Stat 5201, November 18, 2016 Name _____

Seventy two people took the exam. The high score was 84 and the low score 28. The distribution of the scores were 3 in the 80's, 20 in the 70's, 23 in the 60's, 13 in the 50's, 11 in the 40's, 1 in the 30'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

1. There once was an eccentric millionaire who loved to collect vases of all types and ages for his apartment. Since his apartment was not very large, eventually he had to start putting some of the vases into boxes all of the same size. These boxes could contain one or two bigger vases or more smaller vases. Eventually he had to rent storage space to contain all of the boxes. On each box was the year it was filled with vases. When he died the storage space contained around 500 boxes of vases with dates ranging from 1975 to 2015.

His only heir decided to sell all the boxes to just one person. Before making their bid any interested person can spend one hour in the storage space. There is a chart showing where each box is located with its year noted. Your friend, who is an expert on vases, wants to make a bid. She believes that she can check about about 30 boxes in an hour and wants you to develop a sampling plan to help her determine how to select the sample of boxes to use in determining her bid. Describe your plan and briefly explain your reasons for choosing it.

ANS The chart is the sampling frame and for a variety of reasons we should stratify on years. The number and type of vases bought earlier could be different then those bought later and so on. The boxes are clusters of unequal sizes so we should probably use the ratio estimator within strata. If some of the boxes have lots of small vases you might have to just select a sample of vases from the box.

2. In a given region a survey of the $N = 4,450$ assistant nurses is to be conducted to see if any bullying had occurred in their departments. Let p the proportion of departments that have experienced bullying. How large must our sample size be if we want length of our 95%conservative confidence interval to be 0.04

ANS From class

$$n = \frac{n_0}{1 + n_0/N} \quad \text{where} \quad n_0 = (1.96/(2 \times 0.04))^2 = 600$$

So

$$n = \frac{600}{1 + 600/4450} = 528.7$$

or we take 529.

3. A population consists of three strata of sizes 600, 400 and 1000. 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 $n = 200$ taken from the population yielded these results.

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2
1	600	55	25.5	11.6
2	400	35	40.3	6.6
3	1000	110	49.3	16.5

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.

i) Use this information to calculate an approximate 95% confidence interval for the population mean.

ii) Suppose now that the N_h's were know and the observed strata sample variances happened to be your prior guess for the strata variances. How would you have selected the sample?

ANS i) 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.

ii) In this case the optimal allocation would be

$$n_i = n * \frac{N_i \sqrt{s_i^2}}{\sum_{h=1}^3 N_h \sqrt{s_h^2}}$$

4. A company with 1,000 employees wanted to know the total number of days in a given week employees used an on site walking path. A random sample of 200 employees yielded the results in the table below.

number of times used	0	1	2	3
number of employees	80	50	40	30

i) Find an estimate of the proportion of employees who used the walking path at least once during the given week. Give an estimate of the variance of your estimate.

ii) Estimate the total number of times the walking path was used in the given week and give an estimate of the variance of your estimate.

iii) Suppose it is know from another source that 600 of the 1,000 employees did not use the walking path in the given week. Use this information to find a new estimate of the total times the walking path was used in the given week.

ANS i) The estimate is 120/200=3/5 with estimated variance

$$(1 - 200/1000)(1/199)(3/5)(1 - 3/5)$$

ii) The estimate is

$$N\bar{y} = 1000 \frac{0 \times 80 + 1 \times 60 + 2 \times 40 + 3 \times 30}{200}$$

and since $\sum_{i \in smp} y_i = 220$ and $\sum_{i \in smp} y_i^2 = 480$ the estimated variance is

$$1000^2 (1 - 200/1000)(1/199) \frac{480 - (220)^2/200}{199}$$

iii) This is a domain estimation problem where the domain is the total number of employees who walked at least once and the size of the domain is known. The new estimate is

$$400\frac{1\times 50+2\times 40+3\times 30}{120}$$

with estimated variance

$$400^2(1-120/400)\frac{480-(220)^2/120}{119}$$

5. Consider a population with N clusters each of size two. Each clusters consists of two men or a man and a woman or two women. For $i=0,1$ and 2 let N_i be the number of clusters which contain exactly i women. A scientist will use simple random sampling without replacement to take a random sample of size n clusters. They want a sample that will contain more women the men.

i) Given the selected clusters one sampling plan would be to observe both individuals in th selected clusters. Find the probability that for any given selected cluster we see exactly i women.

ii) An alternative sampling plan would be to select a person at random from a selected cluster. If it is a man then observe the other person as well but if it is a woman we do not observe the other person in the cluster. So under this plan there are three possible outcomes, MM, MW and W for a randomly selected cluster. Find their probabilities.

iii) Find conditions on the N_i's for when the second plan will give more probability to the event that we see more women in a sampled cluster than men.

ANS i) Let A_i be the event that there are i women in a select cluster. Then $P(A_i)=N_i/N$. Then

$$P(A_2)>P(A_0)\Longleftrightarrow N_2>N_0$$

ii) Now there are three possible outcomes for each selected cluster, MM, MW and W with probabilities

$$P(MM)=N_0/N,\quad P(MW)=(N_1/N)(1/2)\text{ and}P(W)=(N_1/N)(1/2)+N_2/N$$

iii) For the first sampling plan this happens when $N_2>N_0$. while for the second it happens when

$$\frac{1}{N}(N_1/2+N2)>\frac{N_0}{N}\Longleftrightarrow(N-N_0-N_2)/2>N_0\Longleftrightarrow 3N_0+N_2<N$$

So if N_1 is large this can increase the probability of observing more women even when $N_2>N_0$.