Stat 5201, April 12, 2013 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.

Fifty people took the exam. There were 6 scores in the 90's (high of 99), 3 scores in the 80's, 7 scores in the 70's, 8 scores in the 60's, 12 scores in the 50's, 5 scores in the 40's, 2 scores in the 30's, 3 scores in the 20's and 4 scores in the 10's.

1. A division of a large company has hired 200 new employees in the past year. Typically these employees have more than one supervisor. To learn something about the level of satisfaction of these new hires with their jobs the company plans to take a random sample of 20 of them and have them respond to a questionnaire. The employees selected then choose one of their supervisors to fill out another questionnaire where the supervisor evaluates their work. The results from the two surveys will then be studied. Does this seem like a reasonable course of action? If your answer is yes briefly explain why. If your answer is no briefly explain what you would do differently.

ANS Selecting the employees at random is fine but letting them select their supervisor to fill out the second questionnaire is not a good idea. The company should select the supervisor or perhaps better yet have all the supervisors of each employee fill out the second questionnaire.

2. A shipment of 100 boxes of frozen food (each box contains 8 separate packages of food) was allowed to thaw during transit. The shipper was worried that some of the packages could be spoiled. He took a random sample of 5 boxes and checked all the packages in each box. In 2 of the boxes there were 3 spoiled packages, in 1 of the boxes there was 2 spoiled packages and in 2 of the boxes there was no spoiled packages.

i) Under this sampling plan estimate the total number of spoiled packages in the entire shipment and give its estimated variance.

ii) Suppose instead that the sampling plan took a simple random sample without replacement of size 40 from the population of 800 packages and 8 spoiled packages were found in the sample. Under this sampling plan estimate the total number of spoiled packages in the entire shipment and give its estimated variance.

ANS i) This is a single stage cluster sampling design and the estimate of the population total is

$$\hat{t} = \frac{100}{5}(3+3+2+0+0) = 160$$

$$s^2_{\hat{t}} = 100^2(1-\frac{5}{100})\frac{1}{5}\frac{1}{4}\Big(2(3-1.6)^2 + 1(2-1.6)^2 + 2(0-1.6)^2\Big) = 4370$$

i) Since this is just a SRS of size 40 from a population of 800 we have

$$\hat{t} = 800\frac{8}{40} = 160$$

$$s^2_{\hat{t}} = \frac{800\times 760}{39}\frac{8}{40}\big(1-\frac{8}{40}\big) = 2494.4$$

3. i) Consider a population with 3 strata of sizes N_1, N_2 and N_3. For stratum h let σ_h denote the stratum standard deviation. Although these σ_h's are not known it is believed that $\sigma_2 \doteq (4/3)\sigma_1$ and $\sigma_3 \doteq (3/2)\sigma_2$. Use this information to allocate a sample of n observations to the three strata.

ii) Assume now that there is a fourth strata of size N_4 with a standard deviation of σ_4. Suppose now what is believe is that $\sigma_2 \doteq (5/4)\sigma_1$ and $\sigma_4 \doteq 2\sigma_3$. How can this information be used to allocate a sample of size n to the four strata.

ANS i) Note by assumption

$$\sigma_3 \doteq (3/2)\sigma_2 \doteq (3/2)(4/3)\sigma_1 = 2\sigma_1$$

Let

$$den = N_1\sigma_1 + N_2(4/3)\sigma_1 + N_3 2\sigma_1 = \sigma_1(N_1 + (4/3)N_2 + 2N_3)$$

The the proportion of the sampled allotted to stratum h is $N_h\sigma_h/den$ or

$$N_1/den, (4/3)N_2/den \text{ and } 2N_3/den$$

ii) There is no information on how to allocate the amount of the sample that should be alloted to the first two strata. One could use proportion allocation. However for the first two strata 4/9 should go to stratum 1 and 5/9 to stratum 2. Similarly for the last two strata 1/3 should go to stratum 3 and 2/3 to stratum 4.

4. Consider a population consisting of six units, $y = (y_1, y_2, y_3, y_4, y_5, y_6)$. Consider the following two step sampling design which selects a sample of size 2. In the first step an integer, say i, is selected at random from the labels $1, 2, \ldots, 6$. If $i < 3.5$ the second integer is slected at random from $\{4, 5, 6\}$ while if $i > 3.5$ the second integer is selected at random from $\{1, 2, 3\}$. Consider the estimator of the population total which is three times the sum of the values of the two units which appear in the sample. Show that this is an unbiased estimator of the population total.

ANS Let (i, j) denote a typical sample where $i < 3.5 < j$. Note there are 9 such points and each can arise in two ways. Either i is selected first or j is selected first. Each happen with probability $(1/6) \times (1/3) = 1/18$ so each (i, j) happens with probability 1/9. Now in the sum

$$3 \sum_{i<3.5<j} (y_i + y_j)(1/9) = \frac{1}{3} \sum_{i<3.5<j} (y_i + y_j)$$

y_1 appears exactly 3 times. But this is true for each y_i so the above sum must equal the population total.

5. Consider a population consisting of 20 clusters with a total population size of 940 which is similar to the population plotted on the next page. Suppose we have simple random sample of four clusters from the population which resulted in the following data

cluster	1	2	3	4
size	17	56	23	64
sample size	10	25	12	30
sample mean	32.7	36.1	30.3	33.4
sample variance	26.3	21.4	23.6	29.1

i) Give a point estimate for the population mean.

ii) Give an estimate of variance for your answer in part i). Be sure to state clearly any additional information you need to compute your answer.

ANS Let m_i be the sample size, M_i the cluster size, and $\bar{y}_i$ and s_i^2 be the sample mean and variance for cluster i. The point of the plot was to suggest that since the M_i and m_i vary while the y values of the units do not vary much we should use the ratio estimator.

i) The estimate of the population mean is

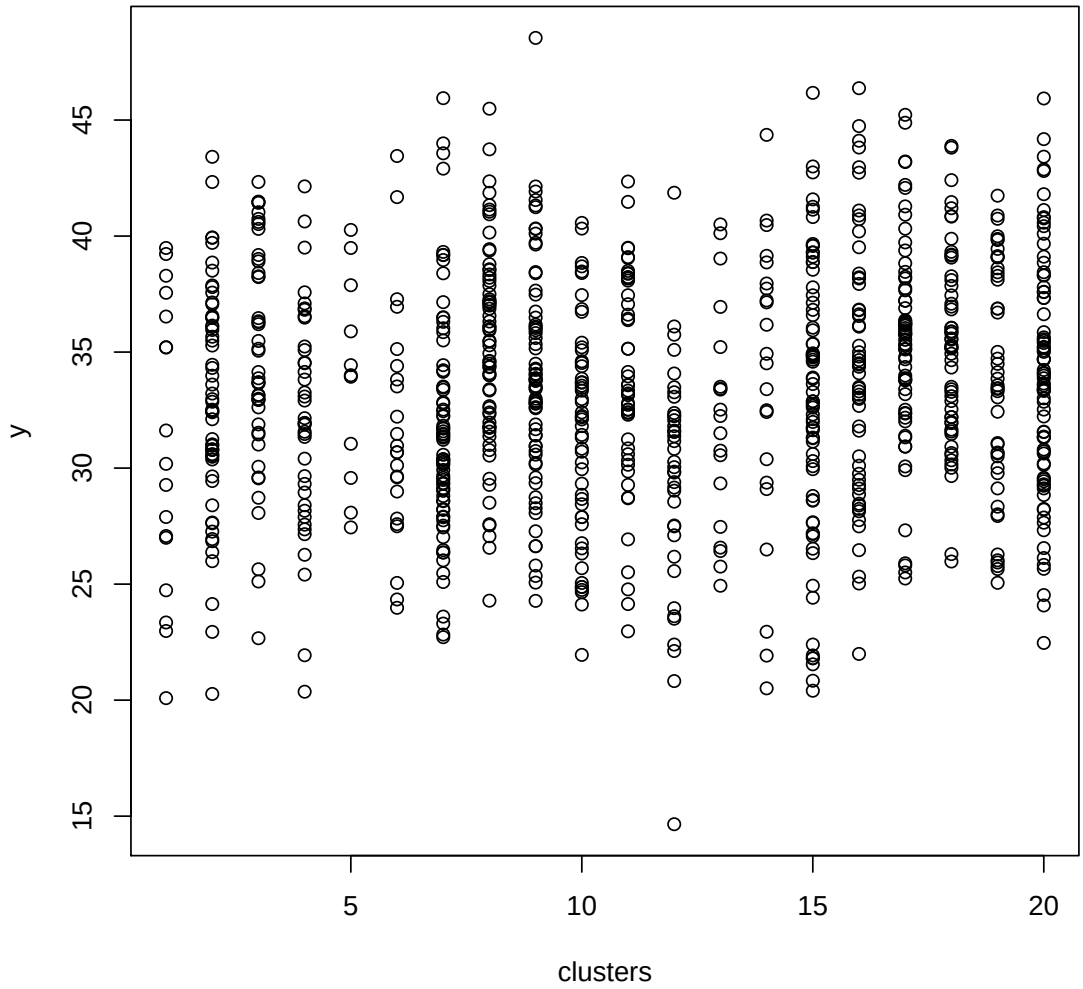
$$\hat{\bar{y}}_r = \sum_{i=1}^{4} M_i \bar{y}_i / \sum_{i=1}^{4} M_i$$

ii) The estimate of variance is

$$\frac{1}{\bar{M}^2}\left\{(1-\frac{4}{20})\frac{s_r^2}{4} + \frac{1}{4\times 20}\sum_{i=1}^{4} M_i^2(1-\frac{m_i}{M_i})^2\frac{s_i^2}{m_i}\right\}$$

where $\bar{M}$ is the population average cluster size if you assume that it is known. Otherwise it is sample average cluster size. Finally

$$s_r^2 = \sum_{i=1}^{4}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2/(4-1)$$


y
15
20
25
30
35
40
45
5
10
15
20
clusters