Unit 5 Classwork

The goals of this assignment are to help you (1) use numerical simulation to analyze and interpret important statistical inference quantities, (2) calculate probabilities using `R`, and (3) estimate probabilities and other quantities through numerical (computer) simulation. Such simulations can be useful. We can use simulations to help confirm that we've calculated a probability "by hand" correctly, or to estimate other quantities, like areas/integrals.

Problem #1

Let $\hat{\theta}$ be an estimator of the parameter θ (e.g., we might think of $\hat{\theta} = \bar{X}$ and $\theta = \mu$, where μ is a population mean). We say that $\hat{\theta}$ is *unbiased* if $E(\hat{\theta}) = \theta$.

1.(a) Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Pois}(\lambda)$. Find an unbiased estimator for λ.

For a Poisson distribution with parameter λ, the mean and variance are both equal to λ. Therefore, we can consider the sample mean $\bar{X}$ as an unbiased estimator for λ.

The expectation of $\bar{X}$ is given by:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Since the expected value is a linear operator, we can move the $\frac{1}{n}$ outside the sum:

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)$$

Since the X_i are iid (independent and identically distributed), we have:

$$= \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$

Each X_i follows a Poisson distribution with parameter λ, so $E(X_i) = \lambda$. Substituting this in:

$$= \frac{1}{n} \sum_{i=1}^{n} \lambda$$

$$= \lambda$$

Therefore, $E(\bar{X}) = \lambda$, and $\bar{X}$ is an unbiased estimator for λ.

1.(b) Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the maximum likelihood estimator (MLE) for λ.

The likelihood function for the exponential distribution is given by:

$$L(\lambda; x_1, x_2, \ldots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The log-likelihood function is:

$$\ln L(\lambda; x_1, x_2, \ldots, x_n) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

To find the maximum likelihood estimator (MLE) for λ, we take the derivative of the log-likelihood with respect to λ and set it equal to zero:

$$\frac{d}{d\lambda} \ln L(\lambda; x_1, x_2, \ldots, x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Solving for λ:

$$\frac{n}{\lambda} = \sum_{i=1}^{n} x_i$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

Therefore, the maximum likelihood estimator (MLE) for λ is:

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{X}}$$

Problem #2

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a population with mean μ and variance σ^2.

2.(a) Show that $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of σ^2. This answers the question of why we divide by $n - 1$ in S^2!

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population with mean μ and variance σ^2.

The sample variance, denoted by S^2, is defined as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

To show that S^2 is an unbiased estimator of σ^2, we need to find the expected value of S^2:

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

Expand the sum:

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

Since X_i are i.i.d., we can rewrite this as:

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

Now, consider the term $E\left((X_i - \bar{X})^2\right)$. Expand it using the fact that $\bar{X}$ is the sample mean:

$$= \frac{1}{n-1}\sum_{i=1}^{n}E\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)$$

Now, use the linearity of expectation:

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(E(X_i^2) - 2E(X_i\bar{X}) + E(\bar{X}^2)\right)$$

The term $E(X_i)$ is μ and $E(\bar{X})$ is also μ. Also, $E(X_i^2)$ is equal to the variance plus the square of the mean, i.e., $\sigma^2 + \mu^2$. Substituting these values:

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2\right)$$

Simplify the expression:

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

Since the mean μ cancels out, we are left with:

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

Now, the sum of σ^2 repeated n times is $n\sigma^2$. Substituting this in:

$$= \frac{1}{n-1} \cdot n\sigma^2$$

The n in the numerator and denominator cancels out:

$$= \sigma^2$$

Therefore, $E(S^2) = \sigma^2$, and S^2 is an unbiased estimator of σ^2. This explains the use of $n-1$ in the denominator when calculating S^2.

2.(b)Assume that $E\left[\sqrt{X}\right] < \sqrt{E(X)}$. Show that S is a *biased* estimator of σ.

