UNIVERSITY OF TORONTO
Faculty of Arts and Science
STA304-**L0101** - Term Test I October 16, 2019

Solutions

INSTRUCTIONS:

- **This test is closed book; it is worth 40 points and you have 50 minutes to complete it**
- **Aids allowed: Non-programmable Calculator and One-sided handwritten 8.5x11 inches aid sheet**
- **No electronic devices with possible internet access, such as phone, tablets and computers are allowed**
- **Work scattered all over the page that cannot be understood will not earn full marks**
- **Please round your final answers to 3 decimal places.**

MARKING SCHEME:

Question	1	2	3	4	Total
Marks	14	10	11	5	40

Q 1. (14 marks) An opinion poll on America's health concern was conducted by Gallup Organization between October 3-5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US. The result was based on telephone interviews of XXXXX adults randomly selected.

(a) (8 marks) Identify the followings items

1mk each ur (viii)

(i) What is the response variable in this survey?

$Y = \begin{cases} 1 \text{ (Yes) AIDS is considered as the most urgent health problem} \\ 0 \text{ (No)} \end{cases}$ ~~Bernoulli~~ so Bernoulli random variable only two yes/no category of response

(ii) What is the parameter of interest?

p = proportion of adults who consider AIDS as the most urgent health problem of the US.

(iii) What was the target population?

US Population the survey results are generilized to all US Population.

(iv) What was the sampled population?

Adults whom could be reached by telephone

(v) What was the sampling unit?

an adult. who could be reached by telephone

(vi) How the survey was conducted?

By telephone interview

(vii) How was the sample selected ?

Random selection from telephone list

(viii) What is the sample statistic?

$\hat{p}$ = 2~~8~~9% proportion of adults who consider AIDS is the most urgent health problem of the sample

(b) (6 marks) In addtion, the report stated that the poll has a margin of ± 3%, 19 times out of 20.

(i) (1 mark) Explain what "margin of ± 3%, 19 times out of 20" means

- 95% margin of error is 0.03 (accepted)
- or we are ~~conf~~ 95% confident that the bound ~~on~~ by for which $\hat{p}$ is expeted to differ from p is 0.03.

(ii) (3 marks) Give an 95% confidence interval for the population proportion of all adults that consider AIDS is the most urgent health problem of the US. Give a practical interpretation of the resulting confidence interval.

$$(0.29-0.03,\ 0.29+0.03) = \boxed{(0.26,\ 0.32)} \quad 2 \text{ mks}$$

- The interval (0.26, 0.32) will contain the proportion of all aduts that consider AIDS is the most health problem ~~in~~ of the US, about 95% of the time 1 mk

(iii) (2 marks) Complet the survey by calculating the sample size that would be necessary to have a margin of error of 3%.

$$n = \frac{N p(1-p)}{(N-1)D + p(1-p)} \quad \text{where } D = \frac{0.03^2}{4}$$

① N unknow ⇒ N consider as large ⇒ $N \sim N-1$

$$\Rightarrow n \approx \frac{p(1-p)}{D + \frac{p(1-p)}{N}} \approx \frac{p(1-p)}{D}$$

② p unknown ⇒ we use $p=0.5$ which provide a conservative sample size

$$n = \frac{0.5 \times 0.5}{\frac{0.03^2}{4}} = 1111.11,$$ we select ⇒ $\boxed{n = 1112}$ 2 mks

Q 2. (10 marks) In this question, parts (a) and (b) are independent.

(a) (6 marks) In a large city school system with 20 elementary schools, the school board is considering the adoption of a new policy that would require elementary students to pass a test in order to be promoted to the next grade. The PTA wants to find out whether parents agree with this plan. Listed below ((i)-(vi)) are some of the ideas proposed for gathering data. Use the following list and answer the question below

(A) Simple Random Sample

(B) Stratified Random Sample

(C) Systemetic Sample

(D) Cluster Sample

(E) Multistage sample

(F) Convenience sample

(G) Voluntary sample

(H) Quota Sample

(I) Census

For each proposed idea below, indicate what kind of sampling strategy is involved.

(i) (1 mark) Put a big ad in the newspaper asking people to log their opinions on the PTA website.

G only those who see the ad will response.

(ii) (1 mark) Randomly select one of the elementary schools and contact every parent by phone.

D one school may not be typical of all.

(iii) (1 mark) Send a survey home with every student, and ask parents to fill it out and return it the next day.

I will have nonresponse bias

(iv) (1 mark) Randomly select 20 parents from each elementary school. Send them a survey, and follow up with a phone call if they do not return the survey within a week.

B elementary school are strata, from each of them we draw SRS

(v) (1 mark) Randomly select one class at each elementary school and contact each of those parent.

