CSITNEPAL

statistics I)

2068,Fourth Batch

Cordial thanks to Mr. Anil Tiwari and Sandip Poudel

2012

Year: 2068

Group A

Attempt any Two (10 x 2 = 20)

1. **Write the importance of sampling over census. Describe systematic sampling. In a population with N = 6 the values of Y are 8, 3, 1, 11, 4 and 7. Calculate the sample Mean $\bar{y}$ for all possible simple random samples without replacement of size 2. Verify that $\bar{y}$is an unbiased estimate of $\bar{Y}$.**

The importance of sampling over census survey are as follows:-

- Sample survey has greater scope
- Sample survey takes minimum time cost and manpower
- Sample survey has greater economy.
- Sample survey has high quality of work.

Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equal probability method, in which every k^{th} element in the frame is selected where, k, the sampling interval is calculated as,

$$K = \frac{N}{n}$$

Where 'n' is the sample size and N is the population size.

Using this procedure each element in the population has a known and each element in the population has a known and equal probability of selection. This makes systematic sampling similar to simple random sampling.

Systematic sampling is to be applied only if the given population is logically homogeneous , because systematic sample units are uniformly distributed over the population.

Solution:

Samples in ascending order: 1,3,4,7,8,11

Here , N=6, all possible samples are:-

(1,3)(1,4)(1,7)(1,8)(1,11)(3,4)(3,7)(3,8)(3,11)(4,7)(4,8)(4,11)(7,8)(7,11)(8,11)

Sample no.	Sample values(Y)	Sample means($\bar{y}$)
1	(1,3)	2
2	(1,4)	2.5
3	(1,7)	4
4	(1,8)	4.5
5	(1,11)	6
6	(3,4)	3.5
7	(3,7)	5
8	(3,8)	5.5
9	(3,11)	7
10	(4,7)	5.5
11	(4,8)	6
12	(4,11)	7.5
13	(7,8)	7.5
14	(7,11)	9
15	(8,11)	9.5
Total		$\sum \bar{y}$ =85

Mean of sample means($\bar{\bar{y}} = \frac{\sum \bar{y}}{15} = 5.66$

we have to show $E(\bar{y}) = \bar{Y}$
to show this we proceed as follows.

Proof:

If $y_1, y_2 \ldots\ldots y_n$ be the 'n' same units drawn from the population units $Y_1, Y_2 \ldots\ldots Y_n$. Then , by definition

Sample mean($\bar{y}$)=$\sum \frac{y_i}{n}$ ... ... (1)

Population mean($\bar{Y}$) $= \sum \frac{Y_i}{N}$

By taking expectation on both size of equation (i) we get,

E($\bar{y}$) $= \frac{1}{n}\sum_{i=1}^{n} E(\bar{y_i}) - - - (ii)$

Here , y_i is the i[th] sample unit expected to occur the i[th] population unit $Y_I . \gamma\ i = 1,2,3 \ldots . n$ with equal probability of occurance $\frac{1}{N}. Then,$

$$E(y_i) = \sum_{i=1}^{n} Y_i . \frac{1}{N} = \frac{1}{N} \sum Y_i = \bar{Y}$$

Thus , equation becomes,

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}\bar{Y} = \frac{1}{n}X\ n.\bar{Y}$$

$\boxed{E(\bar{y}) = \bar{Y}}$ hence proved.

2. **The following data represent the operating times in hours for three types of scientific pocket calculators before a recharge is required:**

Calculator A	**4.9**	**6.1**	**4.3**	**4.6**	**5.3**	**5.5**
Calculator B	**5.4**	**6.2**	**5.8**	**5.5**	**5.2**	**4.8**
Calculator C	**6.4**	**6.8**	**5.6**	**6.5**	**6.3**	**6.6**

Use the Kruskal-Wallis test, at the 0.01 level of significance, to test the hypothesis that the operating times for all three calculators are equal.

3. **The following table shows the scores(Y) made by ten assembly line employees on a test designed to measure job satisfaction. It also shows the scores made on an aptitude test (X_1) and the number of days absent(X_2) during the past year (excluding vacations).**

Y	X_1	X_2
70	**6**	**1**
60	**6**	**2**
80	**8**	**1**
50	**5**	**8**
55	**6**	**9**

85	**9**	**0**
75	**8**	**1**
70	**6**	**1**
72	**7**	**1**
64	**6**	**2**

The summation values are as following:

$\sum Y = 681$, $\sum X1 = 67$ $\sum X2 = 26$ $\sum X_1Y = 467$ $\sum X_2Y = 1510$

$\sum X_1X_2 = 153$ $\sum Y^2 = 47455$ $\sum X_1^2 = 463$ $\sum X_2^2 = 158$

(i) **Calculate the least squares equation that best describes these three variables.**

(ii) **Predict the value of scores when aptitude test is 7 and number of days absent is 6.**

Group B

Answer any eight questions: (8 x 5 = 40)

4. Show that in simple random sampling without replacement sample mean is unbiased estimate of population mean.

we have to show $E(\overline{y}) = \overline{Y}$

to show this we proceed as follows.

