Statistics 5201, November 8, 2019 Name _____ Forty one people took the exam. Here is the distribution of the scores. There were 7 in the 90's, 5 in the 80's, 10 in the 70's, 6 in the 60's, 2 in the 50's, 3 in the 40's, 5 in the 30's, 2 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. In a city with 14,828 households a social scientist was interested in in the number of cars owned by households. A srs without replacement of size 290 yielded the following results

	Owned		Rented	
	Yes	No	Yes	No
≥ 2 cars	141	6	109	34

i) Based on this sample estimate the proportion of renting households in the city that have at least 2 cars. Give the standard error of your estimate.

ii) If the number of renting households in the city is known to be 7526 use this information and the sample to give an estimate of the total number of renting households in the city that own at least 2 cars. Give the standard error, or estimated variance, of this estimate.

ANS i) The estimate is

$$\frac{109}{290}$$

with standard error

$$\sqrt{\frac{1}{289}(1-\frac{290}{14828})\frac{109}{290}\frac{181}{290}}$$

ii) This is a domain estimation problem where the size of the domain is known. Since we are estimating the domain total and the size of the domain is known we get our estimate of variance by conditioning on the number of units in in the sample that fell in the domain, $109+34=143$.

The estimate is

$$7526\frac{109}{143}$$

with standard error

$$7526\sqrt{\frac{1}{142}(1-\frac{143}{7526})\frac{109}{143}\frac{34}{143}}$$

2. In a population it was believed that y_i, the variable of interest, is approximately proportional to an auxiliary variable x_i.

i) Under this assumption what is a reasonable estimate of the population mean of y if in a random sample of size 40 from the population $\bar{y}_{smp} = 87.4$ and $\bar{x}_{smp} = 44.2$ were observed and the population mean of x is 61.3?

ii) What additional information do you need to find an approximate 95% confidence interval for the population mean of y.

Ans Should use the ratio estimate which is $\hat{R} \times 61.3$. where $\hat{R} = 87.4/44.2$.

You need to know $\sum_{i \in smp}(y_i - \hat{R}x_i)^2$ and the population size N unless $40/N$ is negligible.

3. To study the oral health of 200 children in a village dentist A selects a simple random sample (without replacement) of 20 children and counts the number of decayed teeth for each child with the following results

# of cavities per child	0	1	2	3	4	5	6	7	8	9	10
# of children	8	4	2	2	1	1	0	0	0	1	1

i) Using A's results estimate the total number of decayed teeth in the village children and give an estimate of its variance.

ii) Let D be the domain that contains all the children who have at least one decayed tooth. A second dentist, B, examines all the children in the village recording merely those who have at least one decayed tooth. B finds 60 children with no decayed teeth. Use the results of both dentists to get a new estimate of the total number of decayed teeth in the village children and give an estimate of its variance. Which of the two estimators would your prefer and why?

ANS i) The estimate of the total is

$$\hat{Y} = N\bar{y} = 200\frac{42}{20} = 420$$

Note that $\sum_{i \in s} y_i^2 = 252$ so the estimated standard deviation of the estimate is

$$200\sqrt{1 - \frac{20}{200}}\frac{1}{\sqrt{20}}\sqrt{\frac{252 - (42)^2/20}{19}}$$

ii) Since the size of the domain is known our estimate is

$$140\frac{42}{12}$$

with estimated standard deviation of

$$140\sqrt{1 - \frac{12}{140}}\frac{1}{\sqrt{12}}\sqrt{\frac{252 - (42)^2/20}{11}}$$

We would expect the estimate in the second part to be better since it using more information.

4. Consider a stratified population consisting of 3 strata where there are good prior guesses for the true strata variances. In addition the cost of sampling units from each strata is also known approximately. The information is given in the table below.

strata	1	2	3
size	400	600	200
est var	30	20	40
cost	\$6	\$10	\$12

i) Find the optimal allocation of the strata sample sizes if you can spend \$1,000.

ii) Find the minimum total cost necessary to obtain an estimate whose estimated variance will be the value v.

ANS i) Let N_h, σ_h, c_h and n_h be the size, standard deviation, cost and sample size of stratum h. Then for some constant λ the optimal allocation must satisfy

$$n_h = \lambda \frac{N_h \sigma_h}{\sqrt{c_h}}$$

where λ is found from the equation

$$1000 = \sum_h c_h n_h = \lambda \sum_h \sqrt{c_h} N_h \sigma_h$$

ii) Let $W_h = N_h/(N_1 + N_2 + N_3)$ then the minimum total cost is given by

$$c = (\sum_h W_h \sigma_h \sqrt{c_h})^2 / (v + \sum_h W_h^2 \sigma_h^2 / N_h)$$

5. A population consists of a few strata, however this information is not contained in the sampling frame. But the strata sizes, the N_h's are know from other sources. Also the stratum membership can be determined for units in the sample. Suppose a simple random sample without replacement of size n, where n is large, is taken from the population and stratum membership and the y value for each unit in the sample is observed. Let n_h be the number of units that fall in stratum h. Assume that all the observed n_h's are greater than 30.

i) Give a sensible point estimate of the population mean of y.

ii) Give an estimate of the variance for your estimate in the first part. Briefly justify your answer.

ANS This is a post-stratification problem. Note that the n_h's are random variables. Let N_h be the true but known stratum size since the n_h's are random variables. The usual point estimator is

$$\bar{y}_{post} = \sum_h \frac{N_h}{N} \bar{y}_h$$

Since for each h we expect $n_h/N_h \doteq n/N$ we can use as our estimate of variance the estimate of variance under proportional allocation. Hence an approximate estimate of variance is

$$\sum_h (1 - \frac{n_h}{N_h})(\frac{N_h}{N})^2 \frac{s_h^2}{n_h} \doteq (1 - \frac{n}{N}) \sum_h \frac{N_h}{N} \frac{s_h^2}{n}$$

For more discussion see section 4.4 of the text.