CSITNEPAL

Statistics

108(second batch)

Cordial thanks:Mr. Sandip Poudel and Mr. Anil Tiwari

2012

2067

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

1. Describe in detail stratified random sampling method for drawing a random sample of size n from a population of size N. Write down the expression for an unbiased estimator $\overline{y_{st}}$ of population mean $\overline{Y}$ and derive an expression for $Var(\overline{y_{st}})$ when samples were drawn from each stratum by adopting simple random sampling without replacement method. Also find the $Var(\overline{y_{st}})$ under the scheme of proportional allocation of sample sizes to strata.

If the population units are not homogeneous, then Simple Random Sampling (SRS) is not used to draw samples. For each cases, the best method to draw the samples is Stratified Sampling. This method is used for population if it is heterogeneous.

The following steps are followed in this method:

i) Divide the population into number of sub-population called strata.

ii) Population is divided into stratum in such a way that there is homogeneity within stata & heterogeneity between strata.

iii) Each unit of population belongs to only one solution.

iv) Sample mean should differ as possible as it can.

Stratified sampling is the process of dividing the population into number of sub-population called strata and from each strata a sample is drawn independently.

For stratified sampling;

$$\overline{y_{st}} = \sum_{h=1}^{L} \omega_h \overline{y_h}$$

Where, $\overline{y_{st}}$ = unbiased estimator of population mean

$\omega_h = \frac{N_h}{N}$

$\overline{y_h}$ = mean of the sample of h[th] strata

For the expression of $Var(\overline{y_{st}})$ adopting SRSWOR, we know that;

$$\overline{y_{st}} = \sum_{h=1}^{L} \omega_h \overline{y_h}$$

Taking variance on both sides,

$$V(\overline{y_{st}}) = V\left[\sum_{h=1}^{L} \omega_h \overline{y_h}\right] = \sum_{h=1}^{L} \omega_h^2 \, V(\overline{y_h}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2} \cdot \frac{N_h - n_h}{N_h} \cdot \frac{S_h^2}{n_h} \quad \ldots\ldots\ldots (i)$$

$$\boldsymbol{V(\overline{y_{st}}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h(N_h - n_h)S_h^2}{n_h}}$$

Again from equation (i);

$$V(\overline{y_{st}}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$$

$$\boldsymbol{V(\overline{y_{st}}) = \sum_{h=1}^{L} \omega_h^2 (1 - f)\frac{S_h^2}{n_h}}$$

Again from equation (i);

$$V(\overline{y_{st}}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(\frac{N_h - n_h}{N_h . n_h}\right) S_h^2 = \sum_{h=1}^{L} \omega_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

$$\sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{N_h} = \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{N_h^2}{N^2} \cdot \frac{S_h^2}{N_h} = \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{\omega_h S_h^2}{N}$$

$$\boldsymbol{V(\overline{y_{st}}) = \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2} \quad \ldots\ldots\ldots (ii)$$

For proportional allocation method;

$$n_h = \frac{n}{N} N_h$$

Now, from equation (ii);

$$V(\overline{y_{st}}) = \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2$$

$$V(\overline{y_{st}})_{prop} = \sum_{h=1}^{L} \frac{\omega_h^2 S_h^2}{\frac{n}{N} N_h} - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2$$

$$V(\overline{y_{st}})_{prop} = \frac{n}{N} \sum_{h=1}^{L} \frac{N_h^2}{N^2} . \frac{S_h^2}{N_h} - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2$$

$$V(\overline{y_{st}})_{prop} = \frac{1}{n} \sum_{h=1}^{L} \frac{N_h^2}{N} . S_h^2 - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2$$

$$V(\overline{y_{st}})_{prop} = \frac{1}{n} \sum_{h=1}^{L} \omega_h S_h^2 - \frac{1}{N} \sum_{h=1}^{L} \omega_h S_h^2$$

$$\boldsymbol{V(\overline{y_{st}})_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \omega_h S_h^2}$$

Further continuing;

$$\boldsymbol{V(\overline{y_{st}})_{prop} = \left(1 - \frac{n}{N}\right) . \frac{1}{n} \sum_{h=1}^{L} \omega_h S_h^2 = (1 - f) . \frac{1}{n} \sum_{h=1}^{L} \omega_h S_h^2}$$

2. Write down the rationale and method of Wilcoxon matched-pairs signed rank test. Seven prospective graduate students took a test twice with the following scores.

First Attempt	470	530	610	440	600	590	580
Second	510	550	600	490	585	620	598

Attempt							

Compute the value of T^+ where T^+ is the sum of ranks of the positive differences (second attempt – first attempt). Using T^+ as test statistic carry out the test of the following hypothesis at level 0.05

H_0: There is no statistical difference between the first and second attempt score.

H_1: Second attempt score tends to be larger than the first attempt score.

This test is applied to test the hypothesis concerning the difference between two treatment used in the two random samples. It can also be applied to test the hypothesis concerning the effectiveness.

The proves involved in Wilcoxon matched pairs signed rank test is:

Test Null & Alternative hypothesis

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

a) Find the difference: $d_i = x_i - y_i \;;\; i = 1,2,\dots,n$
b) Rank the magnitude of difference d_i either from largest to smallest or vice versa.
c) Attach the signs of $d_i's$ to the respective ranks.
d) Obtain the sum of rank with (+) sign & (-) sign denoted by S(+) & S(-).
e) Then test statistics under H_0 is: $T_{cal} = \min\{S(+), S(-)\}$

Critical value: Tabulated value of Wilcoxon test is obtained from the table of Wilcoxon matched pairs signed rank test.

Decision: If $T_{cal} < T_{tab}$, reject H_0 ; otherwise, accept H_0

As per H_0 and H_1 set in question

Test statistics:

Rank Table:

First attempt	Second attempt	$d_i = x_i - y_i$	Ranking of $\lvert d_i \rvert$	+ sign	- sign

(x_i)	(y_i)				
470	510	40	2	2	
530	550	20	4	4	
610	600	-10	7		7
440	490	50	1	1	
600	585	-15	6		6
590	620	30	3	3	
580	598	18	5	5	
Total				∑=15	∑=14

T_{cal} = min{S(+), S(-)} = min(15,14) = 14

Critical value: For two tail test at 0.05 level of significance, T_{tab} = 21

Decision: $T_{cal} < T_{tab}$, hence we reject H_0.

i.e. H_1 is accepted or second attempt score tends to be larger than the first attempt score.

3. To study the effect of age (X_1 in years) and weight (X_2 in lbs) on systolic blood pressure (Y in mm Hg), the data were recorded for a sample of 15 adult males. The estimated regression model based on data is described below in the box where figures within parenthesis are standard error of the estimate. Further computation shows that:

$$\sum (Y_i - \overline{Y})^2 = 1835.7 \;\&\; \sum (Y_i - \hat{Y}_i)^2 = 1101.3$$

$\hat{Y}_i$=27.4 0.22X_1+ 0.56X_2

(24.68) (0.248) (0.155)

a. Explain the meanings of the estimated slope regression coefficients of the model.

b. What value of Y would you predict if X_1=55 and X_2=175?

c. Compute the value of R^2 and interpret it.

d. Carry out the overall goodness-of-fit test of the model at 5% level of significance.

e. Test the significance of slope regression coefficients at 5% level of significance.

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

4. Describe in detail systematic sampling method when $N = k \times n$. Describe problems that will arise in systematic sampling method when $N \neq k \times n$.

When the complete list of the sampling units are available in systematic order then systematic sampling is applied.

Let we have the list of N sampling units order from 1 to N and we need the sample size n. Then we divide the total population units in n group such that $k = \frac{N}{n}$ i.e. $N = k.n$. Grouping of $N \times n$ array is given as below:

1	2	3	…	r	…	k
k+1	k+2	k+3	…	k+r	…	2k
2k+1	2k+2	2k+3	…	2k+r	…	3k
…	…	…		…		…
(n-1)k+1	(n-1)k+2	(n-1)k+3	…	(n-1)k+r	…	nk

Here, we select a unit randomly from the first row and remaining sample unites are systematically selected suppose unit 'r' is selected randomly from the 1st row then k+r, 2k+r, …, (n-1)k+r units are selected systematically.

For example: N=20, n=5 then k=20/5=4

Here, possible grouping of above sample units is:

1	2	3	4
5	6	6	7
9	10	11	12
13	14	15	16
17	18	19	20

If unit 2 is selected randomly from 1st row then 6, 10, 14 & 18 units are also selected systematically.

But if N=17, n=5 then k=17/5=3.4≈4

Grouping of this sample units is:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17			

Here if we select unit 1 from first row, there would be 5 samples and rest of other units from 1st row gives only 4 sample units. The sample units differ as per the unit of first row. This is the problem that arise in systematic sampling method when $N \neq k \times n$.

5. If V_{srswr} and V_{srswor} correspondingly denote that variance of unbiased estimator of the population mean under simple random sampling with and without replacement method, then show that $(V_{srswr} - V_{srswor}) = \frac{n-1}{Nn} S^2$

We know that,

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

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

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

For SRSWR;

$$V(\overline{x}) = \frac{1}{n^2}\, E\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] + 0$$

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

$$E(x_i-\mu)^2 = \frac{1}{N}\sum(x_i-\mu)^2 = \frac{1}{N}\sum(x_i-\overline{X}) = \sigma^2$$

$$V(\overline{x})_{SRSWR} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$V(\overline{x})_{SRSWR} = \frac{N-1}{N}.\frac{s^2}{n} \quad \ldots\ldots\ldots (ii)$$

For SRSWOR;

From equation (i);

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

$(x_i-\mu)^2$ is the ith sample unit expected to occur from population unit $(x_i-\mu 2\ \forall\ i=1, 2, \ldots, N$ with probability $1N$ to occur.

