Tribhuvan University

Institute of Science and Technology

2069

☆

Solution

Bachelor level/First Year/First Semester/Science	Full Marks: 60
Computer Science and Information Technology Stat.108	Pass Marks: 24
(Statistics)	

Candidates are required to give their answers in their own words as far as practicable.

All notations have the usual meanings.

Group 'A'

Attempt any Two: **(2x10=20)**

1. Describe simple random sampling with and without replacement for drawing a random sample of size n from a population of size N. In both cases show that sample mean is unbiased estimate of the population mean. Derive the variance of the sample mean in both cases. If V_{srswr} and V_{srswor} corresponding denote that variance of sample mean under simple random sampling with and without replacement method, then show that

$$(V_{srswr} - V_{srswor}) = \frac{n-1}{Nn} S^2$$

and write conclusion that you can draw out of the above result.

2. Describe function and method of sign test. A study was designed to determine the effect of a certain movie on the moral attitude of young children. The data below represent a rating from 0 to 20 on a moral attitude scale recorded before and after viewing the movie, where high score associated to high morality. Carry out the test hypothesis that movie had no effect on moral attitude of children against it had using sign test at level 0.1.

Before	14	16	15	18	15	17	19	17	17	16	14	15
After	13	18	16	17	16	19	20	18	19	15	18	16

A sign test is a very simple and easiest non-parametric test which applies to the median. The name sign test comes from the fact that it is based on two signs: plus (+) and minus (-) signs of

i. difference between observations in one sample and the specified value of median and
ii. difference between observations within pairs of two independent samples. The sign test is therefore applied for testing hypothesis concerning
   a) the specified value of median for one population
   b) the median of difference between paired data of two dependent samples

Here, in sign test, there are two types of sample sign test:

i. one sample sign test
ii. two sample sign test

In both sample sign test there are two cases, i.e.,

a) small sample case ($n \leq 25$)
b) large sample case ($n > 25$)

Solution

Null hypothesis (H_0): $M_{d_1} = M_{d_2}$, i.e., movie has no effect on moral attitude of children.

Alternative hypothesis (H_1): $M_{d_1} \neq M_{d_2}$, i.e., movie has effect on moral attitude of children.
Test statistic:

Under H_0, the test statistic is obtained as follows:

Before	14	16	15	18	15	17	19	17	17	16	14	15
After	13	18	16	17	16	19	20	18	19	15	18	16
	+	-	-	+	-	-	-	-	-	+	-	-

Number of + signs (+) = 3
Number of – signs (-) = 9
Number of effective sample n_e = n(+) + n(-)
= 3+9
=12

Test statistic (k) = $\min\{n(+), n(-)\} = \min\{3,9\} = 3$
Critical value:

For 0.1 level of significance,

P-value (P_0) = P($y \leq 3$)=0.073

Decision: since P_0=0.073<0.1=α, we reject H_0 and conclude that movie has effect on moral attitude of children.

3. A sample of n=22 data points was used to estimate β_0+β_1 and β_2 of the multiple regression model:$Y=\beta_0+\beta_1X_1+\beta_2X_2+u$, where Y=sales of the product in thousand Rs, X1=radio advertising expenditure in thousand Rs and X2=newspaper advertising expenditure in thousand Rs. The available estimates of parameters and their standard errors are summarized below:

	β_0	β_1	β_2
Estimate	156.43	13.08	16.08
Standard Error	126.76	1.76	2.96

a) Write the estimated regression model and interpret the coefficient of X1, X2.
b) Predict the value of Y when X1=75 and X2=50.
c) Further computation shows that $\sum(Y_i - \bar{Y})^2$ =2507793 and $\sum(Y_i-\hat{Y}_i)$=49760. Based on this result compute R^2 and carry out overall significance test of the model.
d) Carry out the test of $H_{0:}$ β_1=0 against H_1 β_1.

