Homework 7: 3/13 - 3/17

STA 336 Due 11:59pm Wednesday, April 5 Name:

Instructions: Write-up complete solutions to the following problems and submit answers on Gradescope. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A rubric for homework problems appears on the final page of this assignment.

- Unless otherwise noted, problem numbers are taken from the 4th edition of DeGroot and Schervish's *Probability and Statistics*.

Monday 3/13

Additional Problems

AP1. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed values of a random sample, and let $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ be a bootstrap sample from $\mathbf{x}$.

(a) Let D be the set of unique values in the sample $\mathbf{X}^*$; for example, if $\mathbf{X}^* = (5, 2, 1, 5, 3)$, then $D = \{1, 2, 3, 5\}$. Calculate the probability that x_1 is an element of D. Then show that if n is large, this probability is approximately $1 - e^{-1}$.

(b) Let $\mathbf{x} = \{1, 2, \ldots, 100\}$. Use R to simulate 10^4 bootstrap samples from $\mathbf{x}$. For each bootstrap sample, determine whether the sample contained the number 1. Then calculate the proportion of your bootstrap samples that contained 1, in order to approximate the probability from part (a).

AP2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a population with distribution F that has unknown mean μ and variance σ^2. Let $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ be a bootstrap sample from the ECDF for $\mathbf{X}$; that is, the conditional distribution of X_i^* is DUnif$(X_1, \ldots, X_n)$. Let $\bar{X}^*$ denote the sample mean of $\mathbf{X}^*$:

$$\bar{X}^* = \frac{X_1^* + \cdots + X_n^*}{n}$$

(a) Show that unconditionally, X_1^* has CDF F. *Hint:* To do so, compute the $P(X_1^* \leq x)$ by conditioning on the events that "$X_1^* = X_i$" for $1 \leq i \leq n$, and using the Law of Total Probability.

(b) Use part (a) to calculate $E[X_1^*]$ and $\text{Var}(X_1^*)$.

(c) Show that

$$E[\bar{X}^*|\mathbf{X}] = \bar{X} \qquad \text{and} \qquad \text{Var}(\bar{X}^*|\mathbf{X}) = \frac{1}{n^2}\sum(X_i - \bar{X})^2 = \frac{\hat{\sigma}^2}{n}.$$

(d) Note that both $E[\bar{X}^*|\mathbf{X}]$ and $\text{Var}(\bar{X}^*|\mathbf{X})$ are **statistics**, since they are both functions of the random sample $\mathbf{X}$. Moreover, $\text{Var}(\bar{X}^*|\mathbf{X})$ can be used as an estimator of the variance of the sample mean $\bar{X}$. Show that $\text{Var}(\bar{X}^*|\mathbf{X})$ is a *biased* estimator.

Wednesday 3/14

Additional Problems

AP3. In this problem, we investigate the bootstrap distribution of the median. Run the following code in R to create four samples (of sizes 14, 15, 1000, 1001), each from a Normal population with mean 0 and variance 1. The samples of size 14 and 15 differ by the inclusion of a single value, and similarly, the samples of size 1000 and 1001 differ by the inclusion of a single value.

```
set.seed(316)
sample_14 <- rnorm(14)
sample_15 <- c(sample_14, rnorm(1))
sample_1000 <- rnorm(1000)
sample_1001 <- c(sample_1000, rnorm(1))
```

(a) Create 10^4 bootstrap samples from `sample_14`, as well as 10^4 bootstrap samples from `sample_15`. For each bootstrap sample, compute the median of the sample.

(b) Use R to create two histograms. The first histogram should be of the bootstrap distribution of medians generated using `sample_14`. The second histogram should be of the bootstrap distribution of medians generated using `sample_15`. Comment on similarities and differences between the two histograms.

(c) Now create 10^4 bootstrap samples from `sample_1000`, as well as 10^4 bootstrap samples from `sample_1001`. For each bootstrap sample, compute the median of the sample.

(d) Use R to create two histograms, one for the bootstrap distribution using `sample_1000` and the other using `sample_1001`. Comment on similarities and differences between the two histograms.

(e) Compare your answers to part (b) and part(d). What similarities and differences do you notice based on the relative sample size?

(f) Why does the parity of the sample size (i.e. whether it is even or odd) have an effect on the bootstrap distribution?

AP4. The skewness of a random variable X with mean μ and variance σ^2 is defined as

$$\mu_3 = \text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

and is one measurement of the asymmetry of the distribution. Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ are an iid sample with common CDF F. The sample skewness is the statistic

$$M_3(\mathbf{X}) = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^{3/2}}$$

It is reasonable to use M_3 as the estimator of the skewness μ_3 of a random variable with CDF F. In order to assess the quality of this estimator, we need to approximate the bias and variance of this estimator.

