QUESTION ONE (30 MARKS)

(a) Let π_i be the probability of including unit i in the sample. Define $\alpha_i = \{1,$ 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-Thomson estimator $\bar{y}_{HT} = \sum_{i \in s} \frac{y_i}{N_{\pi_i}}$ is unbiased for $\bar{Y}$ [3 Marks]

(b) (i) Derive the simplification for the variance of the Horvitz-Thomson $Var(\bar{y}_{HT})$ due to Yates and Grundy. [3 Marks]

(ii) Explain a drawback for the variance estimator due to Yates and Grundy. [2 Marks]

(c) In a population with $N = 6$ the values of y_i are $8, 3, 1, 11, 4$ and 7. Calculate the sample mean $\bar{y}$ for all possible simple random samples of size 2. Verify that $\bar{y}$ is an unbiased estimate of $\bar{Y}$ and show that $\text{var}(\bar{y}) = \left(\frac{N-n}{n}\right)\frac{S^2}{n}$; $S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{N-1}$. [4 Marks]

(d) For $N = 4,\ n = 2$ Consider the estimator $\hat{\bar{Y}}_{12} = \frac{1}{2}y_1 + \frac{1}{2}y_2$, $\hat{\bar{Y}}_{13} = \frac{1}{2}y_1 + \frac{2}{3}y_3$, $\hat{\bar{Y}}_{23} = \frac{1}{2}y_2 + \frac{1}{3}y_3$, where, $\hat{\bar{Y}}_{ij}$ is the estimator for the sample that has units (i, j). Prove that $\hat{\bar{Y}}_{ij}$ is unbiased and that $V\left(\hat{\bar{Y}}_{ij}\right) < V(\bar{y})$ if $y_3(3y_2 - 3y_1 - y_3) > 0$. [4 Marks]

(e) Let $Y_1, Y_2, \cdots, Y_N$ be population values from which a simple Random Sample of size n is picked under a no Replacement Scheme. Define an estimator of the population mean as $\hat{\bar{Y}} = \sum_{i=1}^{n}((\frac{y_i}{N_{\pi_i}}))$ where π_i is the inclusion probability of the i^{th} unit in the said population.

(i) Show that the defined estimator in its form is unbiased for the population mean $\bar{Y} = \frac{1}{N}\sum_i^N Y_i$ [5 Marks]

(ii) Derive the variance of the defined estimator and briefly discuss it. [4 Marks]

(f) Suppose that from certain population H, there are r_h respondents with n_h non-respondents. A sampler wishes to assign a value to every missing one. In this case, explain ny use of appropriate mathematical expressions and or equations, how the sampler would perform;

(i) Random Imputation [2 Marks]

(ii) Nearest Neighbour Imputation [3 Marks]

QUESTION TWO (20 MARKS)

(a) Let a population consist of N units and suppose that the population is divided into k mutually disjoint strata in such a way that the i^{th} stratum has N_i units. A surveyor requires a sample of size n from this stratified population where Y_{ij} represents the j^{th} observation in the i^{th} stratum. Being that from the i^{th} stratum, a sample of size n_i is picked with y_{ij} being the j_{th} sampled measurement from the i_{th} stratum.

(i) Let $\bar{y}_w = \sum_i^k \left(\frac{N_i}{N}\bar{y}_i\right)$, Where $\bar{y}_i$ is the mean of the sampled units from the i^{th} stratum. Show that $\bar{y}_w$ is an unbiased estimator of the population mean $\bar{Y}$, derive its variance and discuss the circumstances under which it would vanish. [10 Marks]

(ii) The sampler during the pilot phase tried sampling from this population and realized that using a simple mean $\bar{y} = \frac{1}{n}\sum_i^N Y_i$, $\text{var}(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$

Whereas in using stratified sampling and assigning sample sizes to every stratum optimally, with a fixed overall sample size n.

$$\text{var}(\bar{y}) = \frac{1}{n}\Big(\sum_i^N \frac{N_i}{N} S^2\Big) - \frac{1}{N}\Big(\sum_i^N \frac{N_i}{N}\Big)S^2$$

By comparing the two variances, advice the surveyor, with reasons, on the best scheme to adopt. [10 Marks]

QUESTION THREE (20 MARKS)

(a) Assume that a population has NM elements grouped in N clusters each with size M_i. Suppose that one picks a sample of size n from this population and wishes to estimate the population mean, say, $\bar{Y}$ using this sample. The estimators $\hat{\bar{Y}}_1 = \frac{1}{n}\sum_i^n \bar{y}_i$ and $\hat{\bar{Y}}_2 = \frac{1}{n}\sum_i^n \frac{M_i}{\bar{M}}\bar{y}_i$ are proposed, where $\bar{y}_i$ is the mean of the sampled values from the i^{th} cluster. Derive the variances of the two estimators and choose the best one of the two citing reasons for your choice. [20 Marks]

QUESTION FOUR (20 MARKS)

(a) Briefly explain with the aid of relevant and logical examples, the concept of;

(i) Repeated sampling

(ii) Double sampling

as used in studies of finite population [6 Marks]

(b) Consider a fixed population model in which the population under study is split into two parts, the respondents R and the non-respondents M with r denoting the sample from the respondents and m representing the sample from the non-respondents. Defining $y_1, y_2, y_3, \cdots, y_r$ as the observations from the respondents with a sample mean of $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$ derive an expression for the bias of $\bar{y}_r$ as an estimator of the population mean $\bar{Y}$ and briefly discuss your result. [8 Marks]

(c) One of the main drawbacks of systematic sampling is that it's a Simple Random Sample with only one cluster. Because of this, it is never possible to propose an estimator of the variance based on such a sample. Carefully explain a possible method that can be employed to address this problem and provide a real life example to illustrate your proposition.

[6 Marks]