CSITNEPAL.COM

Statistic solution

2065(First page)

Cordial thanks to Mr. sandip poudel and Mr. Anil tiwari.

2012

2065

Group-A $(2 \times 10 = 20)$

1. Differentiate simple random sampling and stratified random sampling. Show that:

i) In simple random sampling without replacement (SRSWOR), the sample mean is an unbiased estimate of the population mean.

ii) In SRSWOR, the variance of the sample mean is given by

$$var(\overline{y_n}) = \frac{s^2}{n} \cdot \frac{N-n}{N}$$

In simple random sampling, the samples are drawn directly from population units. There may be two cases- SRSWOR and SRSWR. In SRSWOR, the sample unit occurred once is not replaced back in population before making the next draw, and in SRSWR, the sample unit once drawn is again placed on the population unit and then another draw is made. The SRS is useful for homogeneous collection of data.

Whereas stratified random sampling is used when the collected data are found to be heterogeneous. In this method, the population units are divided into different sub-population units called stratum (for each) and samples are drawn from each stratum. The sample units have some homogeneity within stratum but no homogeneity between the strata.

i) Let $x_1, x_2, \dots, x_n$ be the sample observation drawn from the population units $X_1, X_2, \dots, X_N$ of size N.

Then

Sample Mean $(\overline{x}) = \left(\sum_{i=1}^{n} \frac{x_i}{n}\right)$ ... ... (i)

Population Mean $\left(\overline{X}\right) = \frac{(X_1+X_2+\cdots+X_N)}{N} = \sum_{i=1}^{N} \frac{X_i}{N}$

Now taking expectations on both sides of equation (i)

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \rightarrow E(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) \quad \ldots\ldots\ldots (ii)$$

Here, x_i is the i[th] term of sample unit which is drawn from the population unit X_i $(i = 1, 2, \ldots, N)$ with probability of occurance $\frac{1}{N}$.

So, $E(x_i) = X_1\frac{1}{N} + X_2\frac{1}{N} + \cdots + X_N\frac{1}{N} = \sum_{i=1}^{N}\frac{X_i}{N} = \overline{X}\ (or\ \mu)$

Now from equation (ii);

$$E(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}.n\mu = \mu$$

Thus sample mean is an unbiased estimator of population mean.

ii) For SRSWOR, variance of sample mean can be calculated as:

$$V(\overline{x}) = E[\overline{x} - E(\overline{x})]^2 = E\left[\frac{x_1 + x_2 + \cdots + x_n}{n} - \mu\right]^2$$

$$= E\left[\frac{x_1 + x_2 + \cdots + x_n - n\mu}{n}\right]^2 = \frac{1}{n^2}E[x_1 + x_2 + \cdots + x_n - n\mu]^2$$

$$= \frac{1}{n^2}E\left[\sum_{i=1}^{n}(x_i - \mu)\right]^2$$

$$V(\overline{x}) = \frac{1}{n^2}E\left[\sum_{i=1}^{n}(x_i - \mu)^2 + 2\sum_{i<}^{n-1}\sum_{j}^{n}(x_i - \mu)(x_j - \mu)\right] \quad \ldots\ldots\ldots (i)$$

Here $(x_i - \mu)^2$ is the i[th] sample unit drawn from the population unit $(x_i - \mu)^2\ \forall\ i = 1, 2, \ldots, N$ with the probability of occurance $\frac{1}{N}$

$$E(x_i - \mu)^2 = \frac{1}{N}\sum(X_i - \mu) = \sigma^2 \quad \ldots\ldots\ldots (ii)$$

Again $(x_i - \mu)(x_j - \mu)$ is the i[th] sample unit drawn from the population unit $(x_i - \mu)(x_j - \mu)$ with the probability of occurance $\frac{1}{N}.\frac{1}{N-1}$

$$E\left[(x_i-\mu)(x_j-\mu)\right]=\frac{1}{N}.\frac{1}{N-1}\sum_{i<}^{N-1}\sum_{j}^{N}(X_i-\mu)(X_j-\mu)$$

$$=\frac{1}{N}.\frac{1}{N-1}.\frac{(-N\sigma^2)}{2} \quad \ldots\ldots\ldots(iii)$$