Proof:

If $y_1, y_2 \ldots\ldots y_n$ be the 'n' same units drawn from the population units $Y_1, Y_2 \ldots\ldots Y_n$. Then , by definition

Sample mean($\overline{y}$)=$\sum \frac{y_i}{n}$ ... ... (1)

Population mean($\overline{Y}$) $= \sum \frac{Y_i}{N}$

By taking expectation on both size of equation (i) we get,

E($\overline{y}$) $= \frac{1}{n}\sum_{i=1}^{n} E(\overline{y_i}) - - - (ii)$

Here , y_i is the ith sample unit expected to occur the ith population unit $Y_I . \gamma\ i = 1,2,3 \ldots . n$ with equal probability of occurance $\frac{1}{N} . Then,$

$$E(y_i) = \sum_{i=1}^{n} Y_i . \frac{1}{N} = \frac{1}{N} \sum Y_i = \overline{Y}$$

Thus , equation becomes,

$$E(\overline{y}) = \frac{1}{n} \sum_{i=1}^{n} \bar{Y} = \frac{1}{n} X\ n . \bar{Y}$$

$\boxed{E(\bar{y}) = \bar{Y}}$ hence proved.

5. What do you mean by partial correlation coefficient? State the relationship between simple and partial correlation coefficient when there are three variables. If r_{12}= 0.5 , r_{23} = 0.1 and r_{13} = 0.4, compute $r_{12.3}$ and $r_{23.1.}$

Partial correlation coefficient is the relationship between two variables keeping other variables constant.

Let us consider three variables X_1,X_2,X_3.

Then ,

$r_{12.3}$ denotes partial correlation between X_1,&X_2 keeping effect of X_3 constant.

$r_{23.1.}$ denotes partial correlation between X_3,&X_2 keeping effect of X_1 constant.

$r_{13.2}$ denotes partial correlation between X_1& X_3 keeping effect of X_2 constant.

Correlation coefficient between two variables is called simple correlation coefficient. It is denoted by r_{xy}= r_{yx} =r to denote correlation coefficient between two variable X&Y.

Simple correlation is directly related with partial correlation. To know the partial correlation between the variables first we have to know the simple correlation.

For example:

$R_{12.3}$ denotes partial correlation between X_1 & X_2 keeping effect of X_3 constant which can be calculated as

$$r_{12.3}=\frac{(r_{12}-r_{13}.r_{23})}{\sqrt{1-(r_{13})^2}\ \sqrt{1-(r_{23})^2}}$$

where r_{12} ,r_{13} & r_{23} are the simple correlation coefficients.

Here,

We are given,

$r_{12=0.5,}$ r_{23}=0.1 & r_{13}= 0.4

Then,

We have,

$$r12.3=\frac{(r_{12}-r_{13}.r_{23})}{\sqrt{1-(r_{13})^2}\ \sqrt{1-(r_{23})^2}}$$

$$=\frac{0.5-0.1*0.4}{\sqrt{1-(0.4)^2}\ \sqrt{1-(0.1)^2}}$$

=-0.381

Similarly,

$r_{23.1}=\frac{(r_{23}-r_{12}.r_{13})}{\sqrt{1-(r_{12})^2}\ \sqrt{1-(r_{13})^2}}$

$=\frac{0.1-0.5*0.4}{\sqrt{1-(0.5)^2}\ \sqrt{1-(0.4)^2}}$

$=$ -1.01

6. Explain two stage sampling with sample mean and corresponding variance.

If the ultimate sample units are not available then multi stage sampling technique is suitable sampling technique.
Suppose the population consists of NM units
N=number of cluster having M units in each cluster
y_{ij}= value of the variable of interest in j^{th} second stage unit for the i^{th} first stage where j=1,2,…..,M &i=1,2,…….,N

$$\bar{Y}i. = \frac{1}{M}\sum_{i=1}^{M} y_{ij} = Mean\ per\ element\ in\ the\ j^{th}\ column.$$

$$\bar{Y} = \frac{1}{NM}\sum_{j=1}^{M}\sum_{i=1}^{N} y_{ij} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{i.} = population\ mean.$$

$$\bar{\bar{y}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i_0} - - - - - sample\ mean.$$

