Please print in pen:

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: ____________________

Please initial:

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 $\widetilde{a}(\mathcal{P})$.

   (a) [2 points] Show that

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

   (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 $E[\widetilde{a}(\mathcal{S})] = 4.05$

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

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

Please initial:

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

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

where N_h is the size of cluster $h = 1, 2, \ldots, H$. 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, $\pi_u = P(u \in \mathcal{S})$

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

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

Please initial:

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

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

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

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

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

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

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

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

Please initial:

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 $\mathcal{P}_2$.

(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.

$$p\text{-value} = Pr(\qquad\qquad\qquad\qquad)$$

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

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

Please initial:

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:

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

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

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

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

Please initial:

This space is left for rough work

Please initial:

This space is left for rough work

Please initial: