Carleton University
School of Mathematics and Statistics
Sampling Methodology: STAT 3507 – Winter 2021

MIDTERM Examination
March 4, 2021

Name: _____

Student id #: _____

Answer all Questions

1. [**17 points**] As part of an AIDS education program, 150 intravenous drug users seronegative for HIV (Human Immunodeficiency Virus) at a first screening were given instructions on sterilizing their needles with bleach and practicing "safe sex." One year after the program's inception, a sample of 30 of these subjects was taken by numbering the participants from 1 to 150 and by taking all subjects whose numbers are divisible by 5 (e.g., 5, 10, 15, etc.).

    a. [1] State the target population of this study.

    b. [1] Identify the element (observational unit) of the population.

    c. [1] What is the sampling unit?

    d. [1] Propose a sampling frame for this study.

    e. [1] State the survey (study) variable of interest in this study.

    f. [3] Name any three potential sources of non-sampling errors that could occur in this study. Suggest how each of these errors may be reduced.

    g. [3] For this study, explain why a sample survey may be preferred to a census for this survey.

    h. [1] Estimate the probability of an individual selected in the sample?

    i. [1] If subjects 1, 4, 5, 10, 36, and 80 are seropositive for HIV, what is the proportion of seroconverted subjects in the population?

    j. [2] If subjects 1, 4, 5, 10, 36, and 80 are seropositive for HIV, what is the proportion of seroconverted subjects in the sample? Is this an unbiased estimate of the population proportion? Explain.

k. [2] Construct a 95% confidence interval for the proportion of seroconverted subjects in the population.

2. [**4 marks**] The government of Canada is interested in assessing the impact of the recent decline in the Canadian dollar on the economies of the country as well as the provinces and territories. Specifically, the impact on the manufacturing establishments. For the sampling design, a sample of establishments was to be drawn across the country using a simple random sampling (SRS). As the sampling methodologist (statistician) on the panel, explain to the members of the panel, why the use of a stratified random sampling may be more appropriate than the SRS approach.

3. [**14 points**] A sample survey of households in a community containing 2,100 households is to be conducted for the purpose of determining the total number of persons over 18 years of age in the community who have one or more permanent teeth (other than third molars) missing. Since this variable is thought to be correlated with age and income, the strata shown in the accompanying table are formed by using available population data. A stratified random sample of 150 families was selected and the summary of the data is shown in the table below:

	Stratum			
Variable	**1**	**2**	**3**	**4**
Annual family income (x$1000)				
Sample Mean	15	7	15	8
Sample Standard deviation	5	3	3	2
Sample size	50	50	10	40
No. of families	400	700	200	800

a. [2] Estimate the mean ($\overline{Y}$) annual income for the households in the community.

b. [3] Construct a 95% confidence interval for $\overline{Y}$.

c. [1] Estimate the difference in mean annual incomes between households in stratum 1 and stratum 4.

d. [4] Is the mean annual income of households in stratum 1 significantly different from that in stratum 4? Justify your answer.

e. [2] Determine the number of families to be taken from each stratum if proportional allocation is used.

f. [2] Suppose the number of persons over 18 years of age having missing teeth in a family is highly correlated with family income. Is stratified random sampling with proportional allocation likely to yield an estimate having lower variance than that obtained from a simple random sample of the same number of households? Explain.

4. [**10 points**] In a large population survey, 15,000 persons were screened with chest radiographs. Physicians noted possible pulmonary artery enlargement in 250 of these patients. This enlargement was confirmed by a second reading in 205 of these 250 persons. A sample of 185 of the 14,750 chest radiographs in which no enlargement of the pulmonary artery was noted yielded 15 radiographs that were actually positive for pulmonary artery enlargement. These results are summarized in the table below:

	Strata	
	I	II
Sample size	250	185
No. of Persons with Pulmonary artery Enlargement	205	15
Population size	250	14750

a. [2] Estimate the proportion (P) of persons with pulmonary artery enlargement in the population.

b. [4] Obtain a 95% confidence interval of P.

c. [4] Determine the number of persons to select for the next screening if optimal allocation (use the current data where applicable) with $n = 410$.

Formulas

Simple Random Sampling

Total and Mean

$\hat{Y} = N\bar{y}$; $\hat{V}(\hat{Y}) = N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n}$

$\bar{y} = \hat{\bar{Y}}$ $\hat{V}(\bar{y}) = \hat{V}(\hat{\bar{Y}}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$

Proportions

Let $y_i = \begin{cases} 1 & \textit{for a given characteristic} \\ 0 & \textit{otherwise} \end{cases}$

$P = \frac{1}{N}\sum_{i=1}^{N} y_i$; $\sigma^2 = P(1-P)$; $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $\hat{V}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n-1}\left(\frac{N-n}{N}\right)$

Sample size determination

$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$; $D = \frac{B^2}{4}$, *for* $\bar{Y}$ *and* P; $D = \frac{B^2}{4N^2}$, *for* Y

Stratified Random Sampling

Total and mean

$\hat{Y}_{st} = N\bar{y}_{st}$; $\hat{V}(\hat{Y}_{st}) = N^2\left[\frac{1}{N^2}\sum_{i=1}^{H} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}\right]$

$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{H} N_i\bar{y}_i$; $\hat{V}(\bar{y}_{st}) = \left[\frac{1}{N^2}\sum_{i=1}^{H} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}\right]$

Proportions

$\hat{P}_{st} = \frac{1}{N}\sum_{i=1}^{H} N_i\hat{P}_i$; $\hat{V}(\hat{P}_{st}) = \frac{1}{N^2}\sum_{i=1}^{H} N_i^2\left(\frac{\hat{P}_i(1-\hat{P}_i)}{n_i - 1}\right)\left(\frac{N_i - n_i}{N_i}\right)$

Sample size determination

$$n \approx \frac{\sum_{i=1}^{H} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{H} N_i \sigma_i^2}; \qquad a_i = \frac{n_i}{n} \quad ; \quad D = \begin{cases} B^2/4 & for\ \bar{Y}\ or\ P \\ B^2/4N^2 & for\ Y \end{cases}$$

Allocation of the sample

An approximate allocation that minimizes the cost for a fixed value of $V(\bar{y}_{st})$ or minimizes $V(\bar{y}_{st})$ for a fixed cost is given by

$$n_i \approx n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{j=1}^{H} N_j\sigma_j/\sqrt{c_j}}\right)$$

For a fixed $V(\bar{y}_{st}) = \frac{B^2}{4} = D$ then $n \approx \frac{\left(\sum_{i=1}^{H} N_i\sigma_i/\sqrt{c_i}\right)\left(\sum_{i=1}^{H} N_i\sigma_i\sqrt{c_i}\right)}{N^2 D + \sum_{i=1}^{H} N_i\sigma_i^2}$

Neyman Allocation $\quad n_i \approx n\left(\frac{N_i\sigma_i}{\sum_{j=1}^{H} N_j\sigma_j}\right)$ and $n \approx \frac{\left(\sum_{i=1}^{H} N_i\sigma_i\right)^2}{N^2 D + \sum_{i=1}^{H} N_i\sigma_i^2}$

Proportional Allocation $\quad n_i \approx n\left(\frac{N_i}{N}\right)$ and $n \approx \frac{N\left(\sum_{i=1}^{L} N_i\sigma_i^2\right)}{N^2 D + \sum_{i=1}^{L} N_i\sigma_i^2}$