Homework 02 SOLUTIONS

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

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

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 $\text{Unif}[0, 10]$ distribution, computing the observed value of $\widehat{\theta}_n$ for each sample, and then plotting the resulting observed values to obtain an approximation to the sampling distribution of $\widehat{\theta}_n$. 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

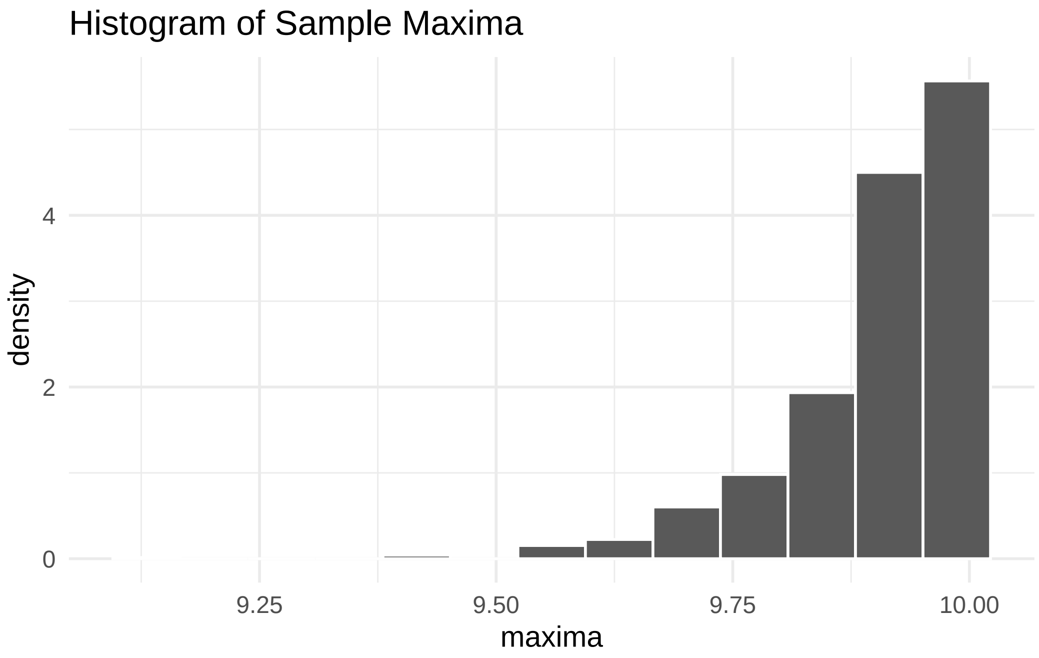
```r
library(tidyverse)
set.seed(123)

n <- 100
B <- 1000

maxima <- c()
for(b in 1:B) {
  maxima <- c(maxima, max(runif(n, 0, 10)))
}

p1 <- maxima %>%
  data.frame() %>%
  ggplot(aes(x = maxima)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 13,
    col = "white") +
  theme_minimal() +
  ggtitle("Histogram of Sample Maxima")

p1
```



Part (b)

