Final Examination, Fall 2010

Instructions. You may use calculators and up to three double-sided notebook sheets, but no other study materials or aids. You need not simplify numerical expressions *except where they are needed for a later part of the same problem.* You may ignore the **fpc** whenever $n/N \leq .005$. Each problem counts the same.

(1). Give brief verbal descriptions (not necessarily using equations) for the following.

(a). Explain what **poststratification** or a *poststratified estimator* is.

(b). What is meant by the **Design Effect** for a sample survey design ?

(2). A survey sample of Households is broken down according to *Urban* and *Rural* addresses, because it is known that different proportions of sampled Households at these two types of addresses will agree to respond to the survey. Then the information on a Household attribute Y_i of interest is recorded for the sampled Households which agree to respond to the survey. (You may assume that the survey was essentially a SRS within the *Urban* and *Rural* subpopulations.) The survey data are summarized, by *Urban* vs. *Rural* address, in the following Table:

	Urban	Rural
# HH's in pop'n	10000	8000
#sampled HH's	113	87
# responding HH's	90	60
Tot. Y_i among responders	1800	360
Tot. Y_i^2 among responders	50400	3600

(a) Using *Urban* and *Rural* as distinct weighting classes, give the best unbiased estimates you can for the Urban and Overall population totals of Y_i. (Do not give standard errors of your estimators for this part.)

(b). Treating the *Urban* sample size as fixed (as though determined nonrandomly from the outset) and treating the Responders as a Small Domain within the *Urban* stratum: what does the ratio 1800/90 estimate (approximately unbiasedly) in this setting, and what is the standard error of this estimator ?

(3). A sampling experiment is conducted by sampling 100 individuals at random out of a target population of 2000 and tabulating their monthly incomes (y_k) and educational levels (x_k = highest grade attained, including college as 13 to 16). The results of the study can be summarized as follows:

$$\overline{x}_s \;=\; 12.8, \quad \overline{y}_s \;=\; 3072, \quad s^2_{y,s} \;=\; (1994)^2, \quad s^2_{x,s} \;=\; (3.5)^2$$

$$s_{xy,s} \;=\; \frac{1}{99} \sum_{k \in s} (x_k - \overline{x}_s)\,(y_k - \overline{y}_s) \;=\; 2805$$

Assume that it is known that the populationwide average educational level $\overline{X}_U = 13.4$.

(a) Give a 95% two-sided confidence interval for the average income of the target population based upon the parameters determined in the whole population from the linear regression model:

$$E(y_k) \;=\; \beta\, x_k \quad , \qquad V(y_k) \;=\; \sigma^2 x_k$$

(b) Give an approximately unbiased 95% two-sided confidence interval for average income based instead upon the parameters determined in the whole population from the linear regression model:

$$E(y_k) \;=\; \beta_0 + \beta_1 x_k \quad , \qquad V(y_k) \;=\; \sigma^2$$

(4). A survey is to be conducted on a population of $N = 200,000$ adults, grouped into 50 blocks of 4000 persons each. First a SRS of 10 blocks will be taken, and then within each sampled block a SRS of 40 people. Suppose for an attribute Y_i of interest, that the overall **population** is thought from previous surveys to have within-block variances all approximately $S_b^2 = 200$, $b = 1, \ldots, 50$, and within-block means $\bar{y}_b$ approximately satisfying

$$\frac{1}{49} \sum_{b=1}^{50} \left(\bar{y}_b \;-\; \bar{Y}\right)^2 \;=\; 2000 \quad , \qquad \bar{Y} \;=\; \frac{1}{50} \sum_{b=1}^{50} \bar{y}_b \;=\; 90$$

(a) Find the population-wide variance $S_Y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$.

(b) For the projected two-stage sample, find the theoretical Coefficient of Variation of the unbiased estimator that would be used to estimate the population total of Y_i.

(5). A sample from a large population ($N > 10^5$) is drawn in the form of 8 independent groups, each of size 100 persons, with all eight drawn by a complex hierarchical design with exactly the same probabilistic mechanism. For each person sampled, the year's total federal and state income taxes paid (in the just-completed tax year) were recorded, and the total of these taxes over all persons in group g are denoted τ_g. Find a 95% Confidence Interval for the Average per-person taxes paid if

$$\sum_{g=1}^{8} \tau_g = 3.2e6 \ , \qquad \sum_{g=1}^{8} \tau_g^2 = 1.4252e12$$

(6). A large stratified survey, in a population of size 100,000, is being planned to measure an attribute Y_i based on three strata of sizes $N_1 = 45000$, $N_2 = 40000$, $N_3 = 15000$. Information about within-stratum standard deviations S_h and costs per observation C_h from previous years' surveys using the same strata can be summarized in the following Table:

Stratum h	S_h	C_h
1	26	25
2	14	49
3	41	16

(a) Suppose that the data-collection budget for the new survey is \$5000. What are the optimum stratumwise sample sizes to use in the new survey, in order to make the mean-squared error as small as possible ?

(b) Find the theoretical standard error of the estimator of population-*average* Y-attribute you will obtain by doing the survey with the stratum sample sizes you found in (a).