Seat

Ni21

Please print in pen: STC 0020

Waterloo Student ID Number:

□□□□□□□□

WatIAM/Quest Login Userid:

□□□□□□□□

UNIVERSITY OF WATERLOO

Examination Test 2 Winter 2020 STAT 341

Times: Friday 2020-03-13 at 09:30 to 10:20

Duration: 50 minutes

Exam ID: 4463921

Sections: STAT 341 LEC 001

Instructors: Nathaniel Stevens

Special Materials

Candidates may bring only the listed aids.

- Calculator - Pink Tie

Instructions:

- You have 50 minutes to complete this test.
- This test consists of 6 questions and 8 pages (including this cover page).
- Pages 7 and 8 contain additional space for rough work. DO NOT use these pages for anything that you would like to have marked. For your convenience, they may be detached from the rest of the test.
- Numeric answers should be rounded to four decimal places (unless the answer is exact to fewer than four decimal places).
- Incorrect answers may receive partial credit if your work is shown. An incorrect answer with no work shown will receive 0 points.

Question	Points
Q1	7
Q2	5
Q3	6
Q4	6
Q5	4
Q6	4
Total	**32**

- Please identify yourself by signing here: ______________________

SOLUTIONS

1. [**7 points**] Consider the population attribute $a(\mathcal{P})$. Based on a random sample $\mathcal{S}$, the population attribute is estimated by $a(\mathcal{S})$ and the corresponding estimator is $\widetilde{a}(S)$

- 2 points for fully correct answer
- 1 point for partially correct answer
- 0 points for very incorrect answer

(a) [2 points] Show that

$$MSE[\widetilde{a}(\mathcal{S})] = Var[\widetilde{a}(\mathcal{S})] + Bias[\widetilde{a}(\mathcal{S})]^2$$

$$MSE[\tilde{a}(S)] = E[(\tilde{a}(S) - a(P))^2]$$

$$= E[((\tilde{a}(S) - E[\tilde{a}(S)]) + (E[\tilde{a}(S)] - a(P)))^2]$$

$$= E[(\tilde{a}(S) - E[\tilde{a}(S)])^2] + (E[\tilde{a}(S)] - a(P))^2 + 2E[(\tilde{a}(S) - E[\tilde{a}(S)])(E[\tilde{a}(S)] - a(P)$$

$$= Var[\tilde{a}(S)] + Bias[\tilde{a}(S)]^2 + 2(E[\tilde{a}(S)] - a(P))\overset{0}{E[\cancel{\tilde{a}(S) - E[\tilde{a}(S)]}]}$$

$$= Var[\tilde{a}(S)] + Bias[\tilde{a}(S)]^2$$

(b) [5 points] Consider estimating the mean of a population with values $\mathcal{P} = \{2, 3, 4, 5, 6\}$ based on a sample of size $n = 4$. The sampling design and sample attribute values for all possible samples are summarized in the table below.

$\mathcal{S}$	$P(\mathcal{S})$	$a(\mathcal{S}) = \overline{y}$
{2,3,4,5}	0.1	3.50
{2,3,4,6}	0.1	3.75
{2,3,5,6}	0.4	4.00
{2,4,5,6}	0.3	4.25
{3,4,5,6}	0.1	4.50

i. [2 points] Show that $E[\widetilde{a}(\mathcal{S})] = 4.05$

$$E[\tilde{a}(S)] = \sum_{S \in P_S} a(S)P(S) = (3.5)(0.1) + (3.75)(0.1) + (4.25)(0.4) + (4.25)(0.3) + (4.5)(0.1)$$

$$= 4.05 \checkmark\checkmark$$

For both of these, -1 if the approach was right but a calculation error was made

ii. [2 points] Show that $Var[\widetilde{a}(\mathcal{S})] = 0.0725$

$$Var[\tilde{a}(S)] = E[\tilde{a}(S)^2] - E[\tilde{a}(S)]^2 = \sum_{S \in P_S} a(S)^2 P(S) - \Big(\sum_{S \in P_S} a(S)P(S)\Big)^2$$

$$= 3.5^2(0.1) + 3.75^2(0.1) + 4^2(0.4) + 4.25^2(0.3) + 4.5^2(0.1) - 4.05^2$$

$$= 0.0725 \checkmark\checkmark$$

iii. [1 point] Calculate $MSE[\widetilde{a}(\mathcal{S})]$

$$MSE[\tilde{a}(S)] = Var[\tilde{a}(S)] + Bias[\tilde{a}(S)]^2$$

$$= 0.0725 + (4.05 - 4)^2$$

$$= 0.075 \checkmark$$

If this is wrong, they get 0.

