1. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(0, \sigma^2)$ population distribution where $\sigma^2 > 0$ is unknown.
(a) Argue

$$Q_1 = \sum_{i=1}^{n} \left(\frac{Y_i}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} Y_i^2$$

is a pivotal quantity and use Q_1 to derive a $100(1-\alpha)\%$ confidence interval for σ^2. Define any notation that you use.
(b) We could also use

$$Q_2 = \frac{(n-1)S^2}{\sigma^2}$$

as a pivotal quantity and derive a $100(1-\alpha)\%$ confidence interval for σ^2 using Q_2. Show how this would be done.
(c) Calculate the expected length of each interval in parts (a) and (b). Can you determine which interval has a smaller expected length?

2. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample with population pdf

$$f_Y(y) = \begin{cases} \frac{2y}{\theta^2}, & 0 < y < \theta \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is unknown.
(a) Show that

$$\widehat{\theta} = \frac{3\overline{Y}}{2}$$

is an unbiased estimator of θ.
(b) Find $V(\widehat{\theta})$. Graph $V(\widehat{\theta})$ versus θ for $\theta > 0$ and $n = 10$.
(c) Find another unbiased estimator of θ that is a function of the maximum order statistic $Y_{(n)}$. Does this unbiased estimator have a smaller variance than the unbiased estimator in part (a)?

3. Suppose Y is a discrete random variable with pmf

y	0	1	2
$p_Y(y)$	p	$3p$	$1-4p$

Treating this as a population pmf, suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $p_Y(y)$. The parameter $0 < p < 1/4$ is unknown.
(a) Show that

$$\widehat{p} = \frac{2 - \overline{Y}}{5}$$

is an unbiased estimator of p.
(b) Show the variance of $\widehat{p}$ is

$$V(\widehat{p}) = \frac{7p - 25p^2}{25n}.$$

(c) How would you estimate the standard error of $\widehat{p}$?

4. In biological applications, it is common to assume a $\mathcal{N}(\theta, \theta^2)$ population distribution for a continuous response (e.g., growth, distance traveled, hormone concentrations, etc.). In this population distribution, the variance is the square of the mean. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\theta, \theta^2)$ population distribution. Consider the two point estimators

$$\widehat{\theta}_1 = \overline{Y} \qquad \text{and} \qquad \widehat{\theta}_2 = cS,$$

where $\overline{Y}$ and S denote the sample mean and sample standard deviation, respectively, and

$$c = \frac{\sqrt{n-1}\,\Gamma(\frac{n-1}{2})}{\sqrt{2}\,\Gamma(\frac{n}{2})}.$$

Both are unbiased estimators of θ.
(a) Suppose $a \in (0,1)$. Show $\widehat{\theta} = a\widehat{\theta}_1 + (1-a)\widehat{\theta}_2$ is also an unbiased of estimator of θ.
(b) Find the value of a that minimizes $V(\widehat{\theta})$.

5. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Pareto population pdf

$$f_Y(y) = \begin{cases} \dfrac{\theta}{y^{\theta+1}}, & y > 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 2$ is unknown.
(a) The sample mean $\overline{Y}$ is an unbiased estimator of what function of θ?
(b) The sample variance S^2 is an unbiased estimator of what function of θ?

6. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson(θ) population distribution, where $\theta > 0$ is unknown.
(a) Explain why both $\overline{Y}$ and S^2 are unbiased estimators of θ.
(b) Find an unbiased estimator of $\tau(\theta) = \theta^2$. *Hint:* Work with $\overline{Y}$.
(c) Which "theorem" from Chapter 7 guarantees

$$\frac{\overline{Y} - \theta}{\sqrt{\theta/n}} \xrightarrow{d} \mathcal{N}(0,1), \text{ as } n \to \infty?$$

(d) In part (c), estimate the standard error $\sqrt{\theta/n}$ with $\sqrt{\overline{Y}/n}$. It follows that

$$Q_n = \frac{\overline{Y} - \theta}{\sqrt{\overline{Y}/n}} \sim \mathcal{AN}(0,1),$$

when n is large; i.e., Q_n is a large-sample pivot. Use Q_n to derive a large-sample $100(1-\alpha)\%$ confidence interval for θ.

7. Suppose we have two independent random samples:

- $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an iid sample from a $\mathcal{N}(\mu_1, \sigma^2)$ population distribution
- $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an iid sample from a $\mathcal{N}(\mu_2, \sigma^2)$ population distribution.

