Indian Statistical Institute, Bangalore
B. Math (III)
Second Semester 2009-2010
Mid-Semester Examination : Statistics (V)
Sample Surveys and Design of Experiments.

Date: 23-02-2010 **Maximum Score 50** **Duration: 3 Hours**

1. Consider the following algorithm to select a unit from the finite population $U = \{1, 2, 3, \cdots, 150\}$. Suppose we generate a $4-digit$ observation $d_4d_3d_2d_1$ on a uniform random variable taking values in the set $\{0000, 0001, 0002, \cdots, 9998, 9999\}$. We render the observation $d_4d_3d_2d_1$ ineffective if

   $a)$ $d_4d_3d_2d_1 = 0000$ or $b)$ $d_4 = 0$ and $d_3d_2d_1 > 900$.

   We then generate a fresh observation. We continue this procedure until such times as

   $a_1)$ $d_4d_3d_2d_1 \neq 0000$ and $b_1)$ $d_4 > 0$ *or* $d_3d_2d_1 \leq 900$; in which case:

   (a) If $001 \leq d_3d_2d_1 \leq 900$ we define $r = d_3d_2d_1 \bmod(150)$. We identify $r = 0$ with 150. We then set $i = r$.

   (b) If $d_3d_2d_1 > 900$ we define $r = d_3d_2d_1 \bmod(150)$. If $d_3d_2d_1 = 000$ we define $r = 000 \bmod(150) = 100$.
   Set $i = [r + (j-1) \times 50] \bmod (150)$ if $d_4 \in A_j$, $j = 1, 2, 3$; where we have $A_1 = \{1, 2, 3\}$, $A_2 = \{4, 5, 6\}$ and $A_3 = \{7, 8, 9\}$. Again, $i = 0$ is identified with 150.

   We then select unit i from the population $U = \{1, 2, 3, \cdots, 150\}$.

   Find the probabilities of selection of different units in the population assigned by our algorithm.

   [12]

2. For a given population of N individuals values $x > 0$ of an auxiliary variable are known. The units in the population have been numbered or labelled according to nondecreasing order of their $x-$values. The population is divided into L clusters such that 'smallest' N_1 units are in the first cluster next N_2 units are in the second cluster and so on. Here $\sum_{h=1}^{L} N_h = N$. Let x_{hj} denote the $x - value$ of the jth unit in the hth cluster $1 \leq j \leq N_h$; $1 \leq h \leq L$. Let $X_h = \Sigma_{j=1}^{N_h} x_{hj}$, $1 \leq h \leq L$. For practical considerations the L clusters are formed so that $X_1, X_2, \cdots, X_L$ are roughly, if not exactly, equal. Such clusters would also be fairly $x-$homogeneous by their very formation. Suppose we use the following two-step selection procedure. We first select a cluster with *probability proportional to* X_h. We then select a unit from the selected cluster with *probability proportional to* x_{hj}. We repeat this two-step procedure n times independently. Based on these n draws (including possible repetitions) suggest an estimator for the population mean $\overline{Y} = \frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}$. Is your estimator unbiased? Obtain and estimate its *Mean Squared Error (MSE).*

   [10]

3. In *simple random sampling with replacement (SRSWR(n))* let $\overline{y}_d$ denote the mean based on distinct units. Obtain an unbiased estimator for $Var\left(\overline{y}_d\right)$.

[6]

4. An *SRSWOR* sample of size n is selected from the population of N units. Let $r_i = \frac{y_i}{x_i}$, $x_i > 0$, $1 \leq i \leq N$, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$ and $\overline{R} = \frac{1}{N}\sum_{i=1}^{N} r_i$.

   Show that $Cov(\overline{x}, \overline{y}) = \frac{N-n}{nN}\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N\overline{X}\ \overline{Y}\right]$.

   Define $e(a,b,c) = a\overline{r}\overline{X} + b\overline{r}\ \overline{x} + c\overline{y}$, where a, b, c are real numbers. Find conditions on a, b, c such that the estimator $e(a,b,c)$ is unbiased for $\overline{Y}$. Hence show that *Hartley-Ross Estimator* $e_{HR} = \overline{r}\overline{X} + \frac{n(N-1)}{N(n-1)}\left(\overline{y} - \overline{r}\ \overline{x}\right)$ is unbiased for $\overline{Y}$.

[10]

5. The purpose of the survey is to estimate $\theta\left(w_1, w_2\right) = w_1\overline{Y}_1 + w_2\overline{Y}_2$ the given linear combination of the stratum means $\overline{Y}_1$ and $\overline{Y}_2$, of two strata into which the population has been divided, w_1, w_2 are real numbers. *SRSWOR* samples of sizes n_1 and n_2 are to be selected from within strata independently. If the cost function is given by $C = c_1 n_1 + c_2 n_2$, find the best values of n_1 and n_2 for estimating θ. In particular consider the cases $a)$ $\theta = \overline{Y}_1 - \overline{Y}_2$, difference between the stratum means and $b)$ $\theta = \overline{Y}$ the population mean.

[12]

6. Obtain π_i and π_{ij}, first and second order inclusion probabilities, $1 \leq i \neq j \leq N$ under *Midzuno-Sen sampling design.*

[08]