Then

$$E(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i-\mu) = \sigma^2$$

And $(x_i - \mu)(x_j - \mu)$ is the sample unit expected to occur from the population unit $(x_i - \mu)(x_j - \mu)$ $\forall\, i = 1, 2, \ldots, N-1\,; j = 1, 2, \ldots, N$ with probability of occurance $\frac{1}{N}.\frac{1}{N-1}$

Then,

$$E\left[(x_i - \mu)(x_j - \mu)\right] = \frac{1}{N}.\frac{1}{N-1}\sum_{i<}^{N-1}\sum_{j}^{N}(x_i - \mu)(x_j - \mu)$$
$$= \frac{1}{N}.\frac{1}{N-1}\left(-\frac{N\sigma^2}{2}\right)$$

Therefore,

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

Now,

$$V(\overline{x})_{SRSWR} - V(\overline{x})_{SRSWOR} = \left(\frac{N-1}{N}\right).\frac{s^2}{n} - \left(1 - \frac{n}{N}\right).\frac{s^2}{n}$$
$$= \left(\frac{N-1}{Nn} - \frac{N-n}{Nn}\right)s^2 = \left(\frac{n-1}{Nn}\right)s^2$$

We can conclude that, when same population units are projected under SRSWR & SRSWOR then the variance of unbiased estimator of the population mean in SRSWR is always greater than that of SRSWOR.

6. Compute the problem of determining if a die is fair or not. For this a die is rolled for 60 times and observed the following outcomes.

Side	1	2	3	4	5	6	Total
No. of times observed	8	9	13	7	15	8	60

Test the hypothesis H_0: the die is fair, i.e. all sides have 1/6 chances of appearing against H_1: the die is unfair at level 0.05.

7. Describe the method of Mann Whitney Test.

Mann-Whitney U-test is most powerful non-parametric test for testing the hypothesis of difference between two independent location of two independent random samples. This is non-parametric alternative of t-test. The testing procedure is:

Set Null & Alternative hypothesis

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

a) Combine the sample observation of both samples so that $n = n_1 + n_2$
b) Rank these n observation either from smallest to largest or vice-versa.
c) Determine the sum of rank assigned to the values of the 1st and 2nd sample separately and denote by R_1 & R_2
d) Combine U-values defined in terms of R_1 & R_2 for each samples as:

$$U_1 = n_1 n_2 + n_1(n_1 + 1)/2 - R$$
$$U_2 = n_1 n_2 + n_2(n_2 + 1)/2 - R$$

e) Calculate test statistic as:

$$U_0 = \min\{U_0, U_1\}$$

Critical Value: Critical value of Mann-Whitney U-test is obtained from Mann-Whitney U-test table for the given n_1, n_2 and specified level of significance.

Decision: If $U_o \leq U_{\alpha,(n_1,n_2)}$ then reject H_0 ; otherwise accept H_0.

8. Suppose that an IQ test is given to eleven randomly selected pairs consisting of one brother and one sister from the same family. To test the null hypothesis that this sample was drawn from a population in which the median IQ of a brother and sister do not differ against the alternative hypothesis that the sister would score higher than brother IQ.

Sister's Score	**129**	**111**	**117**	**120**	**116**	**101**	**107**	**127**	**105**	**123**	**113**
Brother's Score	**115**	**108**	**123**	**104**	**110**	**98**	**106**	**119**	**95**	**130**	**101**

Using sign test carry out the above said hypothesis at 5% level of significance.

H_0: Median IQ of brother do not differ from sister; $M_b=M_s$

H_1: Median IQ of sister is greater than that of brother; $M_b<M_s$

Test Statistics: The table may be constructed after finding out the median of the given data.

Arranging the data in ascending order, we get:

95, 98, 101, 101, 104, 105, 106, 107, 108, 110, 111, 113, 115, 116, 117, 119, 120, 123, 123, 127, 129, 130

Median(Md) = Value of $\left(\frac{N+1}{2}\right)^{th}$ item

= Value of $\left(\frac{23}{2}\right)^{th}$ item

= value of 11.5^{th} item

= $\left(\frac{111+113}{2}\right)$

= 112

Now constructing 2×2 contingence table:

	No. of observer ≤ Md	No. of observer ≥ Md	Total
Sister	4	7	11
Brother	7	4	11
Total	11	11	n=22

Now,

$$X^2_{cal} = \frac{n\left[|ad-bc| - \frac{n}{2}\right]^2}{(a+b)(a+c)(c+d)(b+d)}$$

$$= \frac{22\left[|16-49| - \frac{22}{2}\right]^2}{11\times 11\times 11\times 11} = 0.727$$

Now, $2P_0 = 2 \times 0.727 = 1.455 > \alpha = 0.05$, so we accept H_0. i.e. median IQ of brother does not differ from that of sister.

