cases complete enumeration cannot be conducted because the unit is to be damaged during the process of experimentation for the sake of collection of data. Hence in such cases sampling is the only alternative method.

34. Describe the different types of errors in statistical surveys.

Ans. Two types of errors are likely to occur at the time of conducting statistical surveys: They are named as Sampling error and Non-sampling error.

Sampling error is inherent in sample surveys. Such errors are also called chance errors. It mainly occurs due to the reason that data are not collected from all the units in the population. However sampling error can be minimized by increasing the sample size and by adopting suitable method of randomization at the time of selecting the units.

Non-sampling error can be present in both sample surveys as well as census surveys. The reasons for occurrence of such errors are:

(i) Inadequate coverage of the units from which data are to be collected.

(ii) Faulty questionnaire and defective method of interviewing.

(iii) Lack of properly trained enumerators.

(iv) Improper method of substitution of units due to inability of collecting data from some particular units.

(v) Lack of proper inspection and cross examination of the data collected by the enumerators etc.

THEORITICAL DISTRIBUTIONS

35. Derive the mode of Poisson distribution.

Ans. Mode of any probability distribution is that value of the random variable which has the maximum probability. Let r be the mode of Poisson distribution. Then r is a positive integer such that P (X = r) is the maximum. This means that

$P(X = r) \geq P(X = r-1)$ ......(i)

and $P(X = r) \geq P(X = r+1)$ ...... (ii)

From (i) $\dfrac{P(X=r)}{P(X=r-1)} \geq 1 \Rightarrow \dfrac{e^{-\lambda}\lambda^r/r!}{e^{-\lambda}\lambda^{r-1}/_{(r-1)!}} \geq 1$

$\Rightarrow \dfrac{\lambda}{r} \geq 1 \Rightarrow \lambda \geq r \Rightarrow r \leq \lambda$ ....(iii)

From (ii) $\dfrac{P(X=r)}{P(X=r+1)} \geq 1 \Rightarrow \dfrac{e^{-\lambda}\lambda^r/r!}{e^{-\lambda}\lambda^{r+1}/_{(r+1)!}} \geq 1$

$\Rightarrow \dfrac{r+1}{\lambda} \geq 1 \Rightarrow r+1 \geq \lambda \Rightarrow \lambda - 1 \leq r$

Combining (iii) and (iv), mode of Poisson distribution is that integral value of r which satisfies the condition $\lambda - 1 \leq r \leq \lambda$.

Thus if λ is an integer, Poisson distribution has two modes λ and $\lambda-1$.

If λ is not an integer, the distribution has only one mode which is the integral part of λ.

36. Describe the area property of Normal distribution.

Ans. The area property of normal distribution is as follows :

(i) The area under a normal curve $= P(-\infty < X < \infty)$
$= P(-\infty < Z < \infty) = 1$ or 100 %

(ii) $P(-\infty < X < \mu) = P(\mu < X < \infty) = 0.5$ or 50 %
For a standard normal curve, $P(-\infty < Z < 0)$
$= P(0 < Z < \infty) = 0.5$ or 50 %

(iii) $P(\mu - \sigma < X < \mu + \sigma) = P(1 < Z < 1)$
$= 0.6826$ or 68.26%
$P(\mu - 1.96\,\sigma < X < \mu + 1.96\,\sigma)$
$= P(-1.96 < Z < 1.96) = 0.95$ or 95%
$P(\mu - 2\,\sigma < X < \mu + 2\,\sigma)$
$= P(-2 < Z < 2) = 0.9544$ or 95.44%
$P(\mu - 2.58\,\sigma < X < \mu + 2.58\,\sigma)$
$= P(-2.58 < Z < 2.58) = 0.99$ or 99%
$P(\mu - 3\,\sigma < X < \mu + 3\,\sigma)$
$= P(-3 < Z < 3) = 0.9973$ or 99.73%

(iv) If t is a positive real number then $P(Z \leq -t)$
$= P(Z \geq t) = P(Z < t) = 1 - (Z \geq t)$

37. Derive Poisson distribution as a limiting case of Binomial distribution.

Ans. Binomial distribution tends to a Poisson distribution under the conditions:

(i) $n \to \infty$, (ii) $p \to 0$ and

(iii) np is a constant denoted by λ.

Thus the probability generating function of Poisson distribution can be obtained by applying these conditions on the probability generating function of Binomial distribution.

