Stat 5201, November 17, 2017 Name _____

Fifty nine people took this test and the high score was 100 and the low score was 0. There were 3 in the 80's, 6 in the 70's, 8 in the 60's, 15 in the 50's, 10 in the 40's, 7 in the 30's, 7 in the 20's and 1 in the 10'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 celebrity was interested in the proportion of articles about him available on the Internet that contain mistakes and false information. When he googled his name he found hundreds of articles, too many to read. Can you recommend a sampling plan to select a sample of articles for him to read which will allow you to estimate the proportion of articles that contain mistakes. Can you attach a design weight to an unit in your sample?

**Ans** The population and sampling frame are just the articles that appear when you google his name. Here is one idea. Just select one article at random from each set of five as you move through the list produced by google. The weight attached to each sampled unit is 5 since we will sample 1/5 of the population unless the last group has $k < 5$ articles and then the last sampled unit will get weight $k$.

This is probably easier and better than taking a random sample from the entire list since the order things appear is determined by google in ways that we do not know. Also one could stratify using info about the type of article but this is harder to do because we need to look at all the articles in the list before they are read.

2. Consider a population consisting of 500 distinct units and suppose we take a simple random sample **with replacement** of size 10. Give an expression for the probability that no unit appears twice in the sample.

**Ans** If $N$ is the population size and $n$ the sample size the probability is

$$\frac{N}{N}\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-(n-1)}{N}$$

3. A corporation that owns 950 restaurants wanted to know the total labor costs for a given month. Let $y_i$ denote the labor costs and $x_i$ the total sales of the $i$th restaurant. They took a simple random sample of 30 restaurants and found that $\bar{x}_s = 150,000$ and $\bar{y}_s = 30,000$. Suppose the total sales for that month were 160,000,000.

i) Find a point estimate for the total labor costs for that month.

ii) Give an expression for the 95% confidence interval for this quantity.

**Ans** i) Should use the ratio estimate with $\hat{R} = 30,000/150,000 = 0.2$ Then the estimate total labor costs for the month is $160,000,000/5 = 32,000,000$.

ii) The estimated variance of our estimator is

$$950^2((1-30/950)/30)\sum_{i\in smp}(y_i - \hat{R}x_i)^2/(n-1) = \hat{\sigma}^2$$

and the interval estimate is $32,000,000 \pm 1.96\,\hat{\sigma}$

4. Suppose we have a population with four strata. Let $N_h$ be the size of the $h$th stratum and let $\sigma_h^2$ be our prior guess for its variance. Let $c_h$ be the cost of observing a unit from the $h$th stratum. What is the optimal allocate of $n$ observations if we will do simple random sampling without replacement in each stratum? How much will our sample cost if we want the variance of our estimate of the population mean to be approximately twenty.

**Ans** Optimal allocation, is given by $n_h \propto (N_h\sigma_h)/\sqrt{c_h}$. Here is the formula given in class that relates $c$, the total cost, with the variance, $V$,

$$c = \frac{\left(\sum_h W_h\sigma_h\sqrt{c_y}\right)^2}{V + \sum_h W_h^2\sigma_h^2/N_h}$$

where $W_h = N_h/\sum_j N_j$.

5. A company has six different branches in a larger metropolitan area. They are interested in learning something about how much time it takes to complete a typical transaction. They selected two branches at random and then took a random sample of transactions within the selected branches from the last month. They then determined the number of hours needed to complete each transaction. The data is given below where $M_i$ is the number of transactions and $n_i$ is the sample size.

| Branch | $M_i$ | $n_i$ | $\bar{y}_i$ | $s_i^2$ |
|---|---|---|---|---|
| 1 | 400 | 40 | 33.8 | 12.6 |
| 2 | 600 | 30 | 15.6 | 9.7 |

The number of transactions for the four branches not in the sample were 350, 440, 560 and 640 respectively.

i) Estimate the mean time to complete a transaction for all of the company's transactions for last month.

ii) Give the estimate of variance for your estimator in part i). Your answer should be in the form of an expression that can be computed using a calculator.

**Ans** Here the branches are clusters and this is a two-stage cluster sample where $n = 2$ and $N = 6$. We use the ratio estimator because of the unequal cluster sizes. You need to use the formulas on page 186 of the text.

i)

$$\hat{\bar{y}}_r = \frac{400 \times 33.8 + 600 \times 15.6}{400 + 600}$$

ii) Note the estimate of variance is

$$\frac{1}{\bar{M}^2}\frac{2}{3}\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{2} M_i^2(1 - \frac{n_i}{M_i})\frac{s_i^2}{n_i}$$

where

$$\bar{M} = (400 + 600 + 350 + 440 + 560 + 640)/6$$

and

$$s_r^2 = 400^2(33.8 - \hat{\bar{y}}_r)^2 + 600^2(15.6 - \hat{\bar{y}}_r)^2$$

6. Consider a population which consists of two strata. In addition each person in the population is classified as either male or female. A stratified simple random sample was taken from the population. Information about the population and sample are given in the following table. Note $N_h$ and $n_h$ denote the the stratum size and sample size respectively. In the table the next three items give the number of females in the sample in each stratum, along with their sample means and variances of the $y$ variable of interest.

| Stratum | $N_h$ | $n_h$ | $n_{hw}$ | $\bar{y}_{hw}$ | $s^2_{hw}$ |
|---|---|---|---|---|---|
| 1 | 2,000 | 100 | 55 | 35.6 | 14.6 |
| 2 | 2,500 | 80 | 35 | 41.3 | 9.7 |

i) Suppose it is known that there are 1,200 females in the first stratum. Give the usual 95% confidence interval for the mean of the females in the first stratum.

ii) Redo part i) when the number of females in the first stratum is not known.

iii) Give an expression for the estimate of variance for the usual estimate of the **population total** for all the women in the population when the number of females in each stratum is unknown. Describe briefly what additional information, if any, you need to compute your estimate.

**Ans** i) This is a domain estimation problem, i.e. women in the first stratum where the domain size is known.

$$35.6 \pm 1.96\left( (1 - \frac{55}{1200})\frac{14.6}{55} \right)^{1/2}$$

ii) Since the domain size is not known we replace 55/1200 with 100/2000 in the fpc in the above formula.

iii) When estimating a domain total we do not need to know the size of the domain but our estimate of variance is not based on the $s^2_{hw}$ given above but uses the $y$ values for the women and a 0 value for all the men in the sample. If we have this information then we can compute our estimate of variance in each stratum and then add them up.