Assuming $E\left[\sqrt{X}\right] < \sqrt{E(X)}$, we want to show that the sample standard deviation S is a biased estimator of the population standard deviation σ.

The sample standard deviation is defined as:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

To check for bias, we compare $E(S)$ to σ. First, let's find $E(S)$:

$$E(S) = E\left(\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}\right)$$

Since the square root is a nonlinear function, we can't simply move it inside the sum. However, we can still examine the inequality in terms of expectations:

$$E(S) < E\left(\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}\right) < E(\sigma)$$

The inequality on the left side is true based on the assumption. Now, let's focus on the right side:

$$E\left(\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}\right) < \sqrt{E\left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\right)} = \sqrt{E(S^2)}$$

Now, we know that $E(S^2) = \sigma^2$ from the unbiasedness of S^2 that we showed in the

previous part. Therefore:

$$\sqrt{E(S^2)} = \sqrt{\sigma^2} = \sigma$$

Putting it all together:

$$E(S) < \sigma$$

This implies that S is biased, as the expected value of S is less than the true standard deviation σ. Therefore, S is a biased estimator of σ under the given assumption.

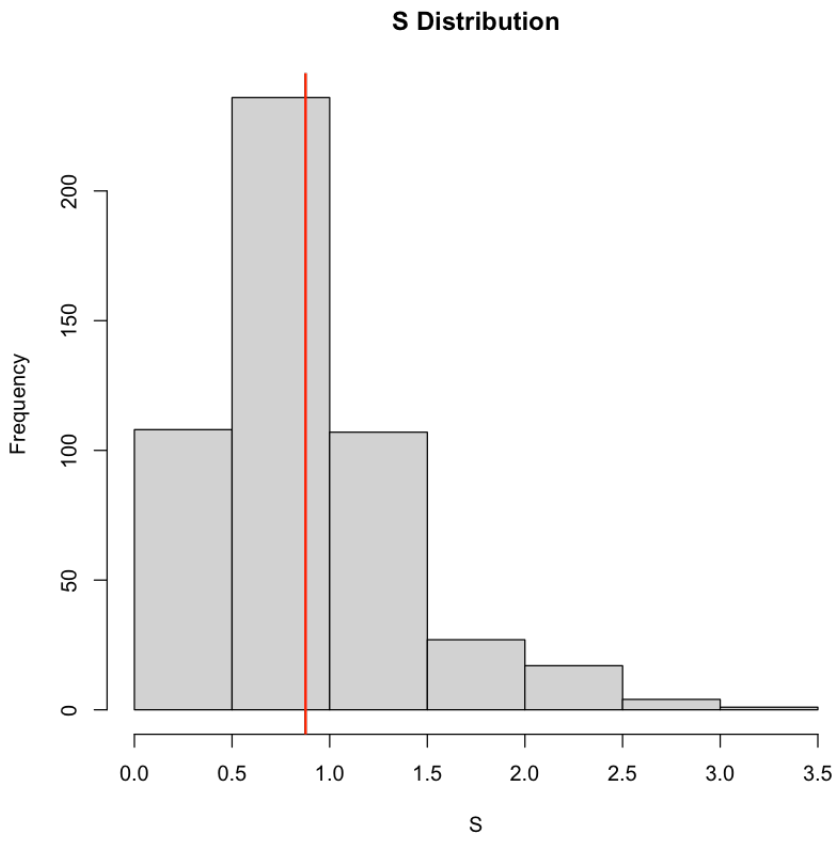
2.(c) Use simulations to provide some evidence that S is biased. Specifically:

1. Generate `m = 500` different samples of size `n = 5` from $\text{Exp}(\lambda = 1)$.
2. Calculate S for each sample.
3. What is the mean of the distribution of S?
4. Creat a histogram of the distribution of S. Comment on the distribution. How does this provide evidence that S is biased?

