STAT 510: Homework 05

David Dalpiaz

Due: Tuesday, October 6, 11:59 PM

General Directions

This assignment is worth 10 points. For each exercise, you may obtain a score of 0, 0.5, or 1.

- To obtain a score of **1**, your answer must be correct, contain valid supporting work, and be reasonably formatted up to and including boxing your answer when possible.
- A score of **0.5** will be given to solutions which show reasonabe effort, but contain errors. (A score of **1** may be granted to some solutions containing errors if they are extremely minor.)
- A score of **0** will be given to a blank solution or a solution that shows no reasonable progress towards the correct solution. Note that if you do not indicate a page for a problem on Gradescope, it will be considered blank.

Please submit your assignment to Gradescope by the due date listed above. You may submit up to 48 hours late with a two point late penalty. After that, no late work will be accepted.

Any grade disputes must be petitioned through Gradescope within one week of receiving a grade.

Please attempt to submit your work as a single PDF and complete the process of indicating which problem is on which page. You may need to merge together PDF files from various sources and scans. We will keep track of best practice for submitting to Gradescope in this Piazza thread.

Homework assignments are meant to be learning experiences. You may discuss the exercises with other students, but you must write the solutions on your own. Directly sharing or copying any part of a homework solution is an infraction of the University's rules on academic integrity. Any violation will be punished as severely as possible.

For this, and all homework assignments, you may use any computational tools that you wish, such as a statistical computing enviroment or integral solver. The course staff is most familiar with `R`, so we will be able to best support `R` users, but you may use any software that you like.

Graded Exercises

Exercise 1 (Distribution of a Bootstrap Sample)

Let $X_1, X_2, \ldots, X_n$ be distinct observations, that is, no ties. Let $X_1^\star, X_2^\star, \ldots, X_n^\star$ denote a bootstrap sample and let

$$\bar{X}_n^\star = \frac{1}{n}\sum_{i=1}^{n} X_i^\star.$$

Find:

- $\mathbb{E}\left[\bar{X}_n^\star \mid X_1, X_2, \ldots, X_n\right]$
- $\mathbb{V}\left[\bar{X}_n^\star \mid X_1, X_2, \ldots, X_n\right]$
- $\mathbb{E}\left[\bar{X}_n^\star\right]$

- $\mathbb{V}\left[\bar{X}_n^\star\right]$

Exercise 2 (Professor Salaries)

The following code loads data about Professor salaries. (Check the documentation for details.) We will be interested in the `salary` variable.

```
salaries = carData::Salaries
```

Define $\theta = T(F) = q_{0.25}$. Create a 95% confidence interval for the 25th percentile of professor salaries using each of the three bootstrap interval methods: Normal, Pivotal, Percentile.

Use as least 2000 bootstrap samples for each interval.

- Note 1: To obtain a "plug-in" estimate for q_p you may simply use the default arguments to R's `quantile()` function.
- Note 2: These salaries are a few years old, but for **Tenure Track** faculty. Not all of your instructors fall into this category.
- Fun Fact: Illinois is a state institution, so salary information is public. We leave it as an exercise to the read to find this data.

Exercise 3 (How Long Will You Survive Cancer?)

For this exercise we will use the `Melanoma` data from the `MASS` package.

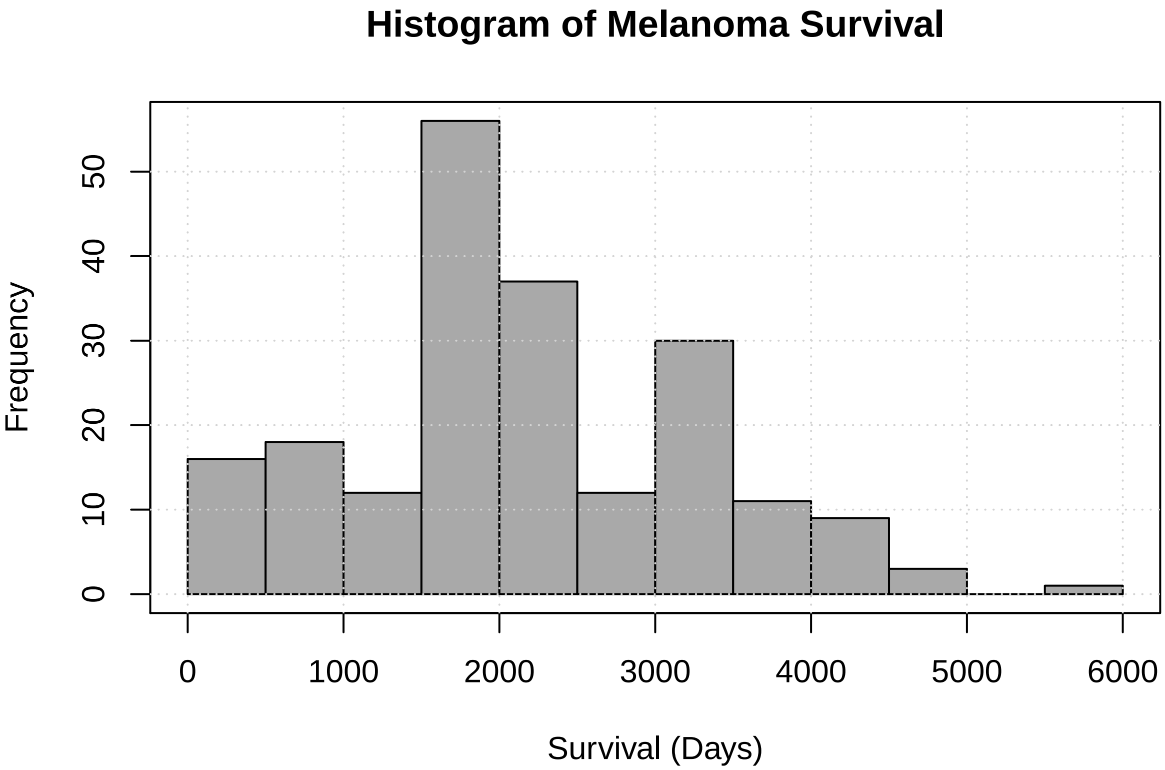
```
head(MASS::Melanoma)
```

```
##   time status sex age year thickness ulcer
## 1   10      3   1  76 1972      6.76     1
## 2   30      3   1  56 1968      0.65     0
## 3   35      2   1  41 1977      1.34     0
## 4   99      3   0  71 1968      2.90     0
## 5  185      1   1  52 1965     12.08     1
## 6  204      1   1  28 1971      4.84     1
```