All parameters are unknown, but note the variance in each population distribution is the same. Let S_1^2 and S_2^2 denote the sample variances from Sample 1 and Sample 2, respectively. Recall the pooled sample variance estimator is

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

(a) Prove

$$Q = \frac{(n_1+n_2-2)S_p^2}{\sigma^2}$$

is a pivotal quantity and give its distribution.
(b) Use the pivot in part (a) to derive a $100(1-\alpha)\%$ confidence interval for σ^2. Define any notation you use.
(c) Derive a $100(1-\alpha)\%$ confidence interval for

$$\theta = a_1\mu_1 + a_2\mu_2.$$

Note that if $a_1 = 1$ and $a_2 = -1$, then $\theta = a_1\mu_1 + a_2\mu_2$ reduces to $\mu_1 - \mu_2$. Therefore, this derivation is more general than what we did in class (but similar in spirit).

8. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with pdf

$$f_Y(y) = \begin{cases} \frac{\alpha_0}{\beta} y^{\alpha_0 - 1}\exp(-y^{\alpha_0}/\beta), & y > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\alpha_0 > 0$ is known and $\beta > 0$ is unknown.
(a) Prove that $Y_i^{\alpha_0} \sim$ exponential(β). *Hint:* Use the transformation method.
(b) Argue $Q = 2T/\beta \sim \chi^2(2n)$, where

$$T = \sum_{i=1}^{n} Y_i^{\alpha_0}.$$

(c) Use the result in part (b) to derive a $100(1-\alpha)\%$ confidence interval for β. Define all notation.

9. Highly active antiretroviral therapy (HAART) refers to the administration of aggressive treatment regimens used to suppress HIV viral replication and to delay the onset of AIDS. HAART is a combination therapy that restores CD4+ T cell numbers in HIV-infected patients. While HAART improves the prognosis of HIV-infected patients, there has been a recent concern that taking HAART may increase the risk of cardiovascular disease. To examine this question, suppose a clinical trial is to be performed with n_1 patients receiving a HAART regimen with a cardiovascular disease drug supplement (Enalapril) and n_2 receiving HAART but no Enalapril.

Assume that patients are randomly assigned to the treatment groups:

- Group 1: HAART with Enalapril
- Group 2: HAART with no Eanlapril.

Patients will then be monitored for the development of cardiovascular disease (CVD). Define

$$\begin{aligned} Y_1 &= \text{number of Group 1 patients who develop CVD} \\ Y_2 &= \text{number of Group 2 patients who develop CVD.} \end{aligned}$$

Assume $Y_1 \sim b(n_1, p_1)$, $Y_2 \sim b(n_2, p_2)$, and Y_1 and Y_2 are independent. The investigators are interested in comparing the binomial probabilities p_1 and p_2.
(a) Prove that

$$\widehat{\theta} = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$$

is an unbiased estimator of $\theta = p_1 - p_2$.
(b) Prove that the variance of $\widehat{\theta}$ is given by

$$V(\widehat{\theta}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

(c) The Central Limit Theorem guarantees $\widehat{\theta}$ is approximately normally distributed when n_1 and n_2 are both large. If $n_1 = n_2 = 100$, $y_1 = 30$, and $y_2 = 20$, compute a 95% confidence interval for θ. Interpret the interval in the context of how the two treatment groups compare.

10. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Rayleigh population pdf

$$f_Y(y) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is unknown.
(a) Find a function of $\overline{Y}$ that is an unbiased estimator of θ.
(b) Find a function of $Y_{(1)}$ that is an unbiased estimator of θ.

11. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma population distribution with shape parameter $\alpha = 2$ and unknown scale parameter $\beta > 0$.
(a) Argue that

$$Q_n = \frac{\overline{Y} - 2\beta}{\sqrt{2\beta^2/n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. Therefore, Q_n is an approximate (i.e., large-sample) pivot.
(b) Use the result in part (a) to show that

$$\left(\frac{\overline{Y}}{2 + z_{\alpha/2}\sqrt{2/n}}, \ \frac{\overline{Y}}{2 - z_{\alpha/2}\sqrt{2/n}} \right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for β. The symbol $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from a $\mathcal{N}(0,1)$ distribution.

12. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with pdf

$$f_Y(y) = \begin{cases} e^{-(y-\theta)}, & y > \theta \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is unknown.

(a) Find an unbiased estimator that is a function of the sample mean $\overline{Y}$.
(b) Find an unbiased estimator that is a function of the minimum order statistic $Y_{(1)}$.
(c) Compare the variances of your two unbiased estimators. Which estimator would you prefer?

13. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with pdf

$$f_Y(y) = \begin{cases} \theta^2 y e^{-\theta y}, & y > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is unknown. Find a function of $\overline{Y}$ that is an unbiased estimator of θ.

14. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) population, where $0 < p < 1$, and let $\widehat{p}$ denote the usual sample proportion. In class, we showed that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is a large-sample $100(1-\alpha)\%$ confidence interval for p. The symbol $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from a $\mathcal{N}(0,1)$ distribution. An alternative large-sample $100(1-\alpha)\%$ confidence interval for p can be constructed by using the following large-sample result:

$$h(\widehat{p}) = \sin^{-1}(\widehat{p}^{1/2}) \ \sim \mathcal{AN}\left(\sin^{-1}(p^{1/2}),\ \frac{1}{4n}\right).$$

(a) Use this fact to construct a large-sample $100(1-\alpha)\%$ confidence interval for

$$h(p) = \sin^{-1}(p^{1/2}).$$

Hint: Consider

$$Q_n = \frac{h(\widehat{p}) - h(p)}{\sqrt{1/4n}}.$$

(b) Transform the endpoints of the interval in (a) to produce a large-sample $100(1-\alpha)\%$ confidence interval for the parameter p. You do this by finding the inverse function $h^{-1}(p)$ and applying this inverse rule to the endpoints of the $h(p)$ interval.

15. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from the population pdf

$$f_Y(y) = \begin{cases} \dfrac{\theta}{y^2}, & y \geq \theta \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is unknown.
(a) Show that pdf of the minimum order statistic $Y_{(1)}$ is given by

$$f_{Y_{(1)}}(y) = \begin{cases} \dfrac{n\theta^n}{y^{n+1}}, & y \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

(b) Show that

$$Q = \frac{\theta}{Y_{(1)}}$$

is a pivotal quantity and use it to derive a $100(1-\alpha)\%$ confidence interval for θ.

16. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma population distribution with shape parameter $\alpha = 2$ and scale parameter θ, where $\theta > 0$ is unknown. That is, the population probability density function (pdf) is

$$f_Y(y) = \begin{cases} \frac{y}{\theta^2} e^{-y/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find a function of the sample mean $\overline{Y}$ which is an unbiased estimator of θ. Also, calculate the variance of your unbiased estimator.
(b) Find two unbiased estimators of $\tau(\theta) = \theta^2$.

17. Suppose $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both μ and σ^2 are unknown. Our goal is to estimate σ^2 using estimators of the form

$$\widehat{\sigma}_c^2 = c\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

where $c > 0$ is a constant (free of μ and σ^2). Note that when $c = 1/(n-1)$, the estimator $\widehat{\sigma}_c^2$ is our usual sample variance S^2 (which is unbiased for σ^2).
(a) When viewed as a function of c, show MSE$(\widehat{\sigma}_c^2)$ is minimized when

$$c = \frac{1}{n+1}.$$

(b) For this part only, suppose μ is known; e.g., $\mu = \mu_0$, and suppose the goal is to write a $1-\alpha$ interval estimator for σ^2. Show that

$$Q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_0)^2$$

is a pivotal quantity and use it to derive the interval. Define all notation as needed.

18. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with probability density function (pdf)

$$f_Y(y) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The population parameter $\theta > 0$ is unknown.
(a) The sample mean $\overline{Y}$ is an unbiased estimator of what function of θ?
(b) Show that

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

is an unbiased estimator of θ.
(c) Derive the standard error of $\widehat{\theta}$.
(d) How would you *estimate* the standard error of $\widehat{\theta}$ in part (b)? Is your estimated standard error an unbiased estimator of the standard error of $\widehat{\theta}$? Explain.

19. SEIR models are used by epidemiologists to describe covid-19 disease severity in a population. The model consists of four different categories:

$$\begin{aligned} \text{S} &= \text{susceptible category} \\ \text{E} &= \text{exposed category} \\ \text{I} &= \text{infected category} \\ \text{R} &= \text{recovered category.} \end{aligned}$$

The four categories are mutually exclusive and exhaustive among living individuals (SEIRD models do include a fifth category for those who have died from disease). A random sample of n individuals is selected from a population (e.g., residents of Richland County) and the category status of each individual is identified. This produces the multinomial random vector

$$\mathbf{Y} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \quad \text{and} \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

The random variables Y_1, Y_2, Y_3, Y_4 record the number of individuals identified in the susceptible, exposed, infected, and recovered categories, respectively. The parameter vector $\mathbf{p}$ is unknown.
(a) Define the sample proportion in the jth category to be

$$\widehat{p}_j = \frac{Y_j}{n},$$

for $j = 1, 2, 3, 4$. Show that $\widehat{p}_1 + \widehat{p}_2$ is an unbiased estimator of $p_1 + p_2$. Note that $p_1 + p_2$ is the (population) proportion of individuals who have not yet contacted covid-19.
(b) The Central Limit Theorem can be used to show the estimator $\widehat{p}_1 + \widehat{p}_2$ is approximately normal with mean $p_1 + p_2$.

1. Derive the variance of $\widehat{p}_1 + \widehat{p}_2$.
2. Suggest a large-sample pivot and then use it to derive the form of a $1-\alpha$ interval estimator for $p_1 + p_2$.