In [ ]:
```r
m <- 500
n <- 5
lambda <- 1

samples <- matrix(rexp(m * n, rate = lambda), nrow = m)
S <- apply(samples, 1, sd)
mean_S <- mean(S)

hist(S, main = "S Distribution", xlab = "S")
abline(v = mean_S, col = "red", lwd = 2)
```



The distribution is right-skewed so it suggests bias.

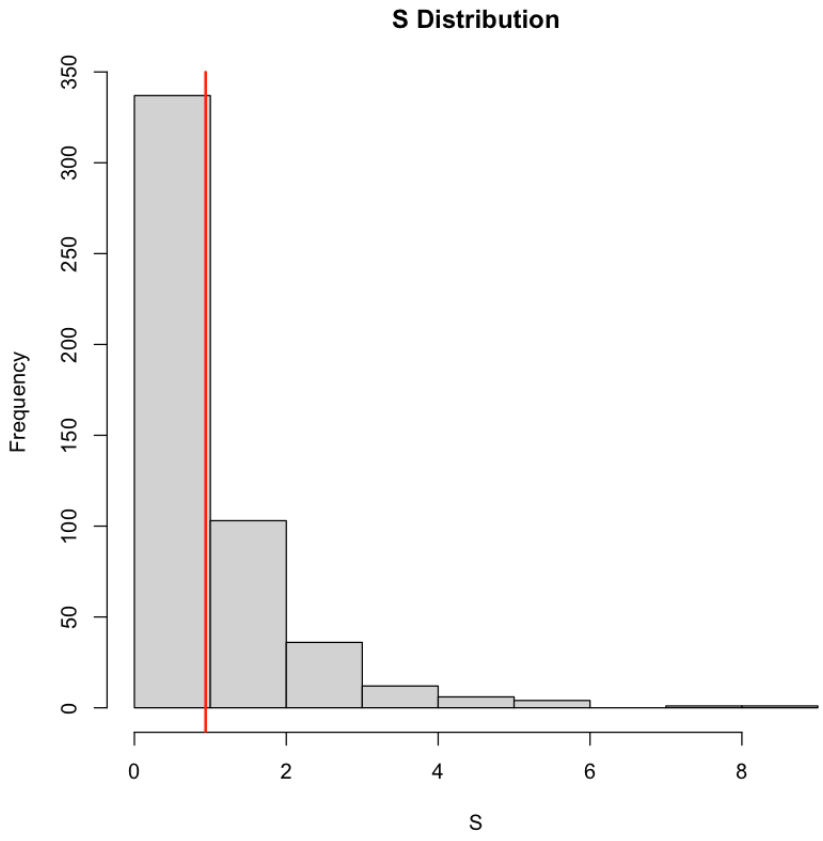
2.(d) Repeat the process described in (e) for S^2. What do you notice?

In [ ]:
```r
samples <- matrix(rexp(m * n, rate = lambda), nrow = m)

S2 <- apply(samples, 1, var)

mean_S2 <- mean(S2)

hist(S2, main = "S Distribution", xlab = "S")
abline(v = mean_S2, col = "red", lwd = 2)
```



The variance S^2 provides an unbiased estimate of the population variance.

Problem #3

A response time is normally distributed with standard deviation of 25 milliseconds. A new system has been installed, and we wish to estimate the true average response time, μ, for the new environment.

Assuming that the response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% confidence has a width of (at most) 10?

In [ ]:
```r
Z <- qnorm(0.975)
sigma <- 25
E <- 10

n <- ceiling((Z * sigma / E)^2)

cat("The required sample size is:", n)
```

```
The required sample size is: 25
```

Problem #4

The dataset contained in the file `winequality-red.csv` is related to a red Portuguese "Vinho Verde" wine. Make sure you download it from Canvas, and load it with the following code:

In [ ]:
```
wine = read.table("winequality-red.csv", header = TRUE)
head(wine)
```

A data.frame: 6 x 2

	X	pH
	<int>	<dbl>
1	1	3.51
2	2	3.20
3	3	3.26
4	4	3.16
5	5	3.51
6	6	3.51

Calculate a 90% confidence interval for the mean pH level.