Using equation (ii) & (iii) in (i), we get;

$$V(\overline{x})=\frac{1}{n^2}\left[\sum_{i=1}^{n}\sigma^2+2\sum_{i<}^{n-1}\sum_{j}^{n}\frac{1}{N}.\frac{1}{N-1}.\frac{(-N\sigma^2)}{2}\right]$$

$$=\frac{1}{n^2}\left[n\sigma^2-n(n-1).\frac{\sigma^2}{N-1}\right]=\frac{n\sigma^2}{n^2}\left[1-(n-1).\frac{1}{N-1}\right]$$

$$=\frac{\sigma^2}{n}\left[\frac{N-1-n+1}{N-1}\right]=\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

And we know that;

$$\sigma^2=\frac{(N-1)}{N}.s^2$$

So,

$$V(\overline{x})=\frac{1}{n}.\frac{(N-1)}{N}.s^2.\frac{(N-n)}{(N-1)}=\frac{s^2}{n}.\frac{(N-n)}{N}$$

2. Test the hypothesis of no difference between the ages of male and female employees of a certain company using Mann-Whitney U test for the sample data. Use 0.10 level of significance.

Males	**31**	**25**	**38**	**33**	**42**	**40**	**44**	**26**
Females	**44**	**30**	**34**	**47**	**35**	**32**	**35**	**47**

Null Hypothesis H_0: $E(X)=E(Y)$; there is no difference between ages of male and female employees of given company.

Alternative Hypothesis H_1: $E(X) \neq E(Y)$; there is significant difference between ages of male and female employees of given company.

Test statistics: Under H_0, test statistics is obtained as follows:

Combined sample observation	Rank (R)	Rank for 1st sample (R_1)	Rank for 2nd sample (R_2)
31	13	13	
25	16	16	
38	7	7	
33	11	11	
42	5	5	
40	6	6	
44	3.5	3.5	
26	15	15	
44	3.5		3.5
30	14		14
34	10		10
47	1.5		1.5
35	8.5		8.5
32	12		12
35	8.5		8.5
47	1.5		1.5
		R_1=76.5	R_2=59.5

Now

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8 \times 8 + \frac{8(8+1)}{2} - 76.5 = 23.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8 \times 8 + \frac{8(8+1)}{2} - 59.5 = 40.5$$

And

$$U_0 = \min\{U_1, U_2\} = \min\{23.5, 40.5\} = 23.5$$

Critical value: From Mann-Whitney U-test table for 0.10 level of significance and for $n_1 = 8$; $n_2 = 8$, we get

$$U_{\alpha,(n_1,n_2)} = U_{0.10,(8,8)} = 15$$

Decision: $U_0 > U_{0.10,(8,8)}$; Hence, H_0 is accepted. i.e. there is no difference between the ages of male and female employees of a company.

3. Suppose you are given following information with n=28. Multiple regression model $Y = 5 + 18X_1 + 20X_2$. Sample size (n) = 28. Total sum of squares (TSS) = 250. Sum of squares due to error (SSE) = 100. Standard error of regression coefficient of X_1 (Sb_1) = 3.2. Standard error of regression coefficient of X_2 (Sb_2) = 5.5

i) Predict the value of Y for X_1=15 and X_2=25.

ii) Test the significance of regression coefficient of X_2.

iii) Compute the multiple coefficient of determination.

Group-B $(8 \times 5 = 40)$

4. Show that in two stage sampling with SRSWOR at both stages. $\overline{y}$ is an unbiased estimator of $\overline{Y}$.

In case of two stage sampling with SRSWOR

$$\bar{\bar{y}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\overline{y_{ij}} = \frac{1}{n}\sum_{i=1}^{n}\overline{y_{i.}}$$

Taking expectations on both sides, we get;

$$E(\bar{\bar{y}}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\overline{y_{i.}}\right] = E_1\left[E_2\left\{\frac{1}{n}\sum_{i=1}^{n}\overline{y_{i.}}\right\}\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}E(\overline{y_{i.}})\right]$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{N}\overline{Y_{i.}}.\frac{1}{N}\right] = E_1\left[\frac{1}{n}.n\sum_{i=1}^{N}\overline{Y_{i.}}.\frac{1}{N}\right] = \frac{1}{N}\sum_{i=1}^{N}E_1(\overline{Y_{i.}}) = \frac{1}{N}N\overline{Y_{..}}$$

$$= \overline{Y}$$

$$\therefore E(\bar{\bar{y}}) = \overline{Y}$$

Thus, $\overline{y}$ is an unbiased estimator of $\overline{Y}$.