(a) Show that if $\mathbf{x}$ are the observed values of the sample, then skewness estimate $M_3(\mathbf{x})$ is equal to the skewness of a random variable which has the DUnif$(x_1, \ldots, x_n)$ distribution.

(b) Write a function in R which will take a vector `x` as input and output the value of the skewness $M_3(\mathbf{x})$ for this vector.

(c) The following sample of size 20 was generated from a skewed distribution with unknown CDF F. Copy-and-paste the following to load the data into R.

```
x <- c(6, 7, 4, 4, 4, 4, 4, 5, 9, 4, 5, 3, 5, 7, 2, 5, 6, 4, 4, 2)
```

Generate 5000 bootstrap samples from `x`. For each sample, compute the sample skewness. Then use these 5000 bootstrap statistics to estimate the **bias** and **standard deviation** of the sample skewness estimator M_3.

Friday 3/17

Additional Problems

AP5. Suppose $X_1, \ldots, X_n$ is an iid sample from $N(\mu, \sigma^2)$ and let $Y_i = e^{X_i}$. The variables $Y_1, \ldots, Y_n$ are said to have the **log-Normal** distribution with parameters μ and σ^2.

By the Central Limit Theorem, the distribution of $\bar{Y}$ should be approximately Normal, if n is large. But the CLT does not specify exactly how large is "large".

Run the following code in R to create a sample $\mathbf{Y}$ of size 25 from the log-Normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

```
set.seed(271828)
y <- exp(rnorm(25))
```

(a) Compute the mean $\bar{y}$ of this sample.

(b) Assume that $\bar{Y}$ is approximately Normal, and create a 95% confidence interval for μ using the techniques of Section 8.5.

(c) Now, use R to simulate 10^4 bootstrap samples from $\mathbf{y}$ and compute the sample mean of each.

(d) Use R to create a histogram of these sample means. Does the histogram appear Normal?

(e) Compute the 95% bootstrap percentile interval for μ.

(f) Compare the upper and lower bounds of the interval in part (b) to those of the interval in part (e). What do you notice?

AP6. In a Dec. 10, 2020 study of the efficacy of the Pfizer vaccine, a total of 43,448 participants received either a vaccine (21720) or a placebo (21728). Among the vaccinated participants, 8 developed cases of Covid-19, while among the placebo (control) group, 162 developed cases of Covid-19. Researchers are interested in estimating the efficacy e of the COVID vaccine, defined as

$$e = 1 - \frac{p_v}{p_c} = \frac{p_c - p_v}{p_c}$$

where p_v is the probability of contracting Covid-19 with the vaccine, while p_c is the probability of contracting Covid-19 without the vaccine. Efficacy indicates the proportionate reduction in disease in the vaccinated group. (For reference, the **relative risk** of the treatment is defined as $R = \frac{p_c}{p_v}$ with $e = 1 - R^{-1}$, and gives the rate of cases in the unvaccinated population per case in the vaccinated population.)

Let $\mathbf{C}$ represent the control sample, where $C_i \sim \text{Bern}(p_c)$ for $1 \leq i \leq 21{,}728$, and let $\mathbf{V}$ represent the vaccinated sample, where $V_i \sim \text{Bern}(p_v)$ for $1 \leq i \leq 21{,}720$. The MLE estimators for p_c and p_v are:

$$\hat{p}_c = \frac{1}{21728}\sum C_i \qquad \hat{p}_v = \frac{1}{21720}\sum V_i$$

We estimate e using the sample efficacy

$$\hat{e} = \frac{\hat{p}_c - \hat{p}_v}{\hat{p}_c}.$$

a. Compute the estimate $\hat{e}$ for this sample.

b. Explain why generating a single bootstrap sample from the control group is equivalent to generating 1 value from a $\text{Bin}(n = 21728, p = \hat{p}_c)$ distribution.

c. Use the `rbinom` function in R to generate 10^4 bootstrap estimates for $(\hat{p}_c^*, \hat{p}_v^*)$. Then compute bootstrap estimates for $\hat{e}^*$ based on each of these 10^4 pairs of bootstrap estimates.

d. Create a histogram of the bootstrap distribution of $\hat{e}^*$ and describe the shape, center and spread of the distribution.

e. Estimate the bias of $\hat{e}$.

f. Create a 95% bootstrap percentile confidence interval for e.

Homework 7: 3/13 - 3/17

STA 336 Due 11:59pm Wednesday, April 5 Name:

General Rubric

Points	Criteria
5	The solution is correct **and** well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information)
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.