Sample Final Questions, Stat 440, Fall '05

Instructions. Bring calculators and up to three double-sided notebook sheets to the Final, but the Exam is closed-book. (Sheafs of pages of notes are not allowed, which is why 3 notebook sheets is the maximum.) As before, you need not simplify numerical expressions *except where they are specifically requested*. This means I will accept arithmetic expressions but not symbolic formulas: for full credit, all numbers should be filled in and only calculator arithmetic remain to be done. *You may ignore the* **fpc** *wherever* $n/N \leq .005$.

(1). *Design effect, definition and calculation in examples.* Find the design effect of a single-stage cluster sample of $n = 100$ from a large $N > 10^6$ population of 4-person households in which the attribute y being sampled has $SSB/SST = .7$.

(2). *Jackknife or Random-groups estimator of variance.* Suppose that it is desired to estimate the average rental per month of housing units within a large urban population of apartments, and that 10 systematic samples of size 80 each are taken (with overall sample fraction $< 1/2000$), yielding totals $\tau_g,\ g = 1, \ldots, 10,$ such that

$$\sum_{g=1}^{10} \tau_g \ = \ 9.6e5 \ , \qquad \sum_{g=1}^{10} \tau_g^2 = 9.4464e10$$

Suppose that these 10 systematic samples can be regarded as independent identically distributed random groups in the population. Find a 95% Confidence Interval for the average monthly rental cost of the housing units in the frame population.

(3). *Two-level sampling estimators and variances, from two-stage cluster formulas.* A survey is to be conducted on a population of 10,000 adults, grouped into 10 'blocks' of 1000 each which is supposed to be much more internally homogeneous than the population as a whole, to learn the average number of years of schooling they have had. First, a SRS of 4 blocks is taken, and then within each sample block a SRS of 20 people. When the sample is drawn, the subsamples in the 4 sampled blocks yield:

Block	2	4	6	9
Mean	15.2	13.5	11.7	14.4
Var	10.2	8.6	13.3	11.1

(a) Find a 95% CI for the average number of years of schooling in the population.

(b) Find an unbiased estimate for the population-wide variance S_Y^2 of the number Y_i of years of schooling for i in the 10,000-element frame.

(4). *Inclusion probability calculation, & HT formula implementation.* A population consists of 40,000 units subdivided hierarchically as follows: units

occur in clusters of 4, and respectively 2500, 4500, and 3000 fall in strata 1, 2, and 3. Suppose that a sample of size 200 units is drawn as follows. Within stratum 1, 35 clusters are sampled SRS, and exactly 2 of the units within each cluster are sampled SRS; within stratum 2, 50 clusters are sampled SRS and exactly 1 of the 4 units within each cluster is randomly sampled; and within stratum 3, 20 clusters are sampled SRS and all 4 of the units within the cluster are sampled.

(a) Find the single inclusion probability for each unit in the frame population.

(b) Find the Horvitz-Thompson estimate of the total property taxes paid annually, if the totals reported from the 70 units (SSU's) sampled in stratum 1, the 50 units sampled in stratum 2, and the 80 units sampled in stratum 3, are respectively 101500, 118750, and 248000.

(5). *Optimal allocation – stratified or cluster estimators.* There are about 80,000 public schools, 23% of them in central cities, 24% in non-central urban areas, 25% in towns, and 28% in rural areas. We regard these 4 types of locations as strata, and wish to estimate the average yearly income earned by teachers. Assume that the standard deviations of yearly income in these strata are respectively $S_1 = 4200$, $S_2 = 3000$, $S_3 = 1900$, and $S_4 = 2400$. Find the optimal stratum sample-sizes for a stratified sample of total size $n = 1000$ schools to estimate average yearly income, and find the theoretical MSE for the unbiased estimator to be derived from that stratified sample.

(6). *Sample size calculation.* A survey of the 9th graders in Maryland is intended to determine the proportion intending to go to a four-year college. A preliminary estimate of $p = .55$ was obtained from a small informal survey. How large must the survey be to provide an estimator with error at most .05 with probability at least 99% ?

(7). *Ratio or regression estimation.* A village contains 175 children. Dr. Jones takes a SRS of 17 of them and counts the cavities in each one's mouth, finding the frequency table:

# Cavities	0	1	2	3	4	5
# Children	5	4	2	3	2	1

Dr. Smith examines all 175 children's mouths and records that 55 have no cavities. Estimate the total number of cavities in the village's children using (a) only Dr. Jones' data, (b) both Dr. Jones' and Dr. Smith's data. (c) Give approximately unbiased estimates for the variances of your (approximately) unbiased estimators in both (a) and (b).
This problem requires a little more arithmetic than I would ask on the Final, but you should make sure you can do it !

(8). *Unequal probability (pps) sampling.* A population of voters is known to consist of three precincts of size 1000, 800, 1200. A Poisson sample is taken,

i.e. individuals are independently chosen to be sampled, with probability .05 in the first precinct, .15 in the second, and .10 in the third, resulting in samples of respective sizes 65, 110, and 105 people in the three precincts. Each sampled individual is asked whether he/she favors extension of the Patriot Act: respectively 35, 60, and 30 people in the three precincts say that they do. Find the unbiased estimator (based on the Poisson sample design) for the overall proportion of voters in the three precincts who support extension of the Patriot Act, and find an approximately unbiased estimate of the coefficient of variation of your estimator.

(9). *With replacement (ppswr) sampling.* A population of 2000 households (containing different numbers of individuals) is sampled to find out how many individuals have lived in the same town for at least 5 years. The sample is done by sampling 60 times with replacement with probabilities

# in HH	1	2	3	4+
# HH's	250	400	620	730
Prob$*10^4$	2.088	4.175	5.219	6.263
# HH sampled (with mult.)	4	9	17	30
# 5+ yr (persons)	3	12	25	77
sum over HH of (#5+ yr)^2	3	18	45	251

That is, the probability with which each house is sampled on each draw is read off from the third line of the Table as a function of the number of persons in the HH.

When the sample is drawn, the count of individuals in each sampled household who have lived in the town for at least five years is recorded, with multiplicity, if sampled more than once, as given in the 4th line of the Table.

(a). Give an unbiased estimate for the total number of individuals in the town who have lived there at least five years, and also give its SE.

(b) (*Trickier than the others.*) Can the estimator be improved by poststratification (treating the sampling as with-replacement but stratified within HH's of sizes as given in the columns of the Table) ? Assume it is known that the total number of sampled individuals in the 4+ group was 167 and the total number of such individuals in the town is 4088.

(10). *Definitions.* Give a two or three sentence (or equation) definition for each of the following terms in sample survey theory.

(a). ICC (Intra-Cluster Correlation Coefficient).
(b). Design Effect.
(c). Systematic Sample.

(d). Nonresponse weighting class.
(e). Model-based variance for the estimator $\hat{t}_y$ of a total based on a SRS.
(f). Raking a two-way table of (aggregated) survey weights.