E. Slud 10/27/05

Sample Problems for Stat 440 In-Class Test

Instructions. The Test will be in-class on Wednesday, November 2, 5–6:15 pm. It will be closed-book, but you may use a (two-sided) notebook sheet of formulas for reference, and you should bring a hand-calculator if possible (or you may share calculators with other students). You need not provide simplified numerical answers except where specifically requested, e.g. when you are asked how much of an MSE improvement one estimator gives over another.

The problems will be similar in scope and difficulty to the Sample Problems given here, but there are more problems here than will be on the test. The coverage of the test is all of the non-starred sections in Chapter 1–5 of the Lohr book, which is essentially the same as the material covered in Lectures so far.

(1). In Lohr's Agricultural Census data `agsrs.dat`, a SRS of n=300 out of N=3078 counties, it is found that a total 160 sampled counties had at least 500 farms in 1992. Consider the following table of results from `agsrs.dat`, related to the sample $\mathcal{S}$, the attribute

$$Y_i = \texttt{farm acreage for county i in 1992},$$

and the domain

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

$$\sum_{i \in D \cap \mathcal{S}} Y_i \;=\; 50468635 \;, \qquad \sum_{i \in D \cap \mathcal{S}} Y_i^2 \;=\; 2.733721e + 13$$

Suppose that you know also that the national total number of counties with at least 500 farms was 1484 in 1992. Give the best 95%confidence interval you can for the national average number of acres per county in counties with more than 500 farms.

(2.) The prevalence of a certain disease is to be measured by means of a sampling study. The target population, of size 10000, is the undergraduate population at a large private college. Based on known demographic characteristics, this population is subdivided into 3 strata, of sizes 5500, 3500, and 1000. The campus health authorities believe before doing the study that roughly 1% of stratum 1, 5% of stratum 2, and 12% of stratum 3 will test positive for the disease. If these guesses are about right, give arithmetic expressions for the width of the (symmetric) 95% confidence intervals for the standard unbiased estimators of the *population-wide fraction who would test positive* based on a sample of fixed size 300 with each of the following designs:
(a) a purely random sample (without replacement);
(b) a stratified sample with proportional allocation; and
(c) the stratified random sample with optimal allocation.

Use the health authorities' guesses as though they were known and correct, and note that the attribute under consideration is 1 or 0 according as an individual does or does not test positive for the disease.

(3.) A sociological experimenter has to make a choice between two sampling designs for estimating the total t of an attribute of interest, based on 3-person households ($M_i = M = m_i = m = 3$). The cost of sampling a single individual within the target population is \$10, and the cost of sampling a cluster (household) is \$25. For the population as a whole, it is known that approximately $MSW/S_Y^2 = 0.3$.

The two possible sampling methods considered are: *(i)* to sample individuals at random, and *(ii)* to sample clusters at random (single-stage cluster sample). Assume that the fixed costs of doing the survey are the same for either method. Also assume that the number N of clusters is very large, so that the ratio n/N is small *and can be treated as negligible.*

(a) To attain equal MSE in estimating t by the two methods, what is the ratio of cost under method (ii) to cost under method (i) ?

(b) For method (ii), if $(cv)_{tU} \approx 0.4$ (this is the coefficient of variation of the 3-person household totals of the y attribute), and if the cost to the experimenter due to estimation error is \$1.6 million multiplied by the squared coefficient of variationof the estimator, then what is the optimal fraction n/N of the clusters to sample ?

(4). 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 $1, 2, 3$. Suppose that the population proportions with which $X_i = h$ are respectively $p_h = .3, .4, .3$ for $h = 1, 2, 3$. Suppose also that we know or think we know that the correlation between X_i and Y_i values is 0.7, and that the variance S_h^2 of Y_i values for the population stratum in which $X_i = h$ is $2 * h$ for $h = 1, 2, 3$, and that the average $\bar{Y}_h$ of Y-attribute values for the three strata is respectively around $5, 10, 25$.

(a) 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) Compare the MSE for the Regression Estimator of $\bar{Y}$ vased on a SRS of size $n = 100$ versus the MSE based on a Stratified Estimator in a stratified survey of sample size $n = 100$ and optimal allocation.

(5). Suppose that a large population of size $N = 9000$ is divided into three strata of respective sizes $N_h = 1000, 3000, 5000$ for $h = 1, 2, 3$.

(a) In a stratified sample if size 18 from this population with proportional allocation, find the inclusion probability for each unit $i \in U_1$ and for each unit $i \in U_3$.

(b) In a 2-stage cluster sample with the strata treated as PSU's, with first-stage SRS of size 2 and with second-stage SRS samples of respective sizes $2, 6, 10$ drawn from the PSU's selected at the first stage, find the inclusion probability for each unit $i \in U_1$ and for each unit $i \in U_3$.

(c) Assuming that the stratum variances S_h^2 of attributes Y_i are respectively $2h$ in the three strata $h = 1, 2, 3$, how large must n be chosen so that (ignoring fpc's) the 95% Confidence Interval width for estimating $\bar{Y}_h$ is at most 1 in all three strata $h = 1, 2, 3$?

Additional topics that might well be covered: two-stage cluster-sampling variance formula; or one- or two-stage ratio estimator.