Solution

Hint:

a) $Y=156.43+13.08X_1+16.08X_2$
b) $Y=156.43+13.08\times75+16.08\times50$
c) We assumed

H_0: $\beta_1 = \beta_2 = 0$
H_1: $\beta_1 \neq \beta_2 \neq 0$
Test statistic:

TSS=SSR+SSE
$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i) + \sum(\hat{Y} - Y_i)$
$2507793 = 49760 + \sum(\hat{Y} - Y_i)$
$\sum(\hat{Y} - Y_i) = 2458033$

SV	SS	df	MSS	F-ratio
SSR	49760	2	24880	0.192
SSE	2458033	19	129370.16	
TSS	2507793			

Critical value:

Calculated $F_{(k-1,n-k)}$df, 0.05≈$F_{(2,19)df,0.05}$=3.52
Decision: $F_{cal}<F_{tab}$ so we reject H_0
$R^2=\frac{TSS}{SSR}$=50.398

d) Solution

We assumed,
H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$
Test statistic:

SV	SS	df	MS	F-ratio
SSR	13.08	2	6.54	72.67
SSE	1.76	19	0.09	

Critical value:
$F_{(2,19)0.05}$=3.52
Decision:
$F_{cal}>F_{tab}$ so we reject H_0.

Group 'B'

Answer any eight questions: **(8x5=40)**

4. In a stratified random sampling, if the cost of survey is constant for each stratum then derive an expression for n_h under optimum allocation.

Do yourself

5. Describe in detail two-stage sampling method. Obtain an expression for an unbiased estimator of population total when sample were drawn by adopting simple random sampling without replacement method, and what would be the expression of this unbiased estimator if M_i=M and m_i=m for all i?

Solution

Let, E_2 and V_2 denote the expectation and variance of a statistics w.r.t. all possible selections of second stage sample and E_1 and V_1 represent the expectation and variance w.r.t. all possible selections of first stage sample.

Let, t be the estimator of the parameter T then, on case of two stage sampling, expected value of t is

$$E(t) = E_1E_2(t) = T$$
$$V(t) = E_1E_2(t-T)^2$$

Consider, $E_2(t_2 - T)^2 = E_2(t^2 - 2tT + T^2) = E_2(t^2) - 2TE_2(t) + T^2$
$= V_2(t) + [E_2(t)]^2 - 2TE_2(t) + T^2$

Now,

$$\begin{aligned} E_1E_2(t-T)^2 &= E_1[V_2(t) + [E_2(t)]^2 - 2TE_2(t) + T^2] \\ &= E_1V_2(t) + E_1[E_2(t)]^2 - 2TE_1E_2(t) + T^2 \\ &= E_1V_2(t) + E_1[E_2(t)]^2 - 2T.T + T^2 \\ &= E_1V_2(t) + E_1[E_2(t)]^2 - T^2 \\ &= E_1V_2(t) + E_1[E_2(t)]^2 - [E_1E_2(t)]^2 \end{aligned}$$

$$= E_1V_2(t) + V_1E_2(t)$$

$$\therefore V(t) = E_1E_2(t-T)^2 = E_1V_2(t) + V_1E_2(t)$$

From N p.s.u. from the population, a simple random sample without replacement of n p.s.u.'s is selected and from the ith selected p.s.u.in S.R.S. without replacement of m_i s.s.u.'s out of M_i s.s.u.'s are selected.

Mean of the sample of ith selected p.s.u.= $\sum_{j=1}^{m_i} \frac{y_{ij}}{m_i} = \bar{y}_i$

Ith selected p.s.u. total $\hat{Y}_i = M_i\bar{y}_i$

Mean of the selected pairs = $\frac{M_1\bar{y}_1+M_2\bar{y}_2+\cdots+M_n\bar{y}_n}{n} = \sum_{i=1}^{n} \frac{M_i\bar{y}_i}{n}$

$$\hat{Y} = N\sum_{i=1}^{n} \frac{M_i\bar{y}_i}{n} = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij}$$

Now, we have to show that,

$$E(\hat{Y}) = Y_i$$

Proof:

$$E(\hat{Y}) = E_1E_2(\hat{Y}) = E_1E_2\left(N\sum_{i=1}^{n} \frac{M_i\bar{y}_i}{n} = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij}\right)$$