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

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

If we consider three variables x_1, x_2 & x_3, then $r_{12.3}$ denotes the partial correlation between x_1 & x_2 by keeping the effect of x_3 as constant and $r_{12.3}$ is called the partial correlation coefficient between x_1 & x_2 with x_3 as constant.

Similarly,

$r_{23.1}$ is partial correlation coefficient between x_2 & x_3 keeping x_1 constant.

$r_{13.2}$ is partial correlation coefficient between x_1 & x_3 keeping x_2 constant.

The relationship between simple and partial correlation coefficients when there are three variables x_a, x_b & x_c is

$$r_{ab.c} = \frac{r_{ab} - r_{bc}.r_{ac}}{\sqrt{1 - r_{ac}^2}.\sqrt{1 - r_{bc}^2}}$$

Solution:

We have, $r_{12} = 0.5$; $r_{23} = 0.1$; $r_{13} = 0.4$

$$\therefore r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}.\sqrt{1 - r_{23}^2}} = \frac{0.5 - 0.1 \times 0.4}{\sqrt{1 - 0.1^2}\sqrt{1 - 0.4^2}} = 0.504$$

$$\therefore r_{23.1} = \frac{r_{23} - r_{13}.r_{12}}{\sqrt{1 - r_{13}^2}.\sqrt{1 - r_{12}^2}} = \frac{0.1 - 0.5 \times 0.4}{\sqrt{1 - 0.5^2}\sqrt{1 - 0.4^2}} = -0.126$$

6. Define cluster sampling with sample mean and variance of sample mean.

The basic assumption of the survey sampling theory is that complete list of individual units or sample frame should be available. However, if the complete list of the samples is not available then theory may not give the reliable estimate. In such event we can use an alternate approach to solve such problem instead of selecting the individual item from the population to make estimate reliable. The selection of group of the elements of the population is called the cluster sampling.

In case of cluster sampling;

$$\overline{y_n} = \frac{1}{n}\sum_{i}^{n} \overline{y_{i.}}$$

Where, $\overline{y_n}$ = mean of sample cluster for n-cluster; $\overline{y_{i.}}$ = mean value of i^{th} cluster

And variance of sample mean is given by

$$V(\overline{y_n}) = \frac{1 - f}{n} S_b^2$$

Where, $V(\overline{y_n})$ = variance of sample mean; S_b^2 = mean square between clusters; n = number of sample clusters; f = n/N; N = no. of clusters that population units are divided into.

7. For what conditions non parametric test is used? Explain some important non-parametric test.

Non-parametric test are those statistical test which do not depend on any assumptions about the form of the population. i.e. those test whose models does not specify the conditions about the parameter if parent population form which samples has been drawn are known as Non-parametric test.

Basic assumptions of N-P test are:

i) The sample observation are independent.

ii) Variables under study are continuous.

iii) Lower order moments exists.

The importance of N-P test is given below:

i) N-P test are important to study variables which are continuous.

ii) N-P test are designed to test the statistical hypothesis.

iii) N-P test are applied to test the data in ordinal and nominal scale.

iv) N-P test does not make any assumptions about parent population.

8. The weights of 5 people before they stopped smoking, in kilogram, are as follows:

Before	**66**	**80**	**69**	**52**	**75**
After	**71**	**82**	**68**	**56**	**73**

Use the signed-rank test for paired observation to test the hypothesis, at 0.05 level of significance, that giving up smoking has no effect on a person's weight against the alternative that one's weight increases if he or she quits smoking.

Null hypothesis (H_0): Smoking has no effect on person's weight.

Alternative hypothesis (H_1): Smoking has effect on person's weight.

Test statistics: Under H_0, test statistics is obtained as follows:

Rank table:

Before (x_i)	After (y_i)	$d_i = x_i - y_i$	Ranking of $\|d_i\|$	+ sign	- sign
66	71	-5	1		1
80	82	-2	3.5		3.5
69	68	1	5	5	
52	56	-4	2		2
75	73	2	3.5	3.5	
Total				$\sum$=8.5	$\sum$=6.5

$T_{cal} = \min\{S(+), S(-)\} = \min\{8.5, 6.5\} = 6.5$

For two tail test at 0.05 level of significance, T_{tab} is 2.78

<INCOMPLETE>

9. A random sample of 15 adults living in a small town is selected to estimate the proportion of voting favoring a certain candidate for major. Each individual was also asked if he or she was a college graduate. By letting Y and N designate the response of 'yes' and 'no' to the education question, the following sequence was obtained.

N	**N**	**N**	**N**	**Y**	**Y**	**N**	**Y**
Y	**N**	**Y**	**N**	**N**	**N**	**N**	

Use the run test at 0.05 level of significance to determine if the sequence supports the contention that the sample was selected at random.

Null hypothesis (H_0): The sequence supports the contention that the sample was selected at random.

Alternative hypothesis (H_1): The sequence of the sample was selected not at random.

Test statistics: Under H_0, test statistics is; r = number of runs.

Here, the sequence of given result is;

N N N N Y Y N Y Y N Y N N N

Number of runs = 7

$n_1 = 9$ ("no" to the education question)

$n_2 = 5$ ("yes" to the education question)

Critical value: For $n_1 = 9$ & $n_2 = 5$ at 5% level of significance $\underline{r} = 3$ & $\overline{r} = 12$

Decision: Here r = 7

Here, r lies between $\underline{r}$ & $\overline{r}$ therefore we accept accept H_0, i.e, The sequence supports the contention that the sample was selected at random.

10. In an experiment to study the dependence of hypertension on smoking habits, the following data were taken on 180 individuals:

	Non smokers	Moderate smokers	Heavy smokers
Hypertension	**21**	**36**	**36**
No hypertension	**48**	**26**	**19**

Test the hypothesis that the presence or absence of hypertension is independent of smoking habits. Use 0.05 level of significance.

Null hypothesis (H_0): Hypertension is independent of smoking habit

Alternative hypothesis (H_1): Hypertension is not independent of smoking habit

Test statistics: Under H_0, test statistics is

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

Calculation table for χ^2

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
21	$69 \times \frac{87}{180} = 33.350$	-12.350	152.523	4.573
36	$62 \times \frac{87}{180} = 29.967$	6.033	36.397	1.215
30	$49 \times \frac{87}{180} = 23.683$	6.317	39.904	1.685
48	$69 \times \frac{93}{180} = 35.650$	12.350	152.523	4.296
26	$62 \times \frac{93}{180} = 32.033$	-6.033	36.397	1.136
19	$49 \times \frac{93}{180} = 25.317$	-6.317	39.904	1.576
Total				∑=14.481

$$\therefore \chi^2_{cal} = 14.481$$

Degree of freedom = (2-1) × (3-1) = 1×2 = 2

Level of significance (α) = 0.05

Critical value: $\chi^2_{tab} = \chi^2_{0.05,2} = 5.99$

Decision: $\chi^2_{cal} > \chi^2_{tab}$

Hence, we reject the null hypothesis H_0 i.e. hypertension is independent of smoking habit.

11. Define dummy variable. What condition should be necessary for fitting logistic regression?

A dummy variable is one that takes the values 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Since in both linear and exponential models population are found to tend to unreasonable limits of -30 to +30. A model which ultimately tends population size is logistic model. In this model, the restriction is as population size increases the growth rate decreases in proportion to size attended.

Here,

$$r[1 - KP(t)] = \frac{P'(t)}{P(t)} \quad \ldots\ldots\ldots (i)$$

Where, $1 - KP(t)$ is the simplest decreasing form and P(t), r & k are constants.

By using partial fractions technique;

$$\frac{1}{r}\left[\frac{1}{P(t)} + \frac{K}{1 - KP(t)}\right] dP(t) = dt$$

$$\therefore \left[\frac{1}{P(t)} + \frac{K}{1 - KP(t)}\right] dP(t) = r.dt$$

On integrating we get;

$$\log P(t) - \log[1 - KP(t)] = rt + \log c$$

$$or, \log \frac{P(t)}{[1 - KP(t)]} = \log e^{rt} + \log c$$

$$or, \log\left[\frac{P(t)}{1-KP(t)}\right] = \log ce^{rt}$$

$$or, \frac{P(t)}{1-KP(t)} = ce^{rt}$$

$$or, P(t) = [1-KP(t)]ce^{rt}$$

$$\therefore P(t) = \frac{1}{K + \frac{1}{c}e^{-rt}}$$

This is the logistic model which shows that:

$$t \to -\infty \ ; \ P(t) = 0$$

$$t \to \infty \ ; \ P(t) = \frac{1}{K}$$

12. Suppose the residuals for a set of data collected over 9 consecutive time periods are as follows:

Time Period	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Residuals	**-2**	**-3**	**2**	**-1**	**0**	**1**	**4**	**-2**	**1**