We'll focus on the `time` variable which is survival time in days.

```
mel_survive = MASS::Melanoma$time
```

```
hist(mel_survive, col = "darkgrey",
     xlab = "Survival (Days)", main = "Histogram of Melanoma Survival")
box()
grid()
```



Let X be the survival time in **years** and define

$$\theta = T(F) = P(X > 5).$$

Create a 95% percentile bootstrap confidence interval for θ, the probability of surviving longer that 5 years. Use at least 20000 bootstrap samples. Also plot a histogram of the bootstrap replicates and overlay the large-sample approximate **estimated** sampling distribution.

Exercise 4 (Deflategate)

On January 18, 2015, Clete Blakeman measured the pressure in pounds per square inch (PSI) of 15 footballs during halftime of the AFC Championship game. Of these footballs, 11 were a sample from the New England Patriots. The remaining 4 were a sample from the Indianapolis Colts. The data follows:

```
pats  = c(11.50, 10.85, 11.15, 10.70, 11.10, 11.60, 11.85, 11.10, 10.95, 10.50, 10.90)
colts = c(12.70, 12.75, 12.50, 12.55)
```

Use the percentile method to create a 95% confidence interval for the difference in medians of the pressure of the Patriot's and Colt's footballs. Use at least 2000 bootstrap samples.

- Note 1: These sample sizes are probably too small.
- Note 2: This is not a rigorous enough analysis to discredit the Patriots. However, disliking the Patriots is totally normal and acceptable! For actual details, see the Wells Report. (Be aware that the report contains text messages that use some not so pleasant language.)
- Note 3: This is not "tidy" data, but for this example, it is much easy to work with.

Exercise 5 (Rank Correlation)

The following loads the `airquality` data and then removes any missing data.

```r
aq = na.omit(airquality)
```

Use the percentile method to create a 90% confidence interval for the Spearman rank correlation between `Ozone` and `Wind`. Use at least 2000 bootstrap samples.

Exercise 6 (Bootstrap Replicates and the Sampling Distribution)

The following code generates data.

```r
set.seed(42)
some_data = rnorm(n = 100, mean = 5, sd = 1)
```

Suppose that we were interested in estimating $\theta = e^{\mu}$ and wanted to consider the estimator

$$\hat{\theta} = e^{\bar{X}}.$$

Generate 2000 (or more) bootstrap replicates of this estimator. Plot a histogram of these bootstrap replicates and overlay the **true** sampling distribution.

Exercise 7 (Bootstrap Coverage)

Perform a simulation study to assess the coverage of the three bootstrap confidence interval methods we have discussed: Normal, Pivotal, Percentile

Let $n = 50$ and

$$T(F) = \int \frac{(x-\mu)^3}{\sigma^3} dF(x).$$

Generate $Y_1, Y_2, \ldots Y_n \sim N(0,1)$ and set $X_i = e^{Y_i}$ for $i = 1, 2, \ldots n$. With this sample $X_1, \ldots X_n$ construct a 95% confidence interval for $T(F)$ using each of the three methods. Use at least 500 bootstrap samples for each, but more is better.

Repeat this process at least 1000 times. Use the results to assess the coverage of the three interval types.

Exercise 8 (More Bootstrap Coverage)

Repeat the above exercise, but also report the average length of the three interval types in addition to their coverage. This time use random samples of size $n = 25$ from a t distribution with 3 degrees of freedom. That is

$$X_1, \ldots X_n \sim t_3$$

Let

$$\theta = T(F) = (q_{0.75} - q_{0.25})/1.34$$

To obtain a "plug-in" estimate for q_p you may simply use the default arguments to R's `quantile()` function.

Exercise 9 (Free Points)

Draw a smiley face! The previous two problems were not fun and we're close to the exam!

Exercise 10 (Free Points)

Write the following: "I will try my best on the exam, but I will also try my best to not stress too much about it!"

- Note: Yes, this is easier said than done.

Exercise 11 (BYOQ: Bring Your Own Question)

Submit your own question with a solution! If accepted, you will receive one **buffer point**. To be accepted:

- Your question must be reasonably challenging. It should be at least as challenging as the "average" question on this assignment.
- Is must be **original**. If, based on some quick searching we can find the *exact* question (that is, you just copy-pasted a question), you will **lose** a point instead of receiving a buffer point. If we find that it is simply a derivative of another exercise (for example, just changing a constant) you will not receive a point.

The instructor may create videos solving these, or they may be circulated for additional practice.