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} Y_{i.} - \bar{Y} - -$$

$$- variance\ between\ the\ first\ stage\ units\ mean.$$

$$S_w^2 = \frac{1}{N-1}\sum_{i=1}^{M} (Y_{ij} - \bar{Y}_{i.})^2 - -$$

$$- variance\ among\ the\ elements\ within\ cluster$$

$$\overline{v(y)} =$$

$$\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)Sw^2 - - - variance\ in\ two\ stage\ sampling.$$

7. Differentiate parametric and non parametric test.

	Parametric test	Non parametric test
1	Parametric test are those which are based on the assumption that the parent population is normal .i.e; those test whose model specific the condition about population parameter from which the samples are drawn.	N-p test are those test whose model doesn't specify the nature of the parent population from which samples are drawn.
2	Parametric test are designed to test statistical hypothesis and for stimulating population parameter	N-p test are designed only to test the statistical hypothesis.
3	Parametric test are applied only on the data which are in internal or ratio scale .	N-p test are applied only to test the data in ofdinal or nominal scale
4	Parametric test are most powerful test.	N-p test are less powerful test.

8. In an industrial production line, items are inspected periodically for defectives. The following is a sequence (from left to right) of defective items, D, and non defective items, N, Produced by this production line:

D	**D**	**N**	**N**	**N**	**D**	**N**	**N**	**D**	**D**
N	**N**	**N**	**N**	**N**	**D**	**D**	**D**	**N**	**N**
D	**N**	**N**	**N**	**N**	**D**	**N**	**D**		

Use run test with a significance level 0.05 to determine whether the defectives are occurring at random or not.

Solution:
Null hypothesis:(H_0)- The defectives occurring is at random order.
Alternative hypothesis (H_1): The defectives occurring is not in random order.
Test statistic: Under H_0 test statistic is, r=no. of runs.
Here sequence of given results in,
DD NNN D NN DD NNNNN DDD NN D NNNN D N D
Here, no.of runs r=13.
& n_1=11(no. of defectives)
n_2=17(no. of non- defectives)
critical value: for n_1=11& n_2=17 at 0.05 level of significance
$\overline{r}$=7 $\overline{r}$ =18

Decision: since no. of runs lies between $\overline{r} and\ \overline{r}$, H_0 is accepted i.e; The defectives occurring is at random order.

9. Use the sign test to see whether there is a difference between the numbers of days required days required to collect an account receivable before and after a new collection policy. Use the 0.05 significance level.

Before	**33**	**36**	**41**	**32**	**39**	**47**	**34**	**29**	**32**
After	**35**	**29**	**38**	**34**	**37**	**47**	**36**	**32**	**30**

Solution:

Null hypothesis(H_0):- $Md_1=Md_2$(There is no difference between no. of days required to collect as account receivable.

Alternative hypothesis(H_1)= $Md_1<Md_2$(There is difference between no. of days required to collect as account receivable.)

Test statistic: Under H_0, test statistic is obtained as follows:-

Before	**33**	**36**	**41**	**32**	**39**	**47**	**34**	**29**	**32**
After	**35**	**29**	**38**	**34**	**37**	**47**	**36**	**32**	**30**
	+	**-**	**-**	**+**	**-**	**0**	**+**	**+**	**-**

Number of + sign x(+)=4

Number of – sign x(-)=4

Number of tie(+)=1

Number of effective sample n_e=n(+)+n(-)

=4+4

=8

Test statistic:(k)= min{ n(+) + n(-)}

=min{4,5}

=4

Critical value: For 0.05 level of significance,

P-value(P_0)=$P(y \leq 4)$

=

Decision: **(please check yourself)**

10.A random sample of 200 married men, all retired, was classified according to education and number of children.

Education	**Number of Children**		
	0-1	**2-3**	**Over 3**
Elementary	**14**	**37**	**32**
Secondary	**19**	**42**	**17**
College	**12**	**17**	**10**

Test the hypothesis, at the 0.05 level of significance, that the number of children is independent of the level of education attained by the father.

Solution:

Null hypothesis(H_0): The number or children is independent of level of education attained by father.

Alternative hypothesis(H_1): The number of children is independent of level of education attained by father.