9. Describe rationale and method of Kruskal-Wallis one-way ANOVA test.

This test is alternative test of One-way ANOVA. Here we test the significance difference between three or more than three treatments classified as one way criteria.

The process of Kruskal-Wallis H test is as follow:

a. Set the Null and Alternative hypothesis.
b. Test Statistics: Under H_0, test statistics is

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)$$

When ranks are repeated,

$$H = \frac{\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)}{1 - \frac{\sum t}{n^3 - n}}$$

$T=t^3-t$; t = no. of times that rank is repeated.

c. Critical value:
   i. Small sample case: i.e. when k=3 & $n_1 \leq 5$
      Critical value is obtained from Kruskal Wallis table as: $P_0 = P(H \geq H_{cal})$
      Decision: If $P_0 < \alpha$, then H_0 is rejected.
   ii. Large sample case: i.e. when k>3 & $n_1>5$
      Then, $H \sim \chi^2_{k-1}$ obtained from chi-square table.

Decision: If $H \geq \chi^2_{k-1}$, then H_0 is rejected.

10. Suppose in a multiple regression model problem, the ANOVA table is as follows. How many independent variables are in the model? What is the sample size? What is the value of R^2? Carry out the overall goodness-of-fit test of the model at 5% level of significance.

Source	SS	DF
Regression	**36**	**2**
error	**64**	**32**

11. Explain the meaning of multicollinearity. How do you detect the problem of multicollinearity in multiple regression?

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

If multi-collinearity is perfect, the regression coefficients are indeterminate and the S.F. are infinite. If multi-collinearity is less than perfect, the scope coefficients though determinate possess large S.F. which means that the coefficients cannot be estimated with great precision or accuracy.

12. Describe the Cobb-Douglas production function model with its application.

Consider the following non-linear model which consists more than two variables as:

$$\gamma = \beta_1 X_2^{\beta_2} X_3^{\beta_3} \dots X_p^{\beta_p} . e^u \quad \dots\dots\dots (i)$$

Where, γ is dependent variable; $X_2, X_3, \ldots, X_p$ are independent variables; $\beta, \beta_2, \beta_3, \ldots, \beta_p$ are regression coefficients. 'u' is the error term and 'e' is the base of natural log.

Taking natural log on both sides;

$$\ln \gamma = ln\beta_1 + \beta_2 lnX_2 + \beta_3 lnX_3 + \cdots + \beta_p lnX_p + u \quad \ldots\ldots\ldots (ii)$$

The model (ii) is linear parameters and can be estimated easily. This model is useful in production function theory and is known as COBB-DOUGLAS production formula. Here we may suppose γ as output, X_2 may be labour input, K_3 may be capital input and so on.

In this model;

$$\beta_k = \frac{\partial \ln l_y}{\partial \ln X_k} = \frac{X_k}{\gamma} \cdot \frac{\partial \gamma}{\partial X_k}$$

This implies that the slope coefficient (say) β_k measures the partial elasticity of the dependent variable w.r.t. the independent variable X_k. i.e. β_k measures the percentage change in γ w.r.t. percentage change in X_k.

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

i) If the sum is 1, then there is a constant return to the scale. i.e. doubling the inputs will double the outputs, tripling the inputs will triple the outputs, and so on.

ii) If the sum is less than 1, there will be decreasing return to scale. i.e. doubling the input will return less than double the output.

iii) If the sum is more than 1, there will be increasing returns to scale. i.e. doubling the inputs will return more than double the outputs.

13. Define partial correlation coefficient. If $r_{12} = 0.33, r_{13} = 0.40, \& r_{23} = 0.76$ then compute $r_{13.2}$ & $r_{23.1}$

Solution:

Here, $r_{12} = 0.33, r_{13} = 0.40, \& \, r_{23} = 0.76$

Now;

$$r_{13.2} = \frac{r_{13} - r_{23}.r_{12}}{\sqrt{1 - r_{23}^2}.\sqrt{1 - r_{12}^2}} = \frac{0.40 - 0.76 \times 0.33}{\sqrt{1 - 0.76^2} \times \sqrt{1 - 0.33^2}} = 0.409$$

$$r_{23.1} = \frac{r_{23} - r_{12}.r_{13}}{\sqrt{1 - r_{12}^2}.\sqrt{1 - r_{13}^2}} = \frac{0.76 - 0.33 \times 0.40}{\sqrt{1 - 0.33^2} \times \sqrt{1 - 0.40^2}} = 0.726$$