Consider,

$$E_2\left(N\sum_{i=1}^{n} \frac{M_i\bar{y}_i}{n} = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij}\right) = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} E_2(y_{ij})$$

where, y_{ij} is expected to have come either from Y_{11} or Y_{12} or Y_{1Mi}, each with same probability $1/M_i$.

$$\therefore E_2(y_{ij}) = \frac{y_{i1}+y_{i2}+\cdots+y_{iM_i}}{M_i} = \sum_{j=1}^{M_i} \frac{Y_{ij}}{M_i} = \bar{Y}_i$$

$$\therefore \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} E_2(y_{ij}) = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} \bar{Y}_i$$

$$= \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i} \times m_i\bar{Y}_i$$

$$= \frac{N}{n}\sum_{i=1}^{n} M_i\,\bar{Y}_i$$

$$= \frac{N}{n}\sum_{i=1}^{n} Y_i$$

Now, taking E_1 on both sides,

$$E_1\frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} E_2(y_{ij}) = E_1\frac{N}{n}\sum_{i=1}^{n} Y_i$$

$$= \frac{N}{n}\sum_{i=1}^{n} E_1\,Y_i$$

$$\therefore E(\hat{Y}) = \frac{N}{n}\sum_{i=1}^{n}\sum_{i=1}^{N} \frac{Y_i}{N} = \frac{1}{n}\sum_{i=1}^{n} Y = \frac{1}{n} \times nY = Y$$

6. Define a run. A true-false examination constructed with the answers running in the following sequence

TFFTFTFTTFTFFTFTFTTF

Does this sequence indicate a departure from randomness in the arrangement of T and F answers? Critical region at 5% level of significance is R$\leq$6 or R$\leq$16 where R=number of runs.

Soln: Here, we assume

H_0 : The samples are in random order.

H_1: The samples are nit in random order.

Test statistic:

TFFTFTFTTFTFFTFTFTTF

No. of T = n_1 = 10

No. of F = n_2 = 10

Then the test statistic for the run test is the number of R.

Critical region:

For pre-assigned level of significance α=0.05 the critical region for the run test is the region below $\underline{r}$ or above $\overline{r}$ where $\underline{r}$ and $\overline{r}$ are the lower and upper critical values of which can be obtained from the number of runs of n_1 and n_2. From the table of run test,

n1 = 10, $\underline{r}$ = 6

n2 = 10 , $\overline{r}$ = 16

Decision: The pre-assigned values α=0.05 does not lies between $\underline{r}$ = 6 and $\overline{r}$ = 16. So, we reject H_0.

7. A survey of voter sentiment was conducted in four wards to compare the fraction of voters candidate A, Random samples of two hundred voters were polled in each of the four ward with result as shown below

	Ward			
	1	2	3	4
Favor A	76	53	59	48
Do not Favor A	124	147	141	152

Do the data present sufficient evidence to indicate that the fraction of voters favoring candidate A differ in the four wards? Use chi=square test at 5% level.

Solution :

Given,

n = 800 & layout are the no. of people favouring A & B.

Probem:

To test,

H_0: The fraction of voters favouring candidate A & B are independent.

H_1: The fraction of voters favouring candidate A & B are dependent.

Calculations:

	Ward				
	1	2	3	4	Total
Favor A	76	53	59	48	236
Do not Favor A	124	147	141	152	564
Total	200	200	200	200	800

$e_{11} = \frac{236*200}{800} = 59$

$e_{12} = \frac{236*200}{800} = 59$

Therefore same e_{13} e_{14} = 59

$e_{21} = \frac{564*200}{800} = 141$

Therefore e_{22} e_{23} e_{24} = 141

<u>Test statistic:</u>

We have,

$\chi^2_{cal} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(oij-eij)^2}{eij}$

$= \frac{(76-59)^2}{59} + \frac{(53-59)^2}{59} + \frac{(59-59)^2}{59} + \frac{(48+59)^2}{59} + \frac{(124-141)^2}{141} + \frac{(147-141)^2}{141} + \frac{(141-141)^2}{141} + \frac{(152-141^2}{141}$

= 10.6

<u>Critica value:</u>

For α = 0.05 & d.f = (r – 1) (c - 1) = (2 – 1) (4 - 1)

= 3

$\chi^2_{\alpha, d.f} = \chi^2_{0.05,3}$ = 7.82

<u>Decision:</u>

Since χ^2_{cal} = 10.69 > $\chi^2_{0.05,d.f}$ = 7.82 , we reject h_0 .