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

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

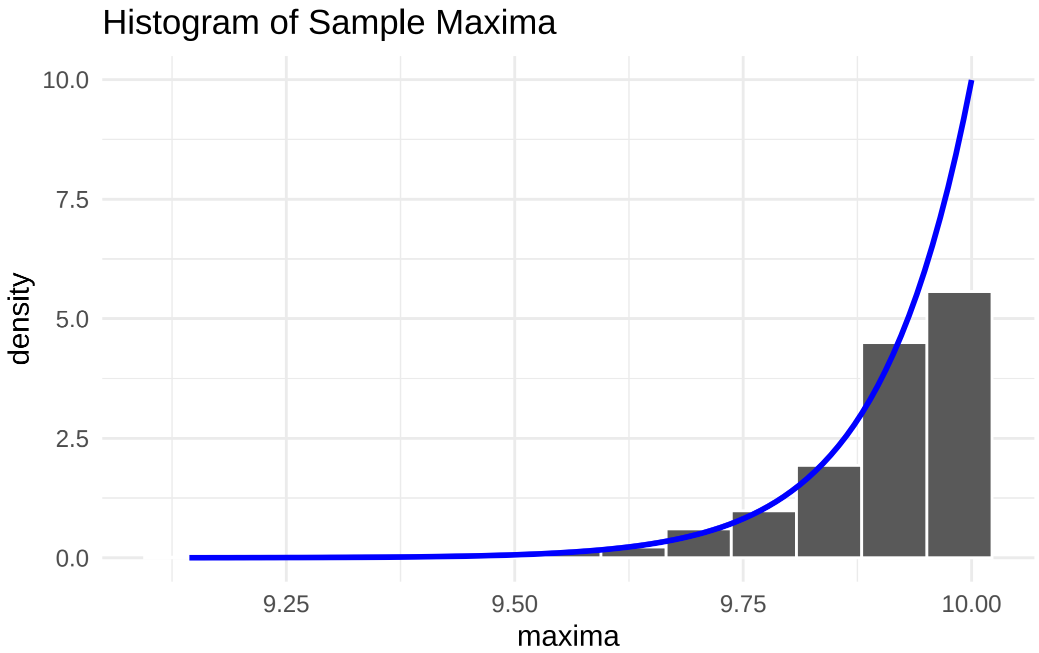
Reproduce your histogram from part (a), and add a call to `stat_function()` to overlay the true sampling distribution of $\widehat{\theta}_n$. Your final plot should look similar (in spirit) to the one created during the demo in Lecture on Tuesday.

ANSWERS TO QUESTION 1(b):

Replace this line with your answers

```
true_dens <- Vectorize(function(x) {
  if((x < 0) | (x > 10)) {
    return(0)
  } else {
    return(100 * x^99 / (10^100))
  }
})
```

```
p1 +
  stat_function(
    fun = true_dens,
    col = "blue",
    linewidth = 1
  )
```



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

$$
\begin{aligned}
\mathbb{E}[\widehat{\theta}_n] &:= \int_{-\infty}^{\infty} x f_{\widehat{\theta}_n}(x) \ \mathrm{d}x \\
&= \int_0^{\theta} x \cdot \frac{n x^{n-1}}{\theta^n} \ \mathrm{d}x = \frac{n}{\theta^n} \int_0^{\theta} x^n \ \mathrm{d}x = \frac{n}{n+1} \cdot \frac{\theta^{n+1}}{\theta^n} = \boxed{\left(\frac{n}{n+1} \right) \theta}
\end{aligned}
$$

$$
\implies \mathrm{Bias}(\widehat{\theta}_n, \theta) = \mathbb{E}[\widehat{\theta}_n] - \theta = \left(\frac{n}{n+1} \right) \theta - \theta = \boxed{-\left(\frac{1}{n+1} \right) \theta}
$$

Part (d)

Compute the empirical mean of the 1000 observed values of $\widehat{\theta}_n$ you generated in part (a). Compare this to the expected value you computed in part (c) above.

ANSWERS TO QUESTION 1(d):

Replace this line with your answers

```
mean(maxima)     # empirical mean
```

```
[1] 9.902243
```

```
(100/101) * 10   # true mean
```

```
[1] 9.90099
```

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*; 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 $\mathscr{P}$ 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.

ANSWERS TO QUESTION 2(a):

Replace this line with your answers

Outcome	Sample Median
(2, 3, 4)	3
(2, 3, 5)	3
(2, 4, 5)	4
(3, 4, 3)	4

Hence, since we are assuming each three-element subset is equally likely, we have that the probability of each of the above outcomes is simply 1/4 meaning the sampling distribution of $\widehat{M}$ is

$\boldsymbol{k}$	3	4
$\mathbb{P}(\widehat{M} = m)$	1/2	1/2

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

$$\begin{aligned} \mathbb{E}[\widehat{M}] &:= \sum_k k \cdot \mathbb{P}(\widehat{M} = k) \\ &= (3)\left(\frac{1}{2}\right) + (4)\left(\frac{1}{2}\right) = \boxed{3.5} \end{aligned}$$

Since this is equal to the population median, $\widehat{M}$ is an unbiased estimator of the population median.