Compute the Durbin Watson statistics. At 0.05 level of significance, is there evidence of autocorrelation among the residuals?

	e_t	e_{t-1}	e_{t-2}	$e_t . e_{t-1}$	$e_t . e_{t-2}$	$e_t^2 =$	et-et-1	$(e_t - e_{t-1=})^2$
1	-2					4		
2	-3	-2		6		9	-1	1
3	2	-3	-2	-6	-4	4	5	25
4	-1	2	-3	-2	3	1	-3	9
5	0	-1	2	0	0	0	1	1
6	1	0	-1	0	-1	1	1	1
7	4	1	0	4	0	16	3	9
8	-2	4	1	-8	-2	4	-6	36

9	1	-2	4	-2	4	1	3	9
Total				$\sum e_t . e_{t-1}$= -8	$\sum e_t . e_{t-2} =$ 0	e_t^2 =3 9		$\sum(e_t - e_{t-1})^2 =$ 91

Test of significance of autocorrelation:

Null Hypothesis (H_0)$\rho = o$ there is no positive auto correlation in the residuals.

Alternative Hypothesis(H_1) ρ>0 there is positive auto correlation in the residuals.

Test statistics, under H_0, the Durbin Watson statistic is

$$d = \frac{\sum_{t=1}^{9}(e_t - e_{t-1})^2}{\sum_{t=1}^{9} e_t^2} = \frac{91}{39} = 2.33$$

From Durbin Watson table for n=9 and k=1 , at 5% level significance d_L=0.84,d_u=1.320,

Decision:since d>d_U,there is no evidience of positive auto correlation in the residuals.

13. Describe multiple regression models with its assumption. Also describe the method of obtaining its parameters.

Multiple regressions are the study of relationship among three or more variables. It is used to estimate the value of dependent variable with the help of known values of two or more independent variables by the use of fitted multiple regression on equation.

Let us consider three variables X_1, X_2 & X_3. Then, multiple regression equation of X_1 on X_2 & X_3 is given as;

$$X_1 = a + b_2X_2 + b_3X_3 \quad \ldots\ldots\ldots (*)$$

Where, X_1 = dependent variable ; a = value of X_1 when $X_2 = X_3 = 0$; b_2 = regression coefficient of X_1 & X_2 ; b_3 = regression coefficient of X_1 & X_3

We have;

$b_2 = \frac{dX_1}{dX_2}$; i.e. rate of change in X_1 w.r.t. unit change in X_2

$b_3 = \frac{dX_1}{dX_3}$; i.e. rate of change in X_1 w.r.t. unit change in X_3

Here, we have to estimate a, b_2 & b_3.

To estimate a, b_2 & b_3 ; the normal equation are:

$$\Sigma X_1 = na + b_2\Sigma X_2 + b_3\Sigma X_3 \quad \ldots\ldots\ldots (i)$$

$$\Sigma X_1X_2 = a\Sigma X_2 + b_2\Sigma X_2^2 + b_3\Sigma X_2X_3 \quad \ldots\ldots\ldots (ii)$$

$$\Sigma X_1X_3 = a\Sigma X_3 + b_2\Sigma X_2X_3 + b_3\Sigma X_3^3 \quad \ldots\ldots\ldots (iii)$$

Then our fitted multiple regression equation of X_1 on X_2 & X_3 is

$$X_1 = \hat{a} + \widehat{b_2}X_2 + \widehat{b_3}X_3$$

Similarly, our fitted multiple regression equation of X_2 on X_1 & X_3 is

$$X_2 = \hat{a} + \widehat{b_1}X_1 + \widehat{b_3}X_3$$

Similarly, our fitted multiple regression equation of X_3 on X_1 & X_2 is

$$X_3 = \hat{a} + \widehat{b_1}X_1 + \widehat{b_2}X_2$$