This problem set covers material from Week 8, dates 4/09- 4/12. Unless otherwise noted, all problems are taken from the textbook. Problems can be found at the end of the corresponding subsection.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Canvas assignment. 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 general rubric for homework problems appears on the final page of this assignment.

Tuesday 4/09

None!

Thursday 4/11

1. 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, we have the conditional distribution $X_i^* | \mathbf{X} \overset{\text{iid}}{\sim} \text{DUnif}\{X_1, \ldots, X_n\}$. 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 "$X_i^* = X_i$" for $1 \leq i \leq n$, and using the Law of Total Probability.

$$\begin{aligned} P(X_i^* \leq x) &= \sum_{i=1}^{n} P(X_i^* \leq x | X_i^* = X_i) P(X_i^* = X_i) \qquad (LoTP) \\ &= \frac{1}{n} \sum_{i=1}^{n} P(X_i^* \leq x | X_i^* = X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} P(X_i \leq x) \\ &= \frac{1}{n} \sum_{i=1}^{n} F(x) = F(x) \end{aligned}$$

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

   Since CDFs uniquely identify a distribution, unconditionally X_1^* has same distribution as X_1. So the mean and variance are μ and σ^2.

(c) Show that

$$\mathbb{E}[\bar{X}^*|\mathbf{X}] = \bar{X} \qquad \text{and} \qquad \text{Var}(\bar{X}^*|\mathbf{X}) = \frac{\hat{\sigma}^2}{n},$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

$$\begin{aligned}
\mathbb{E}[\bar{X}^*|\boldsymbol{X}] &= \frac{1}{n}\mathbb{E}\left[X_1^* + \ldots + X_n^*|\boldsymbol{X}\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^*|\boldsymbol{X}] \\
&= \frac{1}{n}\sum_{i=1}^n \bar{X} = \bar{X} \qquad \textit{since } X_i^*|\boldsymbol{X} \sim \textit{DUnif}\{X_1, \ldots, X_n\}
\end{aligned}$$

$$\begin{aligned}
\textit{Var}(\bar{X}^*|\boldsymbol{X}) &= \frac{1}{n^2}\textit{Var}\left(X_1^* + \ldots + X_n^*|\boldsymbol{X}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \textit{Var}(X_i^*|\boldsymbol{X}) \qquad \textit{since cond. ind.} \\
&= \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}
\end{aligned}$$

(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.
We want to show that $\mathbb{E}[\text{Var}(\bar{X}^*|\boldsymbol{X})] \neq \textit{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

$$\begin{aligned}
\mathbb{E}[\text{Var}(\bar{X}^*|\boldsymbol{X})] &= \mathbb{E}\left[\frac{\hat{\sigma}^2}{n}\right] = \frac{1}{n}\mathbb{E}[\hat{\sigma}^2] \\
&= \frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n^2}(n\sigma^2 - n\mathbb{E}[(\bar{X} - \mu)^2]) \\
&= \frac{1}{n}(\sigma^2 - \textit{Var}(\bar{X})) \\
&= \frac{1}{n}(\sigma^2 - \frac{\sigma^2}{n}) \\
&= \sigma^2\frac{n-1}{n^2} \neq \frac{\sigma^2}{n} = \textit{Var}(\bar{X})
\end{aligned}$$

2. 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 the skewness estimate $M_3(\mathbf{x})$ is equal to the skewness of a random variable Y which has the $\text{DUnif}\{x_1, \ldots, x_n\}$ distribution.

We need to show that if $Y \sim \text{DUnif}\{x_1, \ldots, x_n\}$, then the skewness of Y μ_3 equals $M_3(\boldsymbol{x})$. Note that $\mathbb{E}[Y] = \bar{x}$ and $Var(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \sum_{i=1}^n (x_i - \bar{x})^2 P(Y = x_i) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ and by LoTUS,

$$\mathbb{E}[(Y - \mathbb{E}[Y])^3] = \sum_{i=1}^n (x_i - \bar{x})^3 P(Y = x_i) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3$$

$$\begin{aligned}
\mu_3 = Skew(Y) &= \mathbb{E}\left[\left(\frac{Y - \mathbb{E}[Y]}{\sqrt{Var(Y)}}\right)^3\right] \\
&= \left(\frac{1}{Var(Y)}\right)^{3/2} \mathbb{E}\left[(Y - \mathbb{E}[Y])^3\right] \qquad \textit{since variance constant} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}} \\
&= M_3(\boldsymbol{x})
\end{aligned}$$

(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 4/12

3. 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 $C_i|p_c \overset{\text{iid}}{\sim} \text{Bern}(p_c)$ for $1 \leq i \leq 21{,}728$, and let $\vec{V}$ represent the vaccinated sample, where $V_i|p_v \overset{\text{iid}}{\sim} \text{Bern}(p_v)$ 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 10000 bootstrap estimates for $(\hat{p}_c^*, \hat{p}_v^*)$. Then compute bootstrap estimates for $\hat{e}^*$ based on each of these 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) Create a 95% bootstrap confidence interval for e.

4. Let's investigate the parametric versus nonparametric bootstrap in a specific context using a simulation study. In R, conduct a simulation study that examines the coverage rates of a 95% nonparametric bootstrap interval and a 95% parametric bootstrap interval.

   (a) Let $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \text{Exp}(3)$ data. In R, you should do the following:

      i. Write a function that takes in two inputs: a vector of data `x` and a number of bootstrap simulations `B`. The function should return a vector of the lower and upper bounds of a 95% **nonparametric** bootstrap interval for the rate of the Exponential distribution based on the data `x`.

ii. Write a function that takes in two inputs: a vector of data `x` and a number of bootstrap simulations `B`. The function should return a vector of the lower and upper bounds of a 95% **parametric** bootstrap interval for the rate of the Exponential distribution based on the data `x`.

iii. Set a seed!

iv. Repeat the following $nsim = 1000$ times:

- Generate a *new* set of data $X_1, \ldots, X_{20} \overset{\text{iid}}{\sim} \text{Exp}(3)$ and obtain the two interval estimates with $B = 1500$ bootstrap samples.
- For each set of data and interval, keep track of/record/store whether or not the interval captures the true rate parameter. *Note: this step will take a few seconds to run. I recommend starting with fewer than 1000 iterations just to get your code running.*

v. Obtain the empirical coverage rate of each interval by calculating the proportion of times each interval captures the true rate parameter.

Briefly comment on how the empirical coverage rates compare to each other, as well as how the coverage rates compare to the desired (aka "nominal") coverage rate.

(b) Repeat steps (a)(iii - v) but now with samples of size $n = 100$. Briefly comment on how the coverage rates compare to each other, how the empirical coverage rates compare to the desired (aka "nominal") coverage rate, and how the sample size n affects coverage rates. Based on your findings, also comment on when you might recommend using the parametric versus the nonparametric bootstrap, and why.

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.
Notes:	For problems with code, well-written means only having lines of code that are necessary to solving the problem, as well as presenting the solution for the reader to easily see. It might also be worth adding comments to your code.