STAT 341: Assignment 3

DUE: Monday March 16 by 11:59pm EST

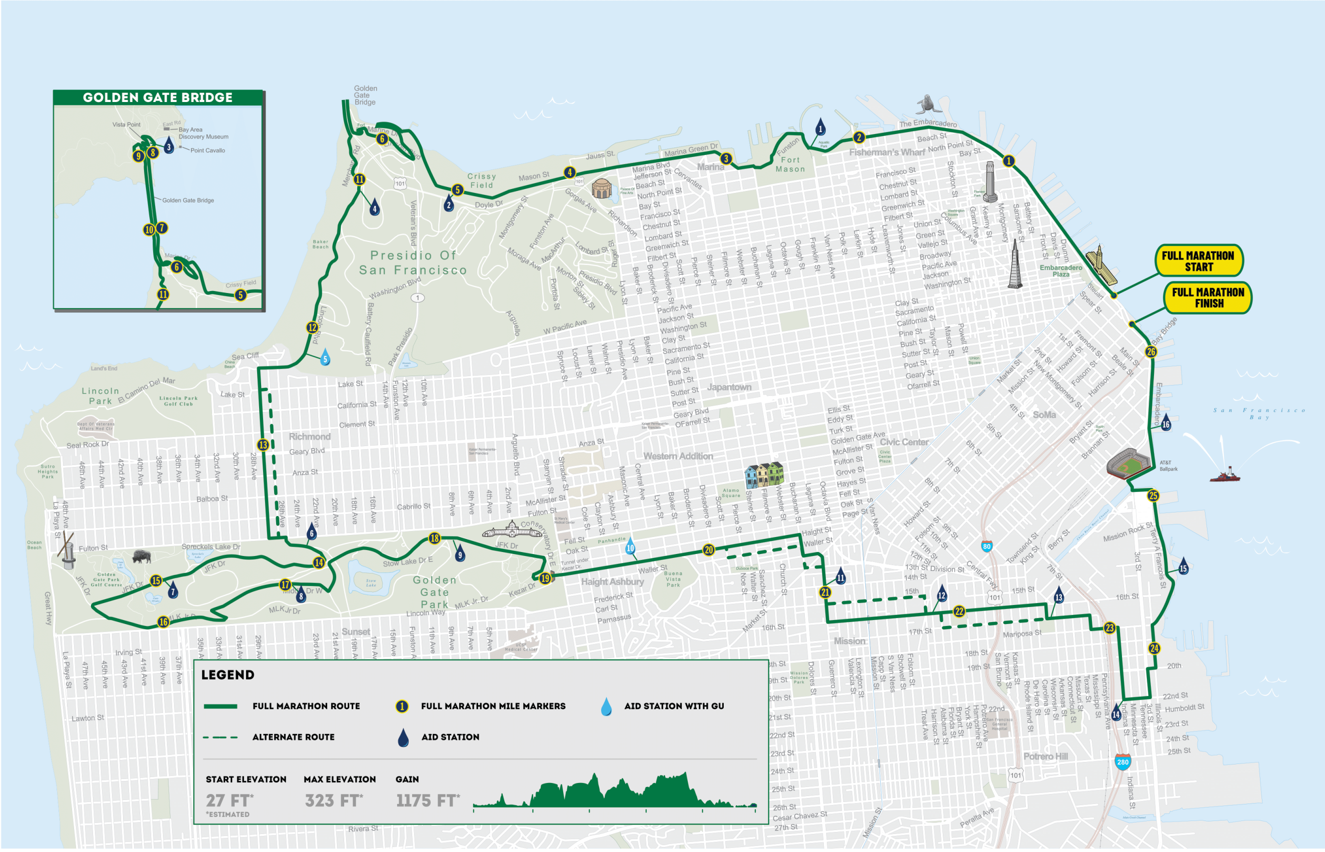
NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. This means that your responses for different question parts should begin on separate pages of your .pdf file. Note that your .pdf solution file must have been generated by R Markdown. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Neither screenshots nor scanned/photographed handwritten solutions will be accepted – these will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will received zero points.
- For interpretation questions: plain text (within R Markdown) is required. Text responses embedded as comments within code chunks will not be accepted.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible.

