

**Stat/Surv 440: Make-up Take-Home for In-Class Test 11/8/07**

**Instructions:** Do as many of these problems as you can, for extra credit on the In-Class Test. The percent score you get, out of 100, will multiply .6 times the number of points you missed on the test and will add to your Test score. (Thus if you got a raw score of 65 on the test, with a recorded score of  $(65/85)*100$ , then the maximum number of points you can add to your test from this take-home is  $.6*(20/85)*100$ , or  $(12/85)*100$ .)

**On this test you may not cooperate with each other, although you may get hints from me (personally or by email).**

(1.) (25 points) The prevalence of a certain disease is to be measured by means of a sampling study. The target population, of size 10,000, 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 3% of stratum 1, 6% of stratum 2, and 10% 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 *estimated population-wide fraction testing 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.

*Note: 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.*

(2.) (25 points) A sociological experimenter has to make a choice between two sampling designs for estimating the total  $t$  of an attribute of interest, based on the following information. 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 there are approximately 3.5 individuals per cluster, along with the values of the ratios

$$\frac{\sum_{i=1}^N (M_i - 1) S_{yU_i}^2}{(K - N) S_{yU}^2} = 0.3 \quad \text{and} \quad \frac{\sum_{i=1}^N (M_i - \bar{M}) M_i \bar{Y}_{U_i}^2}{(N - 1) S_{yU}^2} = 1.44$$

where  $K = \sum_{i=1}^N M_i$  and  $\bar{M} = K/N$ . The two possible sampling methods

are: (i) to sample individuals at random, and (ii) to sample clusters at random. Assume that the fixed costs of doing the survey are the same for either method. Also assume that the population size  $N$  is very large, so that the ratio  $n/N$  is small *and can be treated as negligible*.

(a) Use the given information to find  $S_t^2/S_{yU}^2$  in this problem with unequal cluster-sizes.

(b) For equal variance in estimating  $t$  by the two methods, what is the ratio of cost under method (ii) to cost under method (i) ?

**(3.)** (25 points) A population of 3000 housing units is broken into two strata, one consisting of 300 garden-apartment clusters each containing 6 units, and the other consisting of 1200 miscellaneous units. The attribute of interest concerning this population is annual household income, measured in units of \$10,000. With subscripts indicating attribute and stratum, you are given that  $MSW_{y,1}/S_{y,U_1}^2 = 0.2$ ,  $S_{y,U_1}^2 = 9$ ,  $S_{y,U_2}^2 = 16$ .

A stratified sample is drawn from this population, with 20 clusters sampled at random from the first stratum, and two units sampled at random from each sampled cluster; and 30 units are sampled SRSWOR from the second stratum, then define an unbiased estimator (a formula) for the average household income of the target population of housing units, and give a numerical expression for its standard error.

**(4.)** (25 points) 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

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

and the domain

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

$$\sum_{i \in D \cap \mathcal{S}} Y_i = 50468635 \quad , \quad \sum_{i \in D \cap \mathcal{S}} Y_i^2 = 2.733721e13$$

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.