2. **[5 points]** *Cluster sampling* is a probabilistic sampling mechanism that is applicable when a population $\mathcal{P}$ can be paritioned into H clusters (i.e., sub-populations) $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_H\}$ such that

$$\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_H \quad \text{and} \quad N = N_1 + N_2 + \cdots + N_H$$

where N_h is the size of cluster $K = 1, 2, \ldots, H$. In this setting a sample $\mathcal{S}$ from $\mathcal{P}$ is obtained by randomly selecting (without replacement) $h < H$ clusters and *taking all units* from these h clusters.

(a) [1 point] Derive the (marginal) inclusion probability, $\pi_u = P(u \in \mathcal{S})$

Assume, without loss of generality, that $u \in \mathcal{P}_K$

$P(u \in S) = P(\mathcal{P}_K \text{ is selected})$

$= \frac{h}{H}$ ✓ since the selection of h clusters from H is a SRSWOR.

(b) [2 points] Derive the joint inclusion probability, $\pi_{uv} = P(u \in \mathcal{S}, v \in \mathcal{S})$

<u>If $u, v \in \mathcal{P}_K$:</u>

$P(u,v \in S) = P(\mathcal{P}_h \text{ is selected}) = \frac{h}{H}$ ✓

<u>If $u \in \mathcal{P}_K, v \in \mathcal{P}_j$:</u>

$P(u,v \in S) = P(\mathcal{P}_K \text{ is selected and } \mathcal{P}_j \text{ is selected})$

$= \frac{h(h-1)}{H(H-1)}$ ✓ since selection of clusters is a SRSWOR.

(c) [2 points] Suppose that *two-stage cluster sampling* is employed. Within this paradigm the sample $\mathcal{S}$ is obtained in two stages:

- Randomly select (without replacement) $h < H$ clusters
- From each of those h clusters, randomly select (without replacement) n units.

Assuming $u \in \mathcal{P}_K$, calculate the (marginal) inclusion probability $\pi_u = P(u \in \mathcal{S})$.

$P(u \in S) = P(\mathcal{P}_K \text{ is selected and } u \text{ is selected from } \mathcal{P}_K)$

$= P(u \text{ is selected from } \mathcal{P}_K \mid \mathcal{P}_K \text{ is selected}) P(\mathcal{P}_K \text{ is selected})$

$= \overset{✓}{\frac{n}{N_K}} \times \overset{✓}{\frac{h}{H}}$ since u is selected from $\mathcal{P}_K$ in accordance with SRSWOR and $\mathcal{P}_K$ is selected from $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_H\}$ in accordance with SRSWOR.

minor differences in notation here are okay

* Responses here don't need all of the explanation I provided. Correct formulas are sufficient for full points.

3. **[6 points]** Suppose that $\mathcal{S} = \{1, 3\}$ is a *simple random sample without replacement* from a population $\mathcal{P}$ of size $N = 5$. Relevant inclusion probabilities are shown below

$$\begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}$$

(a) [2 points] Calculate the Horvitz-Thompson estimate of the population average.

$$a_{HT}(S) = \sum_{u\in S} \frac{y_u}{\pi_u} = \frac{1/5}{0.4} + \frac{3/5}{0.4} = 2 \quad ✓✓$$

(b) [2 point] The variance of the Horvitz-Thompson estimator is

$$Var\left[\tilde{a}_{HT}(\mathcal{S})\right] = \sum_{u\in\mathcal{P}}\sum_{v\in\mathcal{P}} (\pi_{uv} - \pi_u\pi_v)\frac{y_u}{\pi_u}\frac{y_v}{\pi_v}$$

State the formula for the estimate of this variance and show that the estimated variance is 15.

Should have been 0.6. Due to this typo, everyone gets full points for this part.

$$\widehat{Var}[\tilde{a}_{HT}(S)] = \sum_{u\in P}\sum_{v\in P}\left(\frac{\pi_{uv}-\pi_u\pi_v}{\pi_{uv}}\right)\left(\frac{y_u}{\pi_u}\right)\left(\frac{y_v}{\pi_v}\right)$$

$$(\text{since } \pi_{uu} = \pi_u) \rightarrow \quad = \sum_{u\in P}\frac{\pi_u(1-\pi_u)}{\pi_u}\left(\frac{y_u^2}{\pi_u^2}\right) + \sum_{\substack{u\in P\\ u\neq v}}\sum_{v\in P}\left(\frac{\pi_{uv}-\pi_u\pi_v}{\pi_{uv}}\right)\left(\frac{y_u}{\pi_u}\right)\left(\frac{y_v}{\pi_v}\right)$$

$$= (1-0.4)\frac{(1/5)^2}{0.4^2} + (1-0.4)\frac{(3/5)^2}{0.4^2} + 2\left(1-\frac{0.4^2}{0.1}\right)\left(\frac{1/3}{0.4}\right)\left(\frac{1/5}{0.4}\right)$$

$$= 0.15 + 1.35 - 0.9$$

$$= 0.6$$

(c) [1 point] Calculate the standard error of the estimate from part (a).