CONTEXT



On Sunday July 29, 2018 more than five thousand individuals ran the San Francisco Marathon which boasts 26.2 miles (42.2km) of scenic views and iconic San Francisco landmarks. In this assignment you will analyze this data, which is available in the `SF_Marathon_2018.csv` file. Each unit in this population is an individual runner, and for each runner we have the following 7 variates recorded:

Variate	Value
`Place`	1 to 5262, identifying the rank of the runner
`Name`	The name of the runner
`Bib`	The runner's bib number
`Time`	The runner's marathon time, recorded as `"H:MM:SS"`
`Pace`	The runner's average pace, recorded as `"MM:SS"` per mile
`Sex`	The runner's sex, recorded as `"F"` or `"M"`
`Age.Division`	The runner's age bracket, recorded in years

QUESTION 1: Data Prep and EDA [7 points]

(a) [1 point] Load the data and save it in a variable called `race`.

(b) [1 point] Replace the entries in the `Time` column of `race` with marathon times recorded in minutes. In other words, convert the character string `H:MM:SS` into a numeric value. For example, `"4:11:32"` would be converted to `251.5333`. To confirm that you have done this correctly print out the first 6 rows of the dataframe `race`. **Note:** You might find the `hour`, `minute`, `second`, and `hms` functions from the `lubridate` package helpful.

(c) [1 point] Remove the two runners for whom `Age.Division = "NoAge"`. Be sure to state which rows were eliminated. **Note:** For the rest of the assignment, the remaining $N = 5262$ runners will be treated as the population $\mathcal{P}$ of interest.

(d) [1 point] Calculate and state the average marathon time. **Note:** For the Questions 2 and 3, we will treat this value as $a(\mathcal{P})$, the attribute of interest calculated on the entire population.

(e) [2 points] Construct a histogram of the marathon finishing times and, with a vertical line, indicate the average finishing time. Be sure to include a relevant title and axis labels.

(f) [1 point] How many minutes did it take `Nathaniel Stevens` to finish the marathon? Round your answer to 4 decimal places.

QUESTION 2: Horvitz-Thompson Estimation – SRSWOR [12 points]

(a) [5 points] Using the following code, take a *simple random sample without replacement* of size $n = 500$ from the population. (This is not worth points)

```
srsSampIndex <- read.table("srsSampIndex.txt")$V1
srsSamp <- race[srsSampIndex, ]
```

i. [2 points] Calculate the Horvitz-Thompson estimate of the average finishing time.

ii. [2 points] Calculate the standard error for this estimate. You may find the following function useful:

```
estVarHT <- function(y_u, pi_u, pi_uv) {
    ## y_u = an n element array containing the variate values for the sample

    ## pi_u = an n element array containing the (marginal) inclusion
    ## probabilities for the sample

    ## pi_uv = an nxn matrix containing the joint inclusion probabilities
    ## for the sample

    delta <- pi_uv - outer(pi_u, pi_u)
    estimateVar <- sum((delta/pi_uv) * outer(y_u/pi_u, y_u/pi_u))
    return(abs(estimateVar))
}
```

iii. [1 point] Calculate an approximate 95% confidence interval for the average finishing time.

(b) [7 points] In this question you will explore the dependency of the Horvitz-Thompson estimator's sampling distribution on sample size. Consider the sample sizes $n \in \{100, 200, \ldots, 1000\}$.

i. [4 points]

- For each sample size n take 50,000 SRSWOR samples from the population.
- For each of the 50,000 samples of a given size, calculate the Horvitz-Thompson estimate of the average finishing time.
- For each sample size n use the 50,000 HT estimates to estimate the sampling bias, sampling variance, and sampling MSE.
- Construct three line-plots (layed out in a 1×3 grid) of bias vs. n, variance vs. n and MSE vs. n. As always, make sure your plots are informatively labelled.

ii. [3 points] Comment on the relationship between bias, variance, MSE and n for the Horvitz-Thompson Estimator.

