Stat 5201, April 15, 2011 Name _____

This test is closed book but you may use both sides of one 8 by 11 formula sheet. You may not use a calculator. It is enough to express any numerical answer as a formula which can easily be evaluated

This test was taken by 36 people. Each problem was worth 20 points. The high score was a 90 and the low score was a 22. There were 3 in the 80's, 11 in the 70's, 7 in the 60's, 7 in the 50's, 3 in the 40's and 3 in the 30's.

1. Within a metropolitan area a large HMO has three clinics. Each clinic has quite a few family practice physicians each of which has many patients. Given a physician and one of their patients by looking at the patient's records, which takes some effort, one can determine the total cost of the patient to the HMO over the past year. Develop a sampling plan that would allow the HMO to estimate the total cost of all the patients treated by their family practice physicians over the last year. Give the formulas you would use to calculate your estimate and an estimate of its variance.

Ans Clinics are strata, physicians are clusters within strata and patients are clusters with physicians. So you should take a random sample of physicians within each stratum and a random sample of patients from each selected physician. So for clinic 1 let N be the number of family practice physicians and n be the size of the sample taken from the first clinic. Let smp denote the physicians selected in the sample. For an $i \in smp$ let M_i be the number of their patients and m_i be the number of their patients selected in the sample. Let y_{ij} denote the total cost for the jth patient in the sample for physician i and $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i$. Then the estimate for the total cost for patients in clinic one is

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in smp} M_i \bar{y}_i$$

with it's estimate of variance given in formula (5.21) of page 184 of the text. The estimate for the HMO is just the sum of the estimates from the three clinics with variance equal to the sum of the strata variances. .

2. A company with 2000 employees 1200 of which are men selected 40 men at random and 40 women at random to participate in a survey. One question asked if they were satisfied with the the company's health benefits' plan. Of the men surveyed 30 were satisfied while only 20 of the women were. Find the 95% confidence interval for proportion of employees who are satisfied with the health plan. If the observed strata sample variances were the true variances what would have been the optimal allocation of the 80 observations for the original sample.

Ans. For this stratified sample the usual unbiased estimate along with its estimated variance is

$$\frac{1200}{2000}\frac{30}{40} + \frac{800}{2000}\frac{20}{40}$$
$$(\frac{3}{5})^2(1 - \frac{40}{1200})\frac{0.75 \times 0.25}{39} + (\frac{2}{5})^2(1 - \frac{40}{800})\frac{0.5 \times 0.5}{39}$$

from which the CI is easily found.

Recall optimal allocation happens when $n_h \propto N_h \sigma_h$ so we have

$$n_1 = \frac{1200\sqrt{0.75 \times 0.25}}{1200\sqrt{0.75 \times 0.25} + 800\sqrt{0.5 \times 0.5}} 80 = 0.565(80) = 45$$

3. In a population it was believed that y_i, the variable of interest, is approximately proportional to an auxiliary variable x_i. Under this assumption what is a reasonable estimate of the population mean of y if in a random sample of size 20 from the population $\bar{y}_{smp} = 67.4$ and $\bar{x}_{smp} = 54.2$ were observed and the population mean of x is 61.3? What additional information do you need to find an approximate 95% confidence interval for the population mean of y.

Ans Should use the ratio estimate which is $\hat{R} \times 61.3$. where $\hat{R} = 67.4/54.2$.

You need to know $\sum_{i \in smp}(y_i - \hat{R}x_i)^2$ and the population size N unless $20/N$ is negligible.

4. A population consists of three strata of sizes 400, 500 and 700.. The strata information for individual units is not contained in the sampling frame however and can only be determined for the units in a sample.. A simple random sample without replacement of size $n = 100$ taken from the population yielded these results.

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2
1	400	25	36.5	22.6
2	500	35	52.3	29.6
3	700	40	27.3	18.5

where n_h is the number of units in the sample that belong to stratum h and N_h is the stratum size which is assumed to be known. Use this information to calculate an approximate 95% confidence interval for the population mean.

Ans This is a poststratitication problem since the n_h's are random variables. The usual point estimator is

$$\bar{y}_{post} = \sum_h \frac{N_h}{N} \bar{y}_h$$

and an approximate estimate of its variance is

$$\sum_h (1 - \frac{n_h}{N_h})(\frac{N_h}{N})^2 \frac{s_h^2}{n_h} \doteq (1 - \frac{n}{N}) \sum_h \frac{N_h}{N} \frac{s_h^2}{n}$$

since for each h we expect $n_h/N_h \doteq n/N$. For more discussion see section 4.4 of the text.

5.. A company with 1,000 employees randomly selects 10 each month to ask them about working conditions.

a) What is the probability that no one from a 5 person office gets included in the survey this year.

b) How many months would a person need to work at the company before they have a probability of at least 0.5 of being included in the monthly survey.

Ans. a. In any given month there are $\binom{1000}{10}$ possible samples and $\binom{995}{10}$ samples which do not include anyone from the office. Since samples are independent across the months the answer is $\left(\binom{995}{10}/\binom{1000}{10}\right)^{12}$

b. The probability that an individual is included in a sample is 10/1000. Hence

$$Pr(\text{employee not in n samples}) = (1 - 10/1000)^n$$

Or $(0.99)^n = 0.5$, or $n \log(0.99) = \log(0.5)$ or $n = 68.97$ or 70 months.