Stat 5201, November 9, 2018 Name _____

Fifty people took this exam. The high score was 92. Here is the distribution of the scores: 1 in the 90's, 9 in the 80's, 12 in the 70's, 10 in the 60's, 10 in the 50's, 4 in the 40's and 1 in the 20's.

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. A large company which has a large fleet of automobiles wanted to know the average milage of the cars they own. The cars are parked in a number of different lots. The lots are of different sizes and located in different areas. To find an estimate of the average milage of their cars they selected 20 lots at random from their list of the lots. Then the milage of each car in the selected lots was found.

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

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

ANS i) The target population is all cars owned by the company. The sampling frame is the list of all lots and the observational unit is the milage for an individual car.

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

$$\hat{R} = \frac{\sum_{i\in smp} Y_i}{\sum_{i\in smp} M_i} \quad \text{and} \quad V(\hat{R}) \widehat{=} \frac{1}{(\bar{M}_{smp})^2} \frac{1-f}{20} \sum_{i\in smp} (Y_i - \hat{R}M_i)^2/(n-1)$$

where $f = 20/N$, N is the total number of lots and $\bar{M}_{smp} = \sum_{i\in smp} M_i/20$.

2. Suppose a university with 5,000 students was interested in how often students used the university health services last semester. A random sample of 100 students yielded the following results:

number of visits	0	1	2	3	4
number of students	70	15	10	4	1

i) Find an estimate of the proportion of students who used the health center at least once during the last semester and give an estimate of the variance of your estimate.

ii) Estimate the total number of visits by students during the past semester and give an estimate of the variance of your estimate.

iii) Suppose it is know from another source that 3800 of the 5,000 students did not use the health center during the past semester. Use this information to find a new estimate of the total number of student visits to the health center during the past center. Also give an estimate of its variance.

ANS i) The estimate is 30/100 with estimated variance

$$(1 - 100/5000)(1/99)(3/10)(1 - 3/10)$$

ii) The estimate is

$$N\bar{y} = 5000 \frac{0\times 70 + 1\times 15 + 2\times 10 + 3\times 4 + 4\times 1}{100}$$

and since $\sum_{i\in smp} y_i = 51$ and $\sum_{i\in smp} y_i^2 = 107$ the estimated variance is

$$5000^2(1 - 100/5000)(1/100)\frac{107 - (51)^2/100}{99}$$

iii) This is a domain estimation problem where the domain is the the group of students who visited the visited the health center during the past semester and where the size of the domain is known. The new estimate is

$$1200\frac{1\times 15+2\times 10+3\times 4+4\times 1}{30}$$

with estimated variance

$$1200^2(1-30/1200)(1/30)\frac{107-(51)^2/100}{29}$$

3. A company was interested in how long it took to complete a job. They created a list of 3,000 recent jobs, 2,000 of which where in the Twin Cities and 1,000 of which where outside the Twin Cities. A stratified random sample was taken and the time, recorded in hours, was found. The results are given below

Location	Number in in sample	Average time	Sample variance
Twin Cities	70	8.6	4.2
outside	30	11.2	8.7

i) Find a 95% confidence interval for the average time needed to complete a job.

ii) Assuming the observed sample variances were the true population variances what would have been the optimal allocation of the original sample.

ANS i) The point estimate is

$$pt=\frac{2000}{3000}8.6+\frac{1000}{3000}11.2$$

with estimate variance

$$vr=(\frac{4}{9})\frac{4.2}{70}(1-70/2000)+(\frac{1}{9})\frac{8.7}{30}(1-30/1000)$$

Then $pt\pm 1.96\sqrt{vr}$ is an approximated 95% confidence interval.

ii) The optimal allocation for the first stratum would be

$$100\frac{2000\sqrt{4.2}}{2000\sqrt{4.2}+1000\sqrt{8.7}}=58.2$$

so a better stratification would have been 58 in the Twin cities and 42 outside.

4. A simple random sample without replacement was taken from a population. You are interested in estimating the the population total for some characteristic y. You note that for each unit in the sample the value of an auxiliary variable x was also observed and you know from past experience that for this population x and y will be positively correlated. A co-worker tells you that they know the population median of x. State a way you could use this information to improve the usual estimator? Be brief but specific.

ANS: One thing to do would be to form two strata: below and above the known median of x. If N is the population size and n the sample size then each unit in the sample has weight N/n. These weights can be renormalized so that both below and above the median of x the weights for the units in the sample sum to $N/2$.

5. A company with $N = 8,550$ employees wants to conduct a survey to see how many of them would be in favor of a proposed change in company policy. Let p be the proportion of employees who would be in favor of the new policy. How large must the sample size, n, be if the length of the resulting 95% conservative confidence interval is no longer than 0.06.

ANS From class

$$n = \frac{n_0}{1 + n_0/N} \quad \text{where} \quad n_0 = (1.96/(2 \times 0.06))^2 = 267$$

So

$$n = \frac{267}{1 + 267/8550} = 259$$

is the sample size.

6. Consider a population $y = (y_1, y_2, \ldots, y_N)$ where the labels carry some information about their y values. Specifically, y_i and y_j tend to have similar values when $|i - j|$ is small. It is believe that most of the y values tend to be small while just a few of them are bigger then some specified number $bd > 0$. Moreover it is believe that the units bigger than bd tend to come in groups of adjacent units. In order to find as many of this large units as possible an investigator devises the following sampling plan. Using simple random sampling without replacement one selects a sample of size n, say s. One then observes the y value for the units in the sample. For an $i \in s$ with $y_i <= bd$ nothing more is done. However if $y_i > bd$ units $i - 1$ and $i + 1$ are then observed. If either of them is greater than bd then the next adjacent unit is observed. One continues sampling in both directions until one finds a unit less than or equal to bd. We call such groups of successive units greater than bd networks. Networks can be of length one or two or so on. Note that if any of the original sample contains a network the final sample will contain more than n units.

Let i^* be a unit with $y_{i^*} > bd$ and suppose it belongs to a network of size $m*$. Find the probability that unit i^* does not appear in the final sample.

ANS The unit i^* will not be in the sample if and only if neither it nor any other members of its network appear in the sample. In this case the sample must only contain units from the $N - m*$ units not in its network. So the probability that it does not appear is

$$\binom{N - m^*}{n} / \binom{N}{n}$$