QUESTION 3: Horvitz-Thompson Estimation – Stratified Random Sampling [24 points]

Stratified randomly sampling is a probabilistic sampling mechanism that is applicable when a population $\mathcal{P}$ can be partitioned into H strata (i.e., sub-populations) $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_H\}$ such that

$$\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \ldots \cup \mathcal{P}_H$$

and

$$N = N_1 + N_2 + \cdots + N_H$$

where N_h is the size of strata $h = 1, 2, \ldots, H$.

In this setting, a sample $\mathcal{S}$ of size n from $\mathcal{P}$ is obtained by taking independent *simple random samples without replacement* from *each* of the H strata. Thus, sample $\mathcal{S}_1$ of size n_1 is taken from strata 1, sample $\mathcal{S}_2$ of size n_2 is taken from strata 2, and so on. Then

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_H$$

where $n = n_1 + n_2 + \cdots + n_H$ and each $\mathcal{S}_h$ is a SRSWOR.

The function `stratRS` defined below may be used to take a stratified randomly sample $\mathcal{S}$ of size n from a population $\mathcal{P}$ of size N. The inputs are:

- `stratLabel`: an N element array containing the integer labels $\{1, 2, \ldots, H\}$ indicating which strata each unit belongs to.
- `stratSampSize`: an H element array containing $\{n_1, n_2, \ldots, n_H\}$, the sample sizes to be drawn from each strata.

```r
stratRS <- function(stratLabel, stratSampSize) {
    H <- length(stratSampSize)
    sampIndex <- list()
    for (h in 1:H) {
        sampIndex[[h]] <- sample(which(stratLabel == h), size = stratSampSize[h],
            replace = FALSE)
    }
    return(unlist(sampIndex))
}
```

(a) Marginal Inclusion Probabilities

i. [2 points] Show that the (marginal) inclusion probability π_u for stratified random sampling is

$$\pi_u = \frac{n_h}{N_h} \qquad \text{if } u \in \mathcal{P}_h$$

ii. [2 points] Write a function called `getInclusionProbStrat` which outputs an N element array containing the inclusion probabilities for each unit in the population $\mathcal{P}$, and which takes as inputs `stratLabel` and `stratSampSize` as defined above.

iii. [1 point] Call `getInclusionProbStrat` using `stratLabel = c(1,1,1,2,2,2,2,2)` and `stratSampSize = c(2,3)` and output the result.

(b) Joint Inclusion Probabilities

i. [4 points] Show that the joint inclusion probability π_{uv} for stratified random sampling is

$$\pi_{uv} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & \text{if } u, v \in \mathcal{P}_h \\ \\ \frac{n_h n_k}{N_h N_k} & \text{if } u \in \mathcal{P}_h, v \in \mathcal{P}_k \end{cases}$$

ii. [2 points] Write a function called `getJointInclusionProbStrat` which outputs an $N \times N$ matrix containing the joint inclusion probabilities for each unit in the population $\mathcal{P}$, and which takes as inputs `stratLabel` and `stratSampSize` as defined above.

iii. [1 point] Call `getJointInclusionProbStrat` using `stratLabel = c(1,1,1,2,2,2,2)` and `stratSampSize = c(2,3)` and output the result.

(c) [2 points] Add a new column to the `race` dataframe called `stratLabel` which assigns a numeric label to each age division. In particular, when `Age.Division = "Under20"` then `stratLabel = 1`; when `Age.Division = "20-24"` then `stratLabel = 2`; when `Age.Division = "25-29"` then `stratLabel = 3`; … when `Age.Division = "65+"` then `stratLabel = 11`. Once you have done this, present the output from the command `table(race$stratLabel)`.