20. If $X \sim \mathcal{N}(0, \sigma^2)$, then $Y = |X|$ follows a half-normal distribution; i.e., the probability density function (pdf) of Y is

$$f_Y(y) = \begin{cases} \dfrac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\dfrac{y^2}{2\sigma^2}\right), & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y)$. The population parameter $\sigma^2 > 0$ is unknown. Note that although σ^2 is the variance of X, it is not the variance of Y.
(a) Show that

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2$$

is an unbiased estimator of σ^2.

(b) Show that

$$V(\widehat{\sigma}^2) = \frac{2\sigma^4}{n}.$$

(c) What is the standard error of $\widehat{\sigma}^2$?
(d) Suggest a point estimator for the standard error. Is your point estimator an unbiased estimator of the standard error?

21. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2$. Both μ and σ^2 are unknown. To estimate μ, we will consider the point estimator $\widehat{\mu}$ which is defined as the value of μ that minimizes

$$Q(\mu) = \sum_{i=1}^{n}(Y_i - \mu)^2 + \lambda\mu^2,$$

where $\lambda \geq 0$ is a known constant.
(a) Show that

$$\widehat{\mu} = \frac{n\overline{Y}}{n + \lambda}.$$

(b) For the point estimator in part (a), show that letting $\lambda = 0$ gives an unbiased estimator of μ.
(c) When $\lambda > 0$, the point estimator in part (a) is biased. Calculate the bias $B(\widehat{\mu})$ and the mean squared error $\text{MSE}(\widehat{\mu})$.
(d) Show there is a value of $\lambda > 0$ that yields

$$\text{MSE}(\widehat{\mu}) < \text{MSE}(\overline{Y}).$$

This is a special instance of the "variance-bias tradeoff" in statistical learning. We might be willing to accept a small amount of bias if doing so provides a larger reduction in variability. *Hint:* What value of λ makes $\text{MSE}(\widehat{\mu})$ as small as possible?

22. Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(β) population distribution where $\beta > 0$ is an unknown parameter. In class, we showed

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}}\right),$$

where $T = \sum_{i=1}^{n} Y_i$ is the sample sum, is a $100(1-\alpha)\%$ confidence interval for β. Recall this interval was derived by starting with the pivotal quantity

$$Q_1 = \frac{2T}{\beta} \sim \chi^2(2n).$$

Pivots are not unique as this problem illustrates. Instead of working with the sample sum, suppose we decided to work with the minimum order statistic $Y_{(1)}$. Recall that

$$Y_1, Y_2, ..., Y_n \sim \text{iid exponential}(\beta) \implies Y_{(1)} \sim \text{exponential}(\beta/n).$$

(a) In one sentence, explain why $Y_{(1)}$ is not a pivot.
(b) Use a moment generating function argument to show

$$Q_2 = \frac{2nY_{(1)}}{\beta} \sim \chi^2(2).$$

(c) Letting $\chi^2_{2,1-\alpha/2}$ and $\chi^2_{2,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of the $\chi^2(2)$ distribution, respectively, use Q_2 to derive a different $100(1-\alpha)\%$ confidence interval for β.

23. In sampling plots of land for counts of an animal species, we assume the number of animals observed in each plot X has a Poisson distribution with mean $\lambda > 0$. The population level parameter λ is unknown. Suppose $X_1, X_2, ..., X_n$ represent the animal counts from a random sample of n plots in a large area. Assume $X_1, X_2, ..., X_n$ are iid from a Poisson(λ) population distribution.
(a) Argue that

$$Q_n = \frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, and hence

$$P\left(-1.96 < \frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} < 1.96\right) \approx 0.95,$$

for large n. A large-sample 95% confidence interval for λ would therefore be the values of λ that satisfy

$$-1.96 < \frac{\overline{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} < 1.96.$$

(b) Suppose that instead of recording the count X on each plot, we merely observe whether at least one animal is present on the plot. In other words, we observe

$$Y = \begin{cases} 1, & X > 0 \\ 0, & X = 0. \end{cases}$$

Show that $P(Y = 1) = 1 - e^{-\lambda}$.
(c) In part (b), suppose $Y_1, Y_2, ..., Y_n$ represent the "success/failure" outcomes observed on the n mutually independent plots. Here, it is understood that

- "success" $\Longrightarrow$ at least one animal observed $(Y = 1)$
- "failure" $\Longrightarrow$ no animals observed $(Y = 0)$.

First write a large-sample 95% confidence interval for

$$p = P(Y = 1) = 1 - e^{-\lambda}.$$

Then, using straightforward algebra, derive a different large-sample 95% confidence interval for λ, that is, different from the one in part (a).
(d) Which interval, the one in part (a) or the one in part (c), do you think is more informative?