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.

ANSWERS TO QUESTION 2(c):

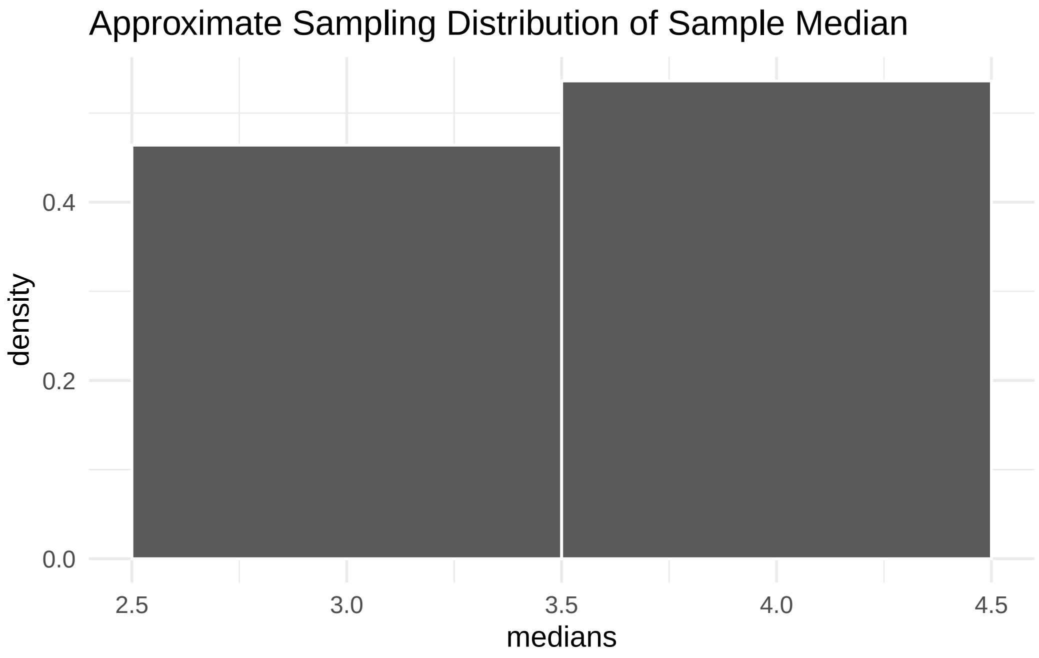
Replace this line with your answers

```r
set.seed(123)

n <- 3          # sample size
B <- 1000       # number of samples
pop <- 2:5      # population

medians <- c()
for(b in 1:B) {
  medians <- c(medians, sample(pop, 3) %>% median())
}

medians %>%
  data.frame() %>%
  ggplot(aes(x = medians)) +
  geom_histogram(
    aes(y = after_stat(density)),
    col = "white",
    bins = 2
  ) +
  theme_minimal() +
  ggtitle(
    "Approximate Sampling Distribution of Sample Median"
  )
```



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.

