Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTERS OF SCIENCE IN STATISTICAL SCIENCE

END OF SEMESTER EXAMINATION

STA 8327 ANALYSIS OF COMPLEX SAMPLES SURVEY DATA

DATE: 25th August 2021 Time: 3 Hours

Instructions

1. This examination consists of **FIVE** questions.

2. Answer **Question 1 (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks)

(a) Discuss the six steps in applied survey data analysis (6 Marks)

(b) Non response is often one source of bias in household surveys.

(i) Giving examples, explain the any three types of non response in a household survey (3 Marks)

(ii) Hence, discuss any three ways to reduce non response (3 Marks)

(c) Explain the differences between Missing Completely at Random (MCAR) and missing at random (MAR) in incomplete data analysis. (2 Marks)

(d) The data below shows invoice amounts (in thousands Kshs) recorded by an auditor for quarter 1-4 of a financial year. A blank space shows that the invoice amount is missing. Using the data below, illustrate how the following imputation strategies are used in missing data analysis:

Quarter1	Quarter2	Quarter 3
25	36	24
45		67
20	55	
		60

66	67	54
	56	50

(i) Mean based imputation (1 Mark)
(ii) Median based imputation (2 Mark)
(iii) Regression imputation (3 Marks)

Question 2 (20 Marks)

(a) Let the population consist of N clusters of M elements each. Suppose n clusters are selected from N clusters by simple random sampling without replacement,

Define Y_{ij} = the value of the characteristic under study for the j^{th} element of the i^{th} cluster

$\bar{Y}_i = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ =the mean per element of the i^{th} cluster

$\bar{Y} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}$ =the population mean

$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2$ the mean square between elements of the i^{th} cluster

$S_b^2 = \frac{1}{N-1}\sum_{j=1}^{M}\left(Y_i - \bar{Y}\right)^2$ =the mean square between cluster mean

Suppose the clusters are equal in size.

Determine the variance of unbiased estimator of $\bar{Y}$ (10 Marks)

(b) Consider the following observations of a random variable Y: 3, 7, 2, 10, 5. Determine the Jackknife estimate of the variance, hence determine its bias (6 Marks)

(c) Assume X=class of employees classes: 1-10; 11-50; 51-100; > 100
Y = turnover
Assume that X is always observed, but Y is affected by nonresponse. Explain the following types of nonresponse

(i) Missing completely at random (2 Marks)
(ii) Missing at random (2 Marks)

Question 3 (20 Marks)

(a) Let π_i be the probability of including unit i in the sample. Define α_i ={1, if the ith unit is in the sample, and 0 otherwise. Further, let Y and X be the survey variable and auxiliary variable respectively. Prove that the Horvitz-Thompson estimator $\bar{y}_{HT} = \sum_{i\in s}\frac{y_i}{N\pi_i}$ is unbiased for $\bar{Y}$,the population mean (4 Marks)

(b) (i) Derive the simplification of the simplification for the variance of the Horvitz –Thompson, $Var\left(\bar{y}_{HT}\right)$ due to Yates and Grundy (8 Marks)

(ii) Explain a drawback for the variance estimator due to Yates and Grundy (1 Mark)

(c) Suppose a prediction model for a random variable $Y_i = \begin{cases} E(Y_i) = \beta x_i \\ Var(Y_i) = \sigma^2 x^2 \\ Cov(Y_i, Y_j) = 0, i \neq j \end{cases}$

where x_i is an auxiliary variable corresponding to the survey variable. Show that the Horvitz-Thompson estimator $\bar{y}_{HT}$ is unbiased under the above model (7 Marks)

Question 4 (20 Marks)

(a) Let $y_1, y_2, y_3, ...y_N$ be a sample drawn from a super population consisting of N independent random variables. Suppose the joint distribution is defined by $Y_i = \begin{cases} E_\xi(Y_i) = \beta x_i \\ Var_\xi(Y_i) = \sigma^2(x_i) \\ Cov_\xi(Y_i, Y_j) = 0, i \neq j \end{cases}$

Determine that the best linear unbiased estimator of the population total T under the super population model, ξ. (8 Marks)

(b) (i) Give two assumptions in computing sample errors in multistage sampling design. (2 Marks)

(ii) Consider a two stage sampling design. Further assume that the total sample frame is divided into a number of strata show that the resultant error is similar to the variance estimate of a sample total for a stratified random sample. (8 Marks)

(c) Explain the difference between model based inference and design based inference in surveys (2 Marks)