8. The final exam scores obtained by two groups of students, where students of group A were taught using method A and those of group B were taught using method B, are summarized below. Use Mann Whitney test to determine whether or not the final exam scores of two groups are different.

Group A	55	59	61	64	64	70	73	75	76	82	83	95
Group B	65	77	80	80	84	86	88	91	91	91		

Solution

Given,

n1=12, n2=10 & layouts are the final exam scores obtained by two group of students.

Problem:

To test

H_0:$\mu_x = \mu_y$: Difference of score between two groups is not significant.

Vs $H_{1:}$ $\mu_x = \mu_y$: Difference of score between two gropus is significant different.

Calculations:

Final Exam score			
Group A	Rank	Group B	Rank
55	1	65	6
59	2	77	11
61	3	80	12.5
64	4.5	80	12.5
64	4.5	84	16
70	7	86	17
73	8	88	18
75	9	91	20
76	10	91	20
82	14	91	20
83	15	total	$Rank_2$ = 153

95	22	
Total	$R_1 = 100$	

Then ,

R_1 = sum of ranks of first sample = 100

R_2 = sum of ranks of second sample = 153

$U_1 = n_1n_2 + n_2(n_2+1) / 2 – R_1$

$= 12 * 10 + 12(12+1) / 2 -100$

$= 120 + 156 / 2 – 100$

$= 98$

$U_2 = n_1n_2 + n_2(n_2+1) / 2 – R_2$

$= 12 * 10 + 10(10+1) / 2 -153$

$= 120 + 110 / 2 – 153$

$= 22$

Test statistic:

Therefore, the statistic U_0 = min of { $U_1 = 98$ or $U_2 = 22$ }

$= 22$

Critical region:

Next for a pre – assigned level of significance $\alpha = 0.05$. We obtain from the Mann-Whitney the critical value $U_\alpha = U_{0.05} = 29$

Decision:

Since $U_0 = 22 < 29 = U_\alpha = _{0.05}$, we reject H_o

⇨ Difference of score between two groups is significantly different.

9. Describe rationale and method of Median test.

Median test is a non-parametric test which is frequently used to test the difference in medians (locations) of two independent distributions. So the median test is used to test whether two independent random samples have been drawn from two populations with same median (or not) or to test the equality of two medians of independent population distributions. Also, the median test is a non-parametric test used to test whether the two treatments applied in an experiment are equally effective or not.

Null hypothesis (H_0): $F_x(x) = F_y(x)$
Alternative hypothesis (H_1): $F_x(x) \neq F_y(x)$

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

i. Combine all the observations of both samples and arrange them in ascending order so that $n = n_1 + n_2$.

ii. Compute the sample median M_d of the pooled data.

iii. Find the number of observation in the first sample less than or equal to M_d, i.e., find the number of x's$\leq M_d$ and denote it by 'a'. This 'a' is the test statistic for testing the null hypothesis, H_0. In this small sample median test, when the null hypothesis H_0 is true, the sampling distribution of random variable A, the number of x's$\leq M_d$, is the hyper geometric distribution with probability mass function (p.m.f.)

$$p(A = a) = \begin{cases} \frac{\binom{n_1}{a}\binom{n_2}{k-a}}{\binom{n_1+n_2}{k}}, \; a = 0,1,2,3,\dots,\min(n_1,k) \\ \qquad 0, \; otherwise \end{cases}$$

where, $k = \frac{n_1+n_2}{2} = \frac{n}{2}$

Critical region:

The critical region is,

$$P_0 = P(A \geq a)$$

Decision:

If $2P_0 \leq \alpha \Rightarrow P_0 \leq \frac{\alpha}{2}$, we reject null hypothesis.