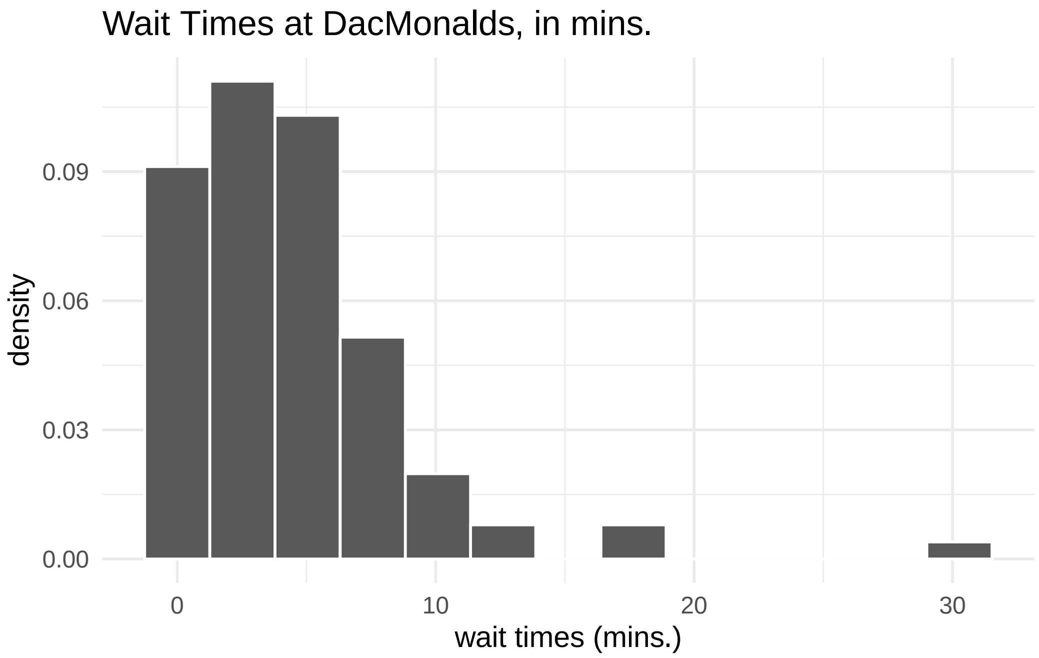
Part (a)

Display a histogram of the wait times that Jaslene observed.

ANSWERS TO QUESTION 3(a):

Replace this line with your answers

```
wait_times <- read.csv("data/wait_times.csv")
wait_times %>%
  data.frame() %>%
  ggplot(aes(x = x)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 13,
    col = "white"
  ) +
  theme_minimal() +
  xlab("wait times (mins.)") +
  ggtitle(
    "Wait Times at DacMonalds, in mins."
  )
```



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

Letting μ denote the true average wait time (in minutes) at *DacMonalds*, we can phrase our hypotheses as

$$\left[\begin{array}{rl} H_0: & \mu = 5 \\ H_A: & \mu \neq 5 \end{array} \right.$$

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

If we adopt an unstandardized test statistic, our rejection region for the mean will take the form

$$\left(-\infty \ , \ \mu_0 - c \cdot \frac{\mathrm{sd}}{\sqrt{n}} \right] \cup \left[\mu_0 + c \cdot \frac{\mathrm{sd}}{\sqrt{n}} \ , \ \infty \right)$$

where $\mu_0 = 5$, n denotes the sample size, sd denotes either the sample standard deviation or the population standard deviation, and c is the quantile of the appropriate distribution. In this case, we do not have access to the population standard deviation meaning we must use the t_{99} distribution; that is, c will be the appropriate quantiles of the t_{99} distributions.

```
crit_quant <- qt(1 - (0.05/2), 99)
5 - crit_quant * sd(wait_times$x) / sqrt(99)
```

```
[1] 4.130085
```

```
5 + crit_quant * sd(wait_times$x) / sqrt(99)
```

```
[1] 5.869915
```

Our rejection region is thus $\boxed{(-\infty, 4.130085] \cup [5.869915, \infty)}$

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

Again continuing to use an unstandardized test statistic, our test statistic will simply be the sample mean:

```r
mean(wait_times$x)
```

```
[1] 4.392019
```

Since this is *not* inside our rejection region, we **fail to reject the null**. That is:

> At a 5% level of significance, there was insufficient evidence to reject the initial claim that the average wait time in *DacMonalds* is 5 minutes.

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

Again, we have to use the t_{99} distribution:

```r
ts_stnd <- (mean(wait_times$x) - 5)/(sd(wait_times$x) / sqrt(99))

# p-value
2 * pt(-abs(ts_stnd), 99)
```

```
[1] 0.1686284
```