Calculation for χ^2

observed value	Expected value	E	(O-E)	$(O-E)^2/E$
14	83*(45/200)=18.675	18.675	-4.675	1.561116
37	83*(96/200)=39.84	39.84	-2.84	0.217989
32	83*(59/200)=24.485	24.485	7.515	1.764851
19	45*(78/200)=17.55	17.55	1.45	0.110658
42	96*(78/200)=37.44	37.44	4.56	0.495086
17	59*(78/200)=23.01	23.01	-6.01	2.124712
12	45*(39/200)=8.775	8.775	3.225	0.866719
17	96*(39/200)=18.72	18.72	-1.72	0.174024
10	59*(39/200)=11.505	11.505	-1.505	0.226503
Total				7.541656

$$\chi^2\ cal = \sum \frac{(O-E)^2}{E} = 7.541656$$

Degree of freedom (D.O.F) =(3-1)*(3-1)=4

Level of significance $(\propto) = 0.05$

Critical value $\chi^2_{cal} = 7.54$ $\chi^2_{0.05}$ =9.49

Decision:- since, $\chi^2\ cal < \chi^2_{tab}$, then H_0 is accepted. i.e, The number or children is independent of level of education attained by father.

11.Write Cobb-Douglas production function with interpretation of the regression coefficients.

Consider the following Non-linear model which consists more than two variables as,

$Y=\beta_1 X_2^{\beta 2}.X_3^{\beta 3} \ldots\ldots X_p^{\beta}.e^4 ----(i)$

Where, Y is dependent variable.

$X_2, X_3 \ldots\ldots$ are independent variable

$\beta_2, \beta_3 \ldots . \beta_p$ are regression coefficients 'μ' is error term & 'e' is the base of natural logarithm.

Taking natural log on both sides

$\log y = ln\beta_1 + \beta_2 lnX_2 + \beta_3 lnX_3 + \cdots .. \beta_n lnX_p + \mu — (ii)$

The model(ii) is linear parameter & can be estimated easily.

This model is useful in production function theory and is known as Cobb-Douglas production function. Here, we may suppose Y as output. X_2 may be labor input,X_3 may be capital input and so on .

In this model,

$P_k = \frac{\partial lny}{\partial lnX_k} = \frac{X_k}{Y}.\frac{\partial y}{\partial X_k}\beta_k$ measures the partial elasticity of the dependent variable w.r.t independent variable X_k i.e; β_k measures the percentage change in Y w.r.t percentage change in Y_k.

The sum $\sum_{i=2}^{\beta}\beta_i$ provides the information about return to scale. i.e the response of dependent variable output to a proportional changes in the explanatory variables(inputs)

This implies that the slope coefficient say β_k measures the partial elasticity of the dependent variable with respect to percentage change in X_k .

The sum $\sum_{i=2}^{P}\beta_i$ provides the information about the returns to scale i.e, the response of dependent variable output to a proportional change in the explanatory variable inputs.

1) If the sum is 1, then, there are constant returns to scale i.e, doubling the inputs will double the outputs, trebling the inputs will triple the output & so on.

2)If the sum is less than 1,then, there will be decreasing returns to scale. i.e; doubling the inputs will results less than double the outputs.

3) If the sum is greater than 1, then, there will be increasing returns to scale. i.e, doubling the inputs will result more than double the output.

12.**Suppose the residuals for a set of data collected over 8 consecutive time periods are as follows:**

Time Period:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Residuals:	**-4**	**-3**	**-3**	**-2**	**1**	**1**	**3**	**7**

Compute the first order autocorrelation.

S.N.	e_t	e_{t-1}	e_{t-2}	$e_t . e_{t-1}$	$e_t . e_{t-2}$	e_t^2 =103	et-et-1	$(e_t - e_{t-1=})^2$
1	-4					16		
2	-3	-4		12		9	1	1
3	-3	-3	-4	9	12	9	0	0
4	-3	-3	-3	9	9	9	0	0
5	1	-2	-3	-2	-3	1	3	9
6	1	1	-3	1	-3	1	0	0
7	3	1	-2	3	-6	9	2	4
8	7	3	1	21	7	49	4	16
Total				$\sum e_t . e_{t-1=}$53	$\sum e_t . e_{t-2}$ =16	e_t^2 =103		$\sum(e_t - e_{t-1})^2$ =30

∴ The first order autocorrelation coefficient is

$$\boldsymbol{\rho = \frac{\sum_{t=2}^{8} e_t . e_{t-1}}{\sum_{t=2}^{8} e_t^2} = \frac{53}{103} = 0.514}$$

13.Explain the term multicollinearity and describe a situation where the problem of multicollinearity arises?

If the model has several variables, it is likely that some of the explanatory variables will be approximately related. This property is known as multicollinearity. Multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the regression.

The situation where the problem of multicollinearity arises are:

1. If the coefficient of multiple determination(R^2) is very high but the test of individual regression coefficient found to be insignificant then there could be the problem of multicollinearity.
2. The high value of standard error associated with individual regression coefficient also indicate the presence of multicollinearity.
3. If inclusion of additional variables make substantial change in the value of individual regression coefficients then the problem of multicollinearity may be present.

However, all above methods just give the tentative idea about the existence of multicollinear problem but not sufficient indicators.