Eric Slud 11/2/05

Stat 440 In-Class Test

Instructions. The Test runs from 5–6:15 pm. It is closed-book, but you may *and should* use a 2-sided notebook sheet of formulas for reference, and a calculator (which may be shared with other students). You need not provide simplified numerical answers except where specifically requested. *Point values for all problem parts are shown. 100 points is a perfect score.*

NB. Some numerical values are in scientific notation, e.g., $9.e10 = 9 * (10^{10})$.

(1). Suppose that we are interested in estimating the population average $\bar{Y}$ for the attributes $Y_1, \ldots, Y_N$ in a large finite population, and that we know for all members of the population a categorical auxiliary variable X_i taking the three possible values $1, 2, 4$. Suppose that the population proportions with which $X_i = h$ are respectively $p_h = .4, .4, .2$ for $h = 1, 2, 4$. Suppose also that we know or think we know that the correlation between X_i and Y_i values is 0.5, and that the variance S_h^2 of Y_i values for the population stratum in which $X_i = h$ is $3 \cdot h$ for $h = 1, 2, 4$, and that the average $\bar{Y}_h$ of Y-attribute values for the three strata is respectively around $5, 9, 16$.

(a) (*15 points*) If the population is divided into strata U_h defined as the individuals i with $X_i = h$, then find $(SSB)/N$ and $(SSW)/N$.

(b) (*15 points*) Find the coefficient of variation of the regression estimator of $\bar{Y}$ based on a SRS of size $n = 100$ in this population, ignoring the *fpc*.

(2). In Lohr's Agricultural Census data `agsrs.dat`, a SRS of n=300 out of N=3078 counties, it is found that a total of 175 sampled counties had at least 500 farms in 1987. Consider the following table of results from `agsrs.dat`, related to the sample $\mathcal{S}$, the attribute

$$Y_i = \texttt{ farm acreage for county i in 1987},$$

and the domain

$$D = \texttt{indices for counties with} \geq 500 \texttt{ farms}:$$

$$\sum_{i \in D \cap \mathcal{S}} Y_i = 52873455 \ , \qquad \sum_{i \in D \cap \mathcal{S}} Y_i^2 = 2.733721e13$$

(a) (*25 pts*) Suppose that you know also that the national total number of counties with at least 500 farms was 1800 in 1987. Give the best 95%confidence interval you can for the national average number of acres per county in counties with at least 500 farms.

(b) (*10 pts extra*) What is the regression estimator in this problem, with attribute $u_i = Y_i \cdot I_{[i \in D]}$ and predictor $x_i = I_{[i \in D]}$?

(3.) A survey is designed to measure the average over a large community of household incomes $t_i = \sum_{j=1}^{4} Y_{ij}$ for 4-person households ($M = 4$). Suppose that the study design is to take a SRS of $n = 150$ households from a large master address list of $N = 10000$ households, and then to sample $m = 2$ members randomly from each sampled household for a personal interview. Assume that the sample yields data

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} Y_{ij} = \$\,3.8e6 \quad , \qquad \sum_{i \in \mathcal{S}} d_i^2 = \$\,5.4e11 \quad , \qquad s_t^2 = \$\,3.6e9$$

where d_i denotes the **difference** between the incomes of the two sampled individuals in the i'th household.

(*25 pts*) Find a 95% confidence interval for average household income in this population.

(4). A campus population of size $N = 9000$ is to be surveyed by a stratified sample for the prevalence of a certain disease, based upon three strata of respective sizes $N_h = 1000, 3000, 5000$ for $h = 1, 2, 3$. The costs of sampling individuals from these strata are estimated to be respectively $40, 20,$ and 10 dollars per person. The campus health authorities believe that roughly 1% of stratum 1, 5% of stratum 2, and 12% of stratum 3 will test positive for the disease.

(a) (*20 pts*) What is the optimal number of individuals to sample in each stratum if the **total** budget for data-collection in the survey is \$2000 ?

(b) *15 pts*) Suppose that the same population were to be sampled by SRS. About how much would the SRS cost if you want to achieve the same MSE as in (a) in estimating the proportion of the population who have the disease ?