(d) [5 points] Using the following code, take a *stratifed random sample* of size $n = 500$ from the population. (This is not worth points)

```r
stratSampIndex <- read.table("stratSampIndex.txt")$V1
stratSamp <- race[stratSampIndex, ]
```

i. [2 points] Calculate the Horvitz-Thompson estimate of the average finishing time. It will be useful to use your `getInclusionProbStrat` function.

ii. [2 points] Calculate the standard error for this estimate. Feel free to use the `estVarHT` function. It will also be useful to use your `getJointInclusionProbStrat` function.

iii. [1 point] Calculate an approximate 95% confidence interval for the average finishing time.

(e) [3 points] Take 50,000 stratified random samples of size $n = 500$ from the population with $\{n_1 = 7, n_2 = 46, n_3 = 97, n_4 = 84, n_5 = 71, n_6 = 57, n_7 = 53, n_8 = 40, n_9 = 24, n_{10} = 13, n_{11} = 8\}$. For each sample calculate the Horvitz-Thompson estimate of the average finishing time. Construct a histogram of these estimates and overlay a vertical line indicating $a(\mathcal{P})$ the true average finishing time.

(f) [2 points] Using the approximate sampling distribution calculated in (e), estimate the sampling bias, sampling variance and sampling mean squared error (MSE) of the stratified random sample Horvitz-Thompson estimator. Compare these results to the $n = 500$ case from Question 2(b) i. and state which sampling mechanism is to be preferred and explain why.

QUESTION 4: Permutation Test [12 points]

Consider the female (`F`) and male (`M`) subpopulations of runners. We may refer to them as $\mathcal{P}_F$ and $\mathcal{P}_M$ and we note that $\mathcal{P} = \mathcal{P}_F \cup \mathcal{P}_M$.

(a) [3 points] Construct two *density* histograms: one of the female finishing times and the other of the male finishing times, and plot them next to each in a 1 × 2 grid. Be sure to include informative titles and axis labels. To enhance comparability ensure the x-axis of each histogram is the same, ensure both histograms have the same number of bins, and indicate (with a vertical line) the average marathon time in each sub-population.

(b) [1 point] State the null hypothesis H_0 that is being tested when comparing these two sub-populations with a permutation test.

(c) [4 points] In this question you will test the hypothesis in (b) using the discrepancy measure

$$D(\mathcal{P}_F, \mathcal{P}_M) = |\overline{y}_F - \overline{y}_M|$$

i. [1 point] Calculate the observed discrepancy.

ii. [1 point] Randomly mix the populations $M = 5,000$ times and construct a histogram of the 5,000 $D(\mathcal{P}_F^\star, \mathcal{P}_M^\star)$ values. Indicate, with a vertical line, the observed discrepancy calculated in i. Note that you may use the `mixRandomly` function from class.

iii. [1 point] Calculate the p-value associated with this test.

iv. [1 point] Based on the p-value calculated in iii. what do you conclude about the comparability of these two populations? In other words, summarize your findings and draw a conclusion about the null hypothesis from part (b).

(d) [4 points] In this question you will test the hypothesis in (b) using the discrepancy measure

$$D(\mathcal{P}_F, \mathcal{P}_M) = \left| \frac{SD(\mathcal{P}_F)}{SD(\mathcal{P}_M)} - 1 \right|$$

i. [1 point] Calculate the observed discrepancy.

ii. [1 point] Randomly mix the populations $M = 5,000$ times and construct a histogram of the 5,000 $D(\mathcal{P}_F^\star, \mathcal{P}_M^\star)$ values. Indicate, with a vertical line, the observed discrepancy calculated in i. Note that you may use the `mixRandomly` function from class.

iii. [1 point] Calculate the p-value associated with this test.

iv. [1 point] Based on the p-value calculated in iii. what do you conclude about the comparability of these two populations? In other words, summarize your findings and draw a conclusion about the null hypothesis from part (b).