Hence for a Poisson distribution,

$$P(X = r) = \lim_{n\to\infty} \binom{n}{r} p^r q^{n-r}$$

$$= \lim_{\substack{n\to\infty \\ p\to 0}} \left[\frac{n!}{r!(n-r)!} p^r q^{n-r} \right]$$

$$= \lim_{\substack{n\to\infty \\ p\to 0}} \left[\frac{n(n-1)(n-2)...(n-r+1)(n-r)!}{r!(n-r)!} \left(\frac{np}{n}\right)^r \left(1 - \frac{np}{n}\right)^{n-r} \right]$$

$$= \lim_{\substack{n\to\infty \\ p\to 0}} \left[\frac{n(n-1)(n-2)\dots(n-r+1)}{r!} \left(\frac{\lambda^r}{n^r}\right)\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-r} \right]$$

$$= \frac{\lambda^r}{r!} \lim_{n\to\infty} \left[\frac{n}{n} \times \frac{(n-1)}{n} \times \frac{(n-2)}{n} \times \dots \times \frac{(n-r+1)}{n} \right] \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n \times \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-r}$$

$$= \frac{\lambda^r}{r!} \lim_{n\to\infty} \left[1 \times \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{r-1}{n}\right) \right] e^{-\lambda} \times 1 = \frac{e^{-\lambda}\lambda^r}{r!}$$

Thus the probability generating function of Poisson distribution is $P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} (r = 0,1,2,3,\dots)$

38. Derive the mean of Binomial distribution.

Ans. The mean of any probability distribution with random variable X is $\mu_1' = E(X)$

So for binomial distribution,

$$\text{Mean} = \mu_1' = E(X) \sum_{r=0}^{n} rP(X=r) = \sum_{r=0}^{n} r\binom{n}{r} p^r q^{n-r}$$

$$= \sum_{r=0}^{n} r \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$= \sum_{r=1}^{n} \frac{n(n-1)!}{(r-1)!(n-r)!} p^r q^{n-r}$$

$$= np \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} p^{r-1} q^{n-r}$$

$$= np \sum_{r=1}^{n} \binom{n-1}{r-1} p^{r-1} q^{n-r}$$

$= np\,(q + p)^{n-1} = np$ (Because $q + p = 1$)

Thus mean of binomial distribution is np.

39. Prove that the mean and variance of a standard normal distribution are 0 and 1 respectively.

Ans. If $X \sim N(\mu, \sigma)$ i.e X is a normal variate having mean = μ and standard deviation = σ, then by definition, a standard normal variable is given by: $Z = \frac{X-\mu}{\sigma}$.

$Z \frac{X-\mu}{\sigma} \Rightarrow X = \mu + Z\sigma$ Hence $E(X) = E(\mu + Z\sigma)$

$\Rightarrow \mu = E(\mu) + \sigma\, E(Z)$

$\Rightarrow \mu = \mu + \sigma E(Z) \Rightarrow \sigma E(Z) = 0$

$\Rightarrow E(Z) = 0$ i.e. Mean of $Z = 0$

Further $V(X) = V(\mu + Z\sigma) \Rightarrow \sigma^2 = \sigma^2 V(Z)$

$\Rightarrow V(Z) = 1$

So the mean and variance of a standard normal distribution are 0 and 1 respectively.

40. What are the values of Quartile deviation and Mean deviation of a normal distribution with mean μ and standard deviation σ ?

Ans. For a normal distribution with mean = μ and standard deviation = σ,

Quartile deviation = $\frac{2}{3}\sigma$ and Mean deviation $= \frac{4}{5}\sigma$ (approximately)

41. Prove that Poisson distribution is a discrete probability distribution.

Ans. In Poisson distribution, the random variable is number of successes which cannot take any value other than positive integers. So Poisson distribution is a discrete distribution. Further the sum of all probabilities of Poisson distribution:

$$= \sum_{r=0}^{\infty} \frac{e^{-\lambda}\lambda^r}{r!} = e^{-\lambda} \sum_{r=0}^{\infty} \frac{\lambda^r}{r!}$$

$$= e^{-\lambda}\left[\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} \dots\dots\right] = e^{-\lambda} \times e^{\lambda} = 1$$

which indicates that it is a probability distribution. Thus Poisson distribution is a discrete probability distribution.

42. Explain how the value of *p* affects the skewness of a Binomial probability curve.

Ans. For a Binomial distribution with parameters n and p, $\mu_2 = npq$ and $\mu_3 = npq\,(q - p)$.

$$\text{So } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\{npq(q-p)\}^2}{(npq)^3} = \frac{(q-p)^2}{npq}$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{npq}}$$

The skewness of any distribution is decided by the sign of γ_1.

For a binomial distribution,

(i) if $2p < 1$ i.e. $p < ½$, $\gamma_1 > 0$

(ii) if $2p > 1$ i.e. $p > ½$, $\gamma_1 > 0$

(iii) if $2p = 1$ i.e. $p = ½$, $\gamma_1 > 0$

This indicates that Binomial distribution is positively skewed for $p < ½$; it is negatively skewed for $p > ½$ and the distribution is symmetrical for $p = ½$.