Actual answer : $SE[\tilde{a}_{HT}(S)] = \sqrt{\widehat{Var}[\tilde{a}_{HT}(S)]} = \sqrt{0.6} \approx 0.7746$

Acceptable response : $SE[\tilde{a}_{HT}(S)] = \sqrt{15} \approx 3.8730$ ✓

(d) [1 point] Calculate an approximate 95% confidence interval for the true population average.

Actual answer : $2 \pm 2\sqrt{0.6} = [0.4508, 3.5492]$

Acceptable response : $2 \pm 2\sqrt{15} = [-5.7460, 9.7460]$ ✓

these will b the answe that are accepted as correct.

4. **[6 points]** This question concerns the anatomy of a significance test meant to compare sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$, containing N_1 and N_2 units respectively.

(a) [1 point] State the null hypothesis H_0 associated with a permutation test that compares $\mathcal{P}_1$ and $\mathcal{P}_2$.

H_0: P_1 and P_2 are randomly sampled from the same population.

Also okay: Something along the lines of P_1 and P_2 are indistinguishd

(b) [1 point] Given an appropriately defined discrepancy measure $D(\mathcal{P}_1, \mathcal{P}_2)$, what types of values provide evidence against H_0? (Circle one).

i. extremely small [✓ ii. extremely large] iii. both

(c) [1 point] By filling in the blank probability expression below, define the p-value associated with this test. Define any notation you introduce.

p-value $= Pr($ $D \geq d_{obs} \mid H_0$ is true $)$

half point off if the conditioning is missed

where d_{obs} is the observed value of the discrepancy D.

(d) [2 points] Explain how the p-value in part (c) is calculated in practice.

- d_{obs} is calculated as $D(P_1, P_2)$ where P_1 and P_2 are the originally observed sub-populations.
- The sub-populations are randomly shuffled M times, each time yielding the pair $\{P_1^*, P_2^*\}$ and a discrepancy value: $D(P^*_{1,i}, P^*_{2,i})$ for $i = 1, 2, \ldots, M$
- The p-value is calculated as the proportion of $D(P^*_{1,i}, P^*_{2,i})$ values atleast as extreme as d_{obs}:

$$\text{pvalue} = \frac{1}{M} \sum_{i=1}^{M} I_{[d_{obs}, \infty)}\left(D(P^*_{1,i}, P^*_{2,i})\right)$$

Some sort of a written discription that conveys this message is fine.
- -1 if almost correct
- -2 if totally wrong

Don't need this formula if it is conveyed correctly with words.

(e) [1 point] In a *true* permutation test, how many discrepancy values is the null distribution composed of?

$\binom{N_1 + N_2}{N_1}$ or, equivalently $\binom{N_1 + N_2}{N_2}$ ✓

5. **[4 points]** Researchers are interested in determining the job-acquisition outcomes of graduates from undergraduate Data Science programs in Canada. In particular, interest lies in estimating the proportion of such students that obtain a job within 3 months of graduation. In order to study this phenomenon, the researchers observe a sample of the 2020 graduates from the University of Waterloo's BMATH in Data Science program.

   (a) [1 point] The **target population** in this scenario is:

   Graduates from undergraduate Data Science programs in Canada ✓

   (b) [1 point] The **study population** in this scenario is:

   ✓

   2020 graduates from UW's BMATH is Data Science program

   (c) [1 point] Define **study error**.

   ✓ Don't need both a statement and formula. Just one is fine.

   The difference between attributes calculated on the target vs. study populations: $a(P_{Study}) - a(P_{Target})$

   (d) [1 point] In the scenario described above, give one possible source of study error.

   Maybe UW students are smarter than other university students, and so their outcomes do not represent all Canadian undergraduates. Or anything similar along these lines

6. **[4 points]** Determine whether the following statements are True or False. In each case circle the correct answer.

   (a) [1 point] Considering all possible samples is the only way to determine the *exact* sampling distribution of an attribute $a(\mathcal{P})$.

   [✓ i. True]
   ii. False

   (b) [1 point] When interest lies in quantifying sampling error, probabilistic sampling is to be preferred over non-probabilistic sampling.

   [✓ i. True]
   ii. False

   (c) [1 point] If we hypothesized that the average from $\mathcal{P}_1$ was larger than the average from $\mathcal{P}_2$, then $D(\mathcal{P}_1, \mathcal{P}_2) = \overline{y}_1 - \overline{y}_2$ is a suitable discrepency measure.

   i. True
   [✓ ii. False]

   (d) [1 point] A large p-value provides evidence in favor of the null hypothesis H_0.

   i. True
   [✓ ii. False]