DS-GA 1002 - Homework 7

Eric Niblock

October 25th, 2020

1. **(Review/Concept Checks)**

   (a) **Suppose $X_1, ..., X_5 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_2)$. Compute the probability that all measurements are within σ of μ:**

$$P(|X_i - \mu| < \sigma \textbf{ for } i = 1, ..., 5)$$

If we just consider one random variable $X \sim \mathcal{N}(\mu, \sigma_2)$, then we have that the probability that X lies within one standard deviation of the mean as,

$$P(|X - \mu| < \sigma) = 2\int_{\mu}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = 2(0.3413) = 0.6826 \qquad (1)$$

And since all $X_1, ..., X_5$ are iid, we have that,

$$P(|X_1 - \mu| < \sigma \cap ... \cap |X_5 - \mu| < \sigma) = (0.6826)^5 = 0.1483 \qquad (2)$$

   (b) **(Bias-Variance Decomposition) Suppose our data is drawn from a parametric model with parameter θ, and suppose that T is an estimator of θ. One method for measuring the quality of T is the expected square loss:**

$$L(\theta) = E[(T - \theta)^2]$$

Prove that $L(\theta)$ can be expressed in terms of the bias of T and the variance of T. [Hint: $(T - \theta)^2 = ((T - E[T]) + (E[T] - \theta))^2$].

It is clear that the bias of T and the variance of T can be expressed as,

$$
\begin{align}
Bias(T,\theta) &= E[T] - \theta \tag{3}\\
Var(T) &= E[T^2] - E[T]^2 \tag{4}
\end{align}
$$

Then, we simply have,

$$
\begin{aligned}
L(\theta) &= E[(T-\theta)^2] = E[((T-E[T]) + (E[T]-\theta))^2]\\
&= E[(T-E[T])^2 + 2(T-E[T])(E[T]-\theta) + (E[T]-\theta)^2]\\
&= E[T^2 + E[T]^2] + E[2(-T\theta - E[T]^2 + E[T]\theta)] + E[(E[T]-\theta)^2]\\
&= E[T^2] + E[T]^2 - 2E[T\theta] - 2E[T]^2 + 2\theta E[T] + (Bias(T,\theta))^2\\
&= E[T^2] - E[T]^2 + (Bias(T,\theta))^2\\
&= Var(T) + Bias^2(T,\theta)
\end{aligned} \tag{5}
$$

So it is true that we can express $L(\theta)$ as a sum of the variance of T and the bias of T squared.

(c) Let X be a random variable with corresponding PDF f_X. Suppose X has a symmetric distribution (i.e., $f_X(x) = f_X(-x)$ for all $x \in \mathbb{R}^n$). Prove that

$$P(|X| \leq a) = 2F_X(a) - 1$$

for all $a > 0$ where F_X is the CDF of X.

First, we begin by rewriting the above expression and simplifying,

$$
\begin{align}
P(-a \leq X \leq a) &= 2F_X(a) - 1 \tag{6}\\
F_X(a) - F_X(-a) &= 2F_X(a) - 1 \tag{7}\\
F_X(a) + F_X(-a) &= 1 \tag{8}\\
P(X \leq a) + P(X \leq -a) &= 1 \tag{9}
\end{align}
$$

But since X is symmetrically distributed about the origin, it is clear that $P(X \leq -a) = P(X \geq a)$. Then, we have,

$$P(X \leq a) + P(X \geq a) = 1 \tag{10}$$

Which is obviously true, since $P(-\infty \leq X \leq \infty) = 1$.

(d) **Two scientists in different labs repeatedly perform the same experiment (about 5 times each) to estimate the quantity of salt produced by a particular chemical reaction. Scientist 1 obtains a 95% confidence interval (in milligrams) [2, 3] whereas scientist 2 obtains a 95% confidence interval (in milligrams) [4, 5]. True or false: this situation is actually impossible since there cannot be a 95% chance the true amount of salt is both less than 3mg and larger than 4mg. Please include a short explanation of why you said true or false.**

False. The confidence interval of 95% implies that if we repeated this experiment many, many times, about 95% of the confidence intervals would contain the true amount of salt within their range. Therefore, these conflicting confidence intervals are able to coexist.

(e) **You have conducted a survey of 100 randomly selected NYU students. You ask each person surveyed for the number of students in the first class they take each week. You average together their 100 answers to get an estimate for the average class size at NYU (i.e., an estimate for the sum of all class sizes divided by the total number of classes). Is this estimate unbiased, biased too high, or biased too low? You can assume that classes taken earlier in the week have roughly the same size as all other classes. Please include a short justification for your answer.**

If the students are reporting the number of students in their first class each week purely by observation (i.e. counting the number of students who are in class), then this estimate is likely to be biased too low, since it is likely that not every student is in attendance during some class period.

If the students are reporting the number of students in their first class via use of the registrar, this is likely to be unbiased.

2. **(Bias in Estimation) There have been many studies on the "hot-hand phenomenon" in sports. In basketball, this occurs when a player has already made several successful shots and is thought to be more likely to make their next shot. Let $X_1, ..., X_n$ denote the n shots taken by a player during a game in chronological order (X_i is 1 for success, and 0 for failure). Let S denote the set of indices of the shots that occur immediately after 3 shots**

were made in a row:

$$S = \{j : 3 < j \leq n \textbf{ and } X_{j-3} = X_{j-2} = X_{j-1} = 1\}$$