In [ ]:
```
pH = wine$pH

confidence_interval <- t.test(pH,conf.level = 0.90)$conf.int

confidence_interval
```

3.3047589487973 · 3.31746744269738

Problem #5

In this example, we will construct a simulation to verify the “coverage properties” of a confidence interval for the mean of a normal distribution.

5.(a) Simulate a matrix with `m = 1000` rows and `n = 100` columns, where each entry is a random number from the population model $N(0,1)$. Interpret each row as a sample from the population.

In [ ]:
```
m <- 1000
n <- 100

simulated_matrix <- matrix(rnorm(m * n, mean = 0, sd = 1), nrow = m, ncol =

head(simulated_matrix)
```

-0.84866386	-0.5160968	-0.78458925	-1.25981691	1.93721076	-0.009922802	-0.383
0.08649426	-1.6862307	0.08174494	0.08936628	0.01574647	-0.881245005	0.384
-0.12386187	0.8820048	1.07284079	-0.44667705	0.23982169	-0.672818807	-0.305
-0.53817811	-0.8781138	-1.85535597	-0.36049315	-0.91329031	1.166658637	0.469
-0.98173336	-1.0014345	-1.06007702	0.38577687	-0.03589847	0.119261765	-0.214
-1.01542317	0.9836105	-0.45411920	0.40661091	-1.27950967	1.526616947	0.555

5.(b) Suppose that we didn't know the population mean, μ, but wanted to estimate it using a confidence interval. For each sample, calculate the 95% confidence for the mean, μ. Assume $\sigma = 1$ is known.

In [ ]:
```r
calculate_t_test <- function(sample) {
  t_result <- t.test(sample,conf.level = 0.95)
  return(t_result$conf.int)
}

confidence_intervals <- apply(simulated_matrix, 1, calculate_t_test)

# View the first few confidence intervals
head(confidence_intervals)
```

-0.09898288	-0.03286238	-0.2165817	0.01353747	-0.1137901	-0.1623933	-0.1481076
0.31795958	0.32888093	0.1437217	0.43544794	0.2416739	0.2219562	0.2272974

5.(c) Why would we use a confidence interval instead of just reporting the sample mean $\bar{x}$?

A confidence interval provides a range of values rather than a single point estimate. This range accounts for the uncertainty inherent in estimating population parameters from a sample. It conveys how precise or uncertain the estimation is.

5.(d) Print and interpret the confidence interval for the first sample (i.e., when `m = 1`).

In [ ]:
```r
first_sample_conf_interval <- confidence_intervals[, 1]

first_sample_conf_interval
```

-0.0989828837916451 · 0.317959582756487

The 95% confidence interval for the mean of the first sample is -0.0989 and 0.3179. This means that if we take many samples and calculate confidence intervals in the same way,

we would expect approximately 95% of those intervals to contain the true mean.

5.(e) Justify why, in part (b), you can use critical values from the normal distribution instead of critical values from the t distribution.

In part (b), it was mentioned that we are assuming a known population standard deviation of 1. When the population standard deviation is known and the sample size is large, it is appropriate to use critical values from the normal distribution rather than the t-distribution.

5.(f) Calculate the proportion of confidence intervals that cover the true μ. Does it match what theory suggests? If it deviates from what theory suggests, explain why.

In [ ]:
```r
true_mean <- 0

mean_covered <- function(conf_interval) {
  return(true_mean >= conf_interval[1] && true_mean <= conf_interval[2])
}

covered_intervals <- apply(confidence_intervals, 2, mean_covered)
proportion_covered <- mean(covered_intervals)

cat("Proportion of Confidence Intervals Covering True Mean:", proportion_cov
```

```
Proportion of Confidence Intervals Covering True Mean: 0.944
```