Stat 5201, April 14, 2010   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 using a calculator.

1. To find the average weight of cows in a region 50 farms were selected at random from a list of all farms. Then the weight of each cow at the 50 selected farms was found.

i) For this survey describe the target population, sampling frame, and observational unit.

ANS The target population is all cows in the region. The sampling frame is the list of all farms and the observational unit is an individual cow.

ii) State the estimate you would use for this survey along with its estimated variance.

ANS This is a single stage cluster sample with unequal cluster (farm) sizes. Since the number of cows per farm could vary a lot we should use the ratio estimator. If Y_i is the total weight of all cows on farm i and M_i is the number of cows on farm i then our estimate along with its estimated variance is

$$\hat{R} = \frac{\sum_{i \in smp} Y_i}{\sum_{i \in smp} M_i} \quad \text{and} \quad \widehat{V(\hat{R})} = \frac{1}{(\bar{M}_{smp})^2} \frac{1-f}{50} \sum_{i \in smp} (Y_i - \hat{R}M_i)^2/(n-1)$$

where $f = 50/N$, N is the total number of farms and $\bar{M}_{smp} = \sum_{i \in smp} M_i/50$.

2. From a list of 10,000 targeted households a random sample of size 200 was selected. In the sample it was determined that 90 of the households owned more that one computer. For this group of 90 the number of magazines they subscribed to was also noted. The results are given in the table to the right.

# of Mags	# of Resps
0	20
1	35
2	20
3	15

i) Find an estimate of the number of households in the targeted population that own more than one computer. Give the standard error of your estimate.

ANS Let p equal the proportion of households in the population than own more than one computer. Then

$$\hat{p} = 90/200 = 0.45 \quad \text{and} \quad \widehat{V(\hat{p})} = (1 - 200/10,000)(0.45 \times 0.55/200$$

and an approximate 95% CI is $\hat{p} \pm \sqrt{\widehat{V(\hat{p})}}$

ii) Estimate the average number of magazines received by households that own more than one computer.

ANS This is a domain estimation problem where the domain is the number of households in the population that own more than one computer. The number of such households, N_d, is unknown. The usual estimate is

$$\bar{y}_d = \frac{0 \times 20 + 35 \times 1 + 2 \times 20 + 3 \times 15}{90} = 4/3$$

iii) Estimate the total number of magazines received by households in the targeted population that own more than one computer.

ANS

$$\hat{t}_d = \frac{n_d}{n} N \bar{y}_d = \frac{90}{200}(10,000 \times 4/3) = 6,000$$

3. Suppose a company employs 5,000 workers of which 3,000 are men and 2,000 are women. It is interested in how far, on the average, employees live from their work. A stratified random sample was taken and the distances were recorded in miles. The results are given below.

Gender	Number in in sample	Average distance	Sample variance
Men	60	15.4	4.2
Women	40	21.6	7.7

a) Find a 95% confidence interval for the average distance from work.
ANS

$$\bar{y}_{str} = \frac{3000}{5000}15.4 + \frac{2000}{5000}21.6 = 17.88$$

$$\hat{V}(\bar{y}_{str}) = \sum_h (1 - n_h/N_h)(\frac{N_h}{N})^2\frac{s_h^2}{n_h}$$
$$= (1 - 60/3000)(3/5)^2(4.2/60) + (1 - 40/2000)(2/5)^2(7.7/40)$$
$$= 0.0247 + 0.0302 = 0.0549$$

b) Assuming the observed sample variances were the true population variances what would have been the optimal allocation of the 100 sampled units?
**ANS** Recall optimal allocation happens when $n_h \propto N_h\sigma_h$. We find that

$$n_1 = \frac{3,000\sqrt{4.2}}{3,000\sqrt{4.2} + 2,000\sqrt{7.7}}100 = 0.53(100) = 53$$

and so we should have sampled 53 men and 43 women.

4. A national chain of stores is thinking of introducing a new product. To estimate the possible demand for the product they plan to test it in a sample of n of their N stores. They believe that in a given store the sales volume for the new product should be roughly a constant proportion of the current total sales volume. Suggest a sensible sampling plan that the corporation can use to select the stores in their sample and an estimate of the total potential sales volume of the new product.

ANS Assuming the stores do not differ to much in size then a srs should work and the ratio estimate can be used. If stores do vary a lot in size then one could stratify on size and then take simple random samples within strata and again use the ratio estimator.
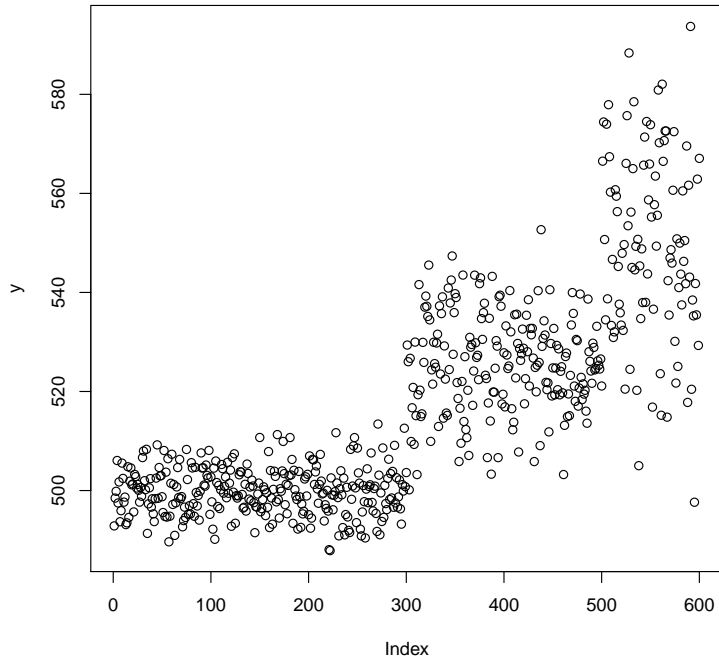
5. A client wants to select a sample of size 50 from a population which she believes should look something like the plot above. Suggest a reasonable sampling design that she can use to draw her sample. Be specific and briefly justify your answer.

ANS From the plot there appears to be 3 strata; the first with 300 units, the next with 200 units and the final with 100 units. The strata variances are obviously increasing as well. From the picture a good guess is that the standard deviation of the first stratum is $1/2$ of the second and $1/4$ of the third. Let σ denote the stratum standard deviation of the first stratum. The we should allocate

$$\frac{300\sigma}{300\sigma + 200(2\sigma) + 100(4\sigma)}50 = \frac{3}{11}50 = 13.6$$

or 14 observations to the first stratum. In the next two strata we should allocate $(4/11)50 = 18.2$ or 18 observations.