Homework 02

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

MEMBER 1 (NetID 1) MEMBER 2 (NetID 2)
MEMBER 3 (NetID 3)

! Important Instructions

- This document contains **all** of the problems for homework 2. Questions are a mix of theoretical and coding.
- Please write your answers in the spaces provided (replacing the text that says "***Replace this line with your answers*** " with your work)
  - If you are comfortable using LaTeX, you may typeset your answers to the theoretical questions.
  - Alternatively, you may write your answers to the theoretical portions on a separate sheet of paper, take a picture of your work, and include a picture in your QMD document.
  - Do NOT try to simply type your answers to the theoretical questions into this document if you are not using LaTeX - we should be able to read all of your equations and computations clearly and easily.
- To prove that you read these instructions fully, please copy and paste the following phrase at the very end of your document: "I have read the instructions fully, and am including the code phrase: meow cat please meow back"
- As always, you must produce a PDF, which you will then submit to Gradescope.
  - After submitting, make sure to **match pages**; failure to do so will result in point penalties on this homework.

Question 1: Sampling Distribution of the Maximum

Consider $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$. Since the parameter θ is the right endpoint of the support, it makes sense to use the sample maximum as an estimator for θ. As such, let us take

$$\hat{\theta}_n := \max_{1 \leq i \leq n} \{X_i\}$$

Part (a)

Suppose $\theta = 10$; i.e. suppose we have $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$. Use `R` to simulate taking 1000 samples of size $n = 100$ from the Unif$[0, 10]$ distribution, computing the observed value of $\hat{\theta}_n$ for each sample, and then plotting the resulting observed values to obtain an approximation to the sampling distribution of $\hat{\theta}_n$. For the `y` aesthetic in your call to `geom_histogram()`, specify `y = after_stat(density)` [don't worry too much about what this does - we'll talk about that later in the course].

ANSWERS TO QUESTION 1(a):

Replace this line with your answers

Part (b)

It can be shown (but you do not need to show this) that the exact distribution of $\hat{\theta}_n$ is given by

$$f_{\hat{\theta}_n}(x) = \frac{nx^{n-1}}{\theta^n} \cdot \mathbb{1}_{\{0 \leq x \leq \theta\}}$$

So, for example, if we take samples of size $n = 100$ the sampling distribution of the sample maximum is given by

$$f_{\hat{\theta}_n}(x) = \frac{100x^{99}}{\theta^{100}} \cdot \mathbb{1}_{\{0 \leq x \leq \theta\}}$$

Reproduce your histogram from part (a), and add a call to `stat_function()` to overlay the true sampling distribution of $\hat{\theta}_n$. Your final plot should look similar (in spirit) to the one created during the demo in Lecture on Tuesday.

ANSWERS TO QUESTION 1(b):

Replace this line with your answers

Part (c)

Using the formula

$$f_{\hat{\theta}_n}(x) = \frac{nx^{n-1}}{\theta^n} \cdot \mathbb{1}_{\{0 \le x \le \theta\}}$$

compute $\mathbb{E}[\hat{\theta}_n]$, and use this to compute the bias of using $\hat{\theta}_n$ as an estimator for θ. [Do not assume $n = 100$ or that $\theta = 10$ anymore; your final expression should be a function of n and θ.]

ANSWERS TO QUESTION 1(c):

Replace this line with your answers

Part (d)

Compute the empirical mean of the 1000 observed values of $\hat{\theta}_n$ you generated in part (a). Compare this to the expected value you computed in part (c) above (after plugging in $n = 100$ and $\theta = 10$).

ANSWERS TO QUESTION 1(d):

Replace this line with your answers

Question 2: Sampling Distribution of the Median

In general, there aren't too many results pertaining to the sample median as an estimator for the population median. However, in certain simple cases, we can still derive some useful results both theoretically and empirically.

Consider a "population" consisting of four values: $\mathcal{P} := \{2, 3, 4, 5\}$. Suppose we take a random sample (X_1, X_2, X_3) of three of these values *without replacement*, and suppose order does not matter [i.e. assume the sample $(2, 3, 4)$ is the same as the sample $(4, 3, 2)$]. Let $\widehat{M}$ denote the sample median of our sample and let m denote the true population median (which is 3.5, in this problem). Assume all 3-element subsets of $\mathcal{P}$ are equally likely to be selected.

Part (a)

Construct the sampling distribution of $\widehat{M}$. Effectively, this amounts to running through all possible samples of size 3 taken from the population, computing the value of $\widehat{M}$ that each outcome maps to, and then constructing a PMF from these values while leveraging the fact that all three-element

subsets are equally likely.

ANSWERS TO QUESTION 2(a):

Replace this line with your answers

Part (b)

Use your answer from part (a) to compute $\mathbb{E}[\widehat{M}]$, and use this to determine whether or not $\widehat{M}$ is an unbiased estimator of m.

ANSWERS TO QUESTION 2(b):

Replace this line with your answers

Part (c)

Simulate taking 1000 samples of size 3, without replacement, from $\mathcal{P}$. Construct an empirical approximation to the sampling distribution of $\widehat{M}$, and display the resulting histogram. (Note: in this problem it could be argued that a *barplot* would be a better visualization, but for how we'll still stick with a histogram.)

ANSWERS TO QUESTION 2(c):

Replace this line with your answers

Question 3: Wait Times

The fast food chain *DacMonalds* advertises: "get your food in under 5 minutes!". To test this claim, Jaslene takes an i.i.d. sample of 100 *DacMonalds* customers and records the amount of time (in minutes) they spent waiting in line. Her data is included in the file `wait_times.csv`, located in the `data/` subfolder.

Part (a)

Display a histogram of the wait times that Jaslene observed.

ANSWERS TO QUESTION 3(a):

Replace this line with your answers

Part (b)

Suppose we interpret *DacMonalds*' claim as "the average wait time of customers is 5 minutes". Using appropriate notation, write down the null and alternative hypotheses, assuming a two-sided alterantive. Make sure to define any parameter(s) fully and clearly, in words.

ANSWERS TO QUESTION 3(b):

Replace this line with your answers

Part (c)

Assuming a 5% level of significance, write down the rejection region of the test of the hypotheses you wrote down in part (b).

ANSWERS TO QUESTION 3(c):

Replace this line with your answers

Part (d)

Using the data in the `wait_times.csv` file, compute the observed value of the test statistic. Use this to conduct (again, assuming a 5% level of significance) a test of the hypotheses you formulated in

part (b). Be sure to state your conclusions in the context of the problem.

ANSWERS TO QUESTION 3(d):

Replace this line with your answers

Part (e)

Compute the p-value of the observed value of the test statistic you computed in part (d).

ANSWERS TO QUESTION 3(e):

Replace this line with your answers

Part (f)

Construct a 90% confidence interval for the true wait times of *DacMonalds* customers.

ANSWERS TO QUESTION 3(f):

Replace this line with your answers