EASTERN UNIVERSITY, SRI LANKA

THIRD EXAMINATION IN SCIENCE - 2005/2006

SECOND SEMESTER(March/April, 2008)

ST 302 - SAMPLING THEORY

(PROPER & REPEAT)

Answer all Questions **Time: Three hours**

Q1. (a) Define a "Sampling Unit" in terms of various context:
Enumeration, Recording, Analysis and Presentation. [20 marks]

(b) Describe six advantages and disadvantages each of using Sampling Techniques. [30 marks]

(c) What is meant by "Sampling Errors" and "Non Sampling Errors"?
Describe six important circumstances, where "Non Sampling errors occur in a sample survey. [50 marks]

Q2. (a) Prove that in Simple Random Sampling without replacement (SRSWOR) the sample mean is an unbiased estimator of the population mean and the variance of the estimator $\overline{y}$ is given by,

$$\text{Var}(\overline{y}) = \left[1 - \frac{n}{N}\right]\frac{S^2}{n}, \text{ where } S^2 = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - \overline{Y}\right]^2.$$

[60 marks]

(b) An industry has 36,000 employees. A random sample of 1000 employees were asked to state the number of days they were absent from work in the previous six months. The results were as follows:

Days off	0	1	2	3	4	5	6	7	8
Number of Employees	451	162	117	112	49	21	5	11	2

(i) Estimate the average number of days "Days off" taken by workman in industry and 95% confidence interval.

(ii) Find a 95% confidence interval for the proportion of employees absent more than 3 day.

[40 mar

Q3. (a) Prove that, in Stratified random sampling, the variance of the estimator $\overline{y}_{st}$ given by,

$$\text{Var}(\overline{y}_{st}) = \sum_{h=1}^{L} \left[\frac{1}{n_h} - \frac{1}{N_h}\right] w_h^2 S_h^2,$$

where $w_h = \dfrac{N_h}{N}$ is the proportion of the total population in stratum h, $S_h^2 = \dfrac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[Y_{hi} - \overline{Y}_h\right]^2$ is the variance in stratum h, n_h is the sample siz in stratum h, L is the number of strata and assume the samples are taken independently from each stratum and in simple random sampling.

If the sampling fraction $f_h = \dfrac{n_h}{N_h}$ are negligible in all strata then show that

$$\text{Var}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{w_h^2 S_h^2}{n_h}.$$

[50 marks]

(b) With two strata, a sampler would like to have $n_1 = n_2$ for administrative convenience instead of using the values given by *Neyman's* allocation. If $\text{Var}(\overline{y}_{st})$ and $\text{Var}(\overline{y}_{st})_{opt}$ denote the variances given when $n_1 = n_2$ and *Neyman's* allocation respectively, show that the fractional increase in the variance is,

$$\frac{\text{Var}(\overline{y}_{st}) - \text{Var}(\overline{y}_{st})_{opt}}{\text{Var}(\overline{y}_{st})_{opt}} = \left[\frac{r-1}{r+1}\right]^2,$$

where $r = \dfrac{(n_1)_{opt}}{(n_2)_{opt}}$ as given by *Neyman's* allocation and ignore the sampling fraction.

[50 marks]

Q4. (a) Define a "Linear Systematic Sample" and show that its sample mean is an unbiased estimator of the population mean. Show also that the variance of the estimated mean $Var(\overline{y}_{sys})$ is given by,

$$\text{Var}(\overline{y}_{sys}) = \left[\frac{N-1}{N}\right] S^2 - \left[\frac{(n-1)k}{N}\right] S^2_{wsy},$$

where $S^2_{wsy} = \frac{1}{k(n-1)} \sum_{r=1}^{k} \sum_{i=1}^{n} [Y_{ri} - \overline{y}_r]^2$ is the sum of squares among units which lie within the same systematic sample, $S^2 = \frac{1}{(N-1)} \sum_{r=1}^{k} \sum_{i=1}^{n} \left[Y_{ri} - \overline{Y}\right]^2$ and $\overline{Y}$ is the population mean.

[40 marks]

(b) The data in following table are small artificial population which exhibits a fairly steady rising trend. Each column represents a systematic sample and the rows are the strata. Compare the precision of systematic sampling, random sampling and stratified sampling.

Data for 10 systematic samples with $n = 4$, $k = 10$, $N = nk = 40$.

Strata	1	2	3	4	5	6	7	8	9	10	Total
I	0	1	1	2	5	4	7	7	8	6	41
II	6	8	9	10	13	12	15	16	16	17	122
III	18	19	20	20	24	23	25	28	29	27	233
IV	26	30	31	31	33	32	35	37	38	38	331

[60 marks]

Q5. (a) For a stratified random sampling without replacement, the variance of the estimated proportion p_{st}, of units in a population possessing a certain attribute is

$$\text{Var}(\overline{p}_{st}) = \sum_{h=1}^{L} \frac{w_h^2 P_h Q_h}{n_h} [1 - f_h].$$

The cost(in suitable units) of data collection in a stratified sample survey is given by the fraction

$$C = c_0 + \sum_{h=1}^{L} c_h n_h,$$

where c_h is the cost per individual observation in stratum h and c_0 is th
cost of the survey.

(i) Show that the sample allocation that minimizes $V + \lambda C$, where V
variance of the estimated proportion (p_{st}) and λ is a positive const
given by,

$$n_h = w_h \sqrt{\frac{P_h Q_h}{\lambda C_h}},$$

(ii) Show how to choose λ so that the optimal allocation minimizes the
cost of sampling for fixed variance V.

[40 m

(b) A survey is to be conducted to determine the proportion of households
city living in rented houses. The 2026 households in the city are divide
into four strata. The following data are given.

Stratum	Population Size	Estimated proportion renting	Sampling cost per household
1	1190	0.75	9
2	523	0.50	9
3	215	0.20	16
4	98	0.12	16

(i) For the above data, evaluate $n_h = w_h \sqrt{\frac{P_h[1-P_h]}{\lambda C_h}}$ for $h = 1, 2, 3, 4$
terms of λ.

(ii) Assuming it is required to estimate the proportion of households living
rented houses to within 0.1 of the true value with 95% confidence. D
termine the appropriate value for λ. Hence find the total sample size a
optimal strata allocations that minimizes the total cost of sampling.

(iii) What is the total survey cost?

[60 mark

Q6. In order to estimate the total cattle population in a district consisting of 1238 villages, a simple random sample of 16 villages was selected. The number of cattle recorded in the survey, together with the most recent census figures, are given below.

Village	Number of Cattle: Survey	Number of Cattle: Census	Village	Number of Cattle: Survey	Number of Cattle: Census
1	654	623	9	292	371
2	696	690	10	555	298
3	530	534	11	2110	2045
4	315	293	12	592	1069
5	78	69	13	707	706
6	640	842	14	1890	1795
7	692	475	15	1123	1406
8	210	161	16	115	118

The census showed that there were 680,900 cattle in the 1238 village. Estimate the total cattle population from the survey data using

(a) The ratio estimator.

(b) The regression estimator.

Also estimate and compare the efficiencies of these estimators relative to an estimator based on the survey information alone.

Construct an appropriate 95% confidence interval for the number of cattle in the 1238 village for the above two estimates. [100 marks]