B Same as in (IV)

(vi) (1 mark) Go through the district's enrollment records, selecting every 40th parent. PTA volunteers will go to those home to interview the people chosen.

C

(b) (4 marks) An university administration is considering a variety of ways to sample students for a survey. For each of these proposed survey designs, identify the type of bias.

Please, write in the blank to the right with letter (A-D) corresponding to the correct answer.

(i) (2 marks) Publish an advertisement inviting students to visit a website and ansver questions.

(A) NonResponse bias

[✓ (B) voluntary response bias]

(C) Convenience response bias

(D) Bad sampling frame Bias

(ii) (2 marks) Set up a table in the student union and ask students to stop and answer a survey

(A) Undercoverage bias

(B) voluntary response bias

[✓ (C) Convenience response bias]

(D) Bad sampling frame Bias

Q 3. (11 marks) In a city of 72,500 people, a simple random sample without replacement of four households is selected from the 25,000 households in the population to estimate the average cost on food per household for a week. The first household in the sample had 4 people and spent a total of \$150 in food that week. The second household had 2 people and spent \$100. The third, with 4 people, spent \$200. The fourth, with 3 people, spent \$140.

(a) (3 marks) Identify the sampling units, the variable of interest, and any auxiliary information

Sampling units = households 1mk

Variable of interest = y = cost of food. 1 mk

Auxiliary variable = x = number of people in the household 1mk

(b) (2 marks) Suggest two types of estimators for estimating the mean expenditure per household for a week's food in the city. Summarize some properties of each estimator.

Simple Estimator $\hat{\mu} = \bar{y}$ (unbiased) 1 mk

Ratio Estimator $\hat{\mu}_r = \frac{\bar{y}}{\bar{x}} (\mu_x)$ biased Estimator 1 mk

or Regression, or difference Estimator both biased.

(c) (6 marks) The data obtained from the survey were recorded in variables y and x, analyzed in R and produced results presented below (see next page). Based on the R output, answer the following questions.

(i) (2 marks) Estimate mean expenditure using the simple estimator, and estimate the variance of the estimator.

From R output $\hat{\mu} = \bar{y} = \frac{590}{4} = \boxed{147.5}$ 1mk

$\hat{V}(\hat{\mu}) = \boxed{422.849}$ 1 mk

(ii) (2 marks) Estimate mean expenditure using the ratio estimator, and estimate the variance of the estimator.

From R output $\hat{\mu}_r = r \cdot \mu_x = 45.38 \times 2.9 = \boxed{131.6}$ 1mk

$\hat{V}(\hat{\mu}_r) = \boxed{119.6}$ 1 mk

(iii) (2 marks) Based on the data, which estimator appears preferable in this situation?

Based on the correlation of 0.867 or estimated variance, $\hat{\mu}_r$ is better than $\hat{\mu}$. 2 mks

Q 4. (5 marks) Suppose we choose a simple random sample without replacement of n units out of N units in the population of bernoulli random variables $\{0,1\}$ data values. Let $p = \frac{1}{N}\sum_{i=1}^{N} y_i$ be the population proportion, and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - p)^2$ denotes the population variance. Let $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{p})^2$ be the sample proportion and the sample variance, respectively.

(a) (2 marks) Show that $\widehat{p}$ is unbiased estimator of p.

$$E(\hat{p}) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum p = \frac{1}{n}(np) = p$$

$\hat{p}$ is unbiased for p.

(b) (1 mark) Using the result $\sum_{i=1}^{N}(y_i - \overline{y})^2 = \sum_{i=1}^{N} y_i^2 - N\overline{y}^2$, show that σ^2 can be writen as $\sigma^2 = p(1-p)$.

$$\sigma^2 = \frac{1}{N}\sum(y_i - \bar{y})^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - p)^2 = \frac{1}{N}\left[\sum y_i^2 - N\bar{y}^2\right]$$
$$= \frac{1}{N}\left[\sum y_i - Np^2\right] = \frac{1}{N}\left[Np - Np^2\right] = p - p^2 = \boxed{p(1-p)}$$

1 mk

(c) (1 mark) Using the result $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$, show that s^2 can be writen as $s^2 = \frac{n}{n-1}\widehat{p}(1-\widehat{p})$.

$$s^2 = \frac{1}{n-1}\left[\sum y_i^2 - n\bar{y}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i - n\bar{y}^2\right]$$
$$= \frac{1}{n-1}\left[n\hat{p} - n\hat{p}^2\right] = \frac{n}{n-1}\left[\hat{p} - \hat{p}^2\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

1 mk

(d) (1 mark) Is the sample variance s^2 unbiased estimator for σ^2?

- under SRS without replacement $E[s^2] = \frac{N}{N-1}\sigma^2$
  $\Rightarrow$ No: s^2 is biased estimator
- No (also accepted)

End !