STAT 521: Midterm Exam Solution Name:

Problem 1: (20 pts)

A residential area has 5000 private houses. We want to estimate the proportion of houses with more than four persons living in them. The estimator is required to have standard error not exceeding 0.01.

1. (10pt) If we are going to use SRS design to select a sample of size n, how large a sample is needed to meet the accuracy requirement?

   Wish to find the minimum integer n satisfying

   $$\sqrt{\frac{1}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}P(1-P)} \leq 0.01$$

   The maximum of $P(1-P)$ is achieved at $P=0.5$. Thus, with $N=5000$, we can solve

   $$\left(\frac{1}{n}-\frac{1}{5000}\right)\frac{5000}{4999}0.25 \leq (0.01)^2$$

   The minimum sample size is 1667.

2. (10pt) How does your answer to the above question change if we use an extra information that the true proportion will lie in the range 0.1 to 0.3?

We can apply the same inequality at $P=0.3$ to obtain $n \geq 1479$. Thus, the answer is 1479.

Problem 2: (20 pts)

Suppose that the elementary schools in a city are grouped into 30 school districts, with each school district containing four schools. Suppose that a simple one-stage cluster sample of three school districts is taken for the purpose of estimating the number school children in the city who are color-blind, and that the accompanying data are obtained from this sample.

Sample		No. of	No. of
School District	School	Children	Color-blind children
1	1	130	2
	2	150	3
	3	160	3
	4	120	5
2	1	110	2
	2	120	4
	3	100	0
	4	120	1
3	1	89	4
	2	130	2
	3	100	0
	4	150	2

1. Estimate and obtain a 95% confidence interval for the total number of color-blind children in the city.

   The sampling design is a simple random cluster sampling of size $n_I = 3$ from the clustered population of size $N_I = 30$. From the cluster sample, we observe $t_1 = 2+3+3+5 = 13$, $t_2 = 2+4+0+1 = 7$, and $t_3 = 4+2+0+2 = 8$. Thus, the estimated total is

$$\hat{T} = \frac{N_I}{n_I}\sum_{i=1}^{3} t_i = 10 \cdot (13+7+8) = 280.$$

   The standard error of $\hat{T}$ is

$$\widehat{SE}(\hat{T}) = \sqrt{\frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)s_t^2} = \sqrt{\frac{30^2}{3}\left(1-\frac{3}{30}\right)10.33} = 52.82$$

   Thus, 95% CI for T is

$$\left[\hat{T} - 1.96\widehat{SE}(\hat{T}), \hat{T} + 1.96\widehat{SE}(\hat{T})\right] = [176.47, 383.53].$$

2. Estimate and obtain a 95% confidence interval for the proportion of color-blind children in the city.

   The parameter of interest can be written as

$$P = \frac{\sum_{i=1}^{N_I} t_i}{\sum_{i=1}^{N_I} x_i},$$

   where x_i is the number of children in the i-th school district. We can use

$$\hat{P} = \frac{\sum_{i=1}^{3} t_i}{\sum_{i=1}^{3} x_i} = \frac{13+7+8}{560+450+469} = 0.0189$$

   to estimate P. The standard error of $\hat{P}$ is

$$\widehat{SE}(\hat{P}) = \sqrt{\frac{1}{n_I}\left(1-\frac{n_I}{N_I}\right)\frac{s_d^2}{\bar{x}^2}} = \sqrt{\frac{1}{3}\left(1-\frac{3}{30}\right)\frac{4.416169}{493^2}} = 0.00233$$

   where

$$s_d^2 = \frac{1}{3-1}\sum_{i=1}^{3}(t_i - \hat{P}x_i)^2 = 4.416169.$$

   Thus, 95% CI for P is

$$\left[\hat{P} - 1.96\widehat{SE}(\hat{P}), \hat{P} + 1.96\widehat{SE}(\hat{P})\right] = [0.0144, 0.0235].$$

Problem 3: (30 pts)

A single stage cluster sampling was used to estimate the total number of children in a household in a given finite population of households. Clusters were created by forming M adjacent households and a simple random sampling of clusters are used to select samples from the population. The following is the ANOVA table obtained from the sampled households. (The value of M can be computed from the ANOVA table.)

ANOVA table

Source	d.f.	Sum of Squares
Between Clusters	100	3,000
Within Clusters	909	9,090
Total	1,009	12,090

(a) What is your estimate for the intracluster correlation coefficient?

Mean sum of squares:

$$\begin{aligned} S_b^2 &= 3000/100 = 30 \\ S_w^2 &= 9090/909 = 10. \end{aligned}$$

Also,

$$\hat{\sigma}_y^2 = \frac{1}{M} S_b^2 + \left(1 - \frac{1}{M}\right) S_w^2 = 12.$$

Thus,

$$\hat{\rho} = 1 - \frac{S_w^2}{S_y^2} = 1 - \frac{10}{12} = 0.1667.$$

(b) Estimate the variance of the estimated average number of children in a household in the population. (Ignore the finite population correction term.)

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{n_I M} S_b^2 = \frac{1}{101 \cdot 10} \times 30 = 0.0297.$$

(c) What is the effective sample size of this sampling design? Explain it.

Design effect is $1 + (M-1)\hat{\rho} = 1 + 9/6 = 2.5$. Thus, effective sample size is $n^* = n/deff = 1010/2.5 = 404$. The above cluster sampling has the same efficiency of the SRS design of size $n^* = 404$ in terms of variance.

Problem 4: (30 pts)

Among the 7,500 employees of a company, we wish to know the proportion P of them that owns at least one vehicle. For each individual in the sampling frame, we have the value of his income. We then decide to construct three strata in the population: individuals with low income (stratum 1), with medium income (stratum 2), and with high income (stratum 3). Within stratum h, simple random sampling without replacement of size n_h is performly independently from the population of size N_h. Let p_h be the estimated proportion of individuals in stratum h owing at least one vehicle. The results are given in the following table.

	$h=1$	$h=2$	$h=3$
N_h	3,500	2,000	2,000
n_h	500	300	200
p_h	0.13	0.45	0.50

(a) Find the unbiased estimate $\hat{P}$ of P.

$$\begin{aligned}\hat{P} &= \sum_{h=1}^{3} \frac{N_h}{N} p_h \\ &= \frac{1}{7500}(3500 \cdot 0.13 + 2000 \cdot 0.45 + 2000 \cdot 0.5) \\ &= 0.314\end{aligned}$$

(b) Compute the estimated variance of $\hat{P}$ in (a).

$$\begin{aligned}\hat{V}(\hat{P}) &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} p_h(1 - p_h) \\ &\doteq 1.73 \times 10^{-4}\end{aligned}$$

(c) What is the optimal allocation for the stratum sample sizes under $n_1 + n_2 + n_3 = 1,000$? (You may assume that the costs are the same for each stratum.) Compute the estimated variance under the optimal allocation and compare it with the variance in (b).

Optimal sample size allocation:

$$n_h^* \propto N_h S_h = N_h \sqrt{P_h(1-P_h)}$$

Thus,

$$\begin{aligned} n_1^* &= 371 \\ n_2^* &= 314 \\ n_3^* &= 315 \end{aligned}$$

and

$$\begin{aligned} \hat{V}(\hat{P}|n_h^*) &= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{1}{n_h^*} \left(1 - \frac{n_h^*}{N_h}\right) \frac{n_h^*}{n_h^* - 1} p_h(1-p_h) \\ &\doteq 1.54 \times 10^{-4}. \end{aligned}$$

Thus, it is smaller than the variance in (b).