10. Describe the method of estimation of the parameters α and β of the growth model $Y_t = \alpha e^{\beta t}$, where the values of Y_t are available for t=1,2,.....,n? What does the estimated β measure?

Do yourself

11. Describe the method of formulating a multiple regression model when the dependent variable Y is binary in nature.

Soln: A multiple regression equation of dependent variable X on two independent variables Y_1 and Y_2 is an equation for estimating a dependent variable X from two independent variables X_1 and X_2.

The multiple regression equation of dependent variable Y on two independent variables X_1 and X_2 are given by,

$$Y = a + b_1 X_1 + b_2 X_2$$

12. Write a suitable multiple regression model if the dependent variable(Y) is output and one independent variable(X_1) is labor input and other independent variable (X_2) is capital input. Describe the properties of the model.

Soln: It is logical extension of the simple linear regression analysis in multiple regression analysis instead of a single independent variable, two or more independent variables are used to estimate the unknown values of a dependent variable.

$$Y= a + b_1X_1 + b_2X_2$$

Where,

Y = dependent variable
a = constant

b_1 = coefficient of X_1

b_2 = coefficient of X_2

X_1 = independent input variable

X_2 = independent input variable

Properties of the model are:

a) To establish a regression equation, which provides estimates of the dependent variable from the values of two or more independent variables.

b) To obtain measures of error involves in using this regression as a basis for estimation of the dependent variables (i.e. to examine the multiple regression standard error of estimate.).

c) To measure the coefficient of multiple determination or the proportion of variation in the dependent variable which is explained by the independent variable.

13. Define multiple correlation coefficient. If r_{12}=0.679, r_{13}=0.502 and r_{23}= -0.092, then compute $R_{1.23}$, where r_{ij} is the simple correlation coefficient between $X_{1,}$ and X_2 and X3.

The study of the relationship among three or more variables at a time is called multiple correlation. In multiple correlation, all the given variables are studied at one time by taking one variable as dependent and all the remaining variables as independent.

The multiple correlation coefficient between the three variables x_1, x_2, x_3 when x_1 is dependent and x_2 and x_3 are independent variable is denoted by $R_{1.23}$ which is calculated by following formula:

$$R_{1.23} = \sqrt{\frac{{r_{12}}^2+{r_{13}}^2-2r_{12}.r_{13}.r_{23}}{1-{r_{23}}^2}} \quad \ldots\text{(i)}$$

Similarly,

$$R_{2.13} = \sqrt{\frac{{r_{21}}^2+{r_{23}}^2-2r_{12}.r_{13}.r_{23}}{1-{r_{13}}^2}}$$

and,

$$R_{3.12} = \sqrt{\frac{{r_{31}}^2+{r_{32}}^2-2r_{12}.r_{13}.r_{23}}{1-{r_{12}}^2}}$$

We have,

$r_{12} = 0.679$

$r_{13} = 0.502$

$r_{23} = -0.092$

Substituting these values in (i), we get,

$R_{1.23} = 1.85$

14. What is the auto correlation? How do you estimate and test it?

One of the important assumptions that we made in simple or multiple regression model id that there is no correlation between the error terms, i.e. $cov(e_i, e_j) = 0$ for $i \neq j$. This assumption is often violated in time series model, i.e. $cov(e_i, e_j) \neq 0$ for $i \neq j$. Error terms for time periods not too far apart may be correlated due to several factors. This property is known as serial or autocorrelation.

We estimate it by Durbin-Watson test. The steps for D.W test are as follows:

- Set up H_0 and H_1:
  H_0: $\varrho = 0$ i.e. there is no auto correlation between error terms.
  H_1: $\varrho > 0$ (right tailed test)

H_1: $\varrho < 0$ (left tailed test)

- Compute the test statistic

  Under H_0, the D.W test statistic d is

  $$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t{}^2}$$

- For the given sample size (n) and given number of explanatory variables find out d_L and d_u values at α level of significance.
- Make decision comparing d with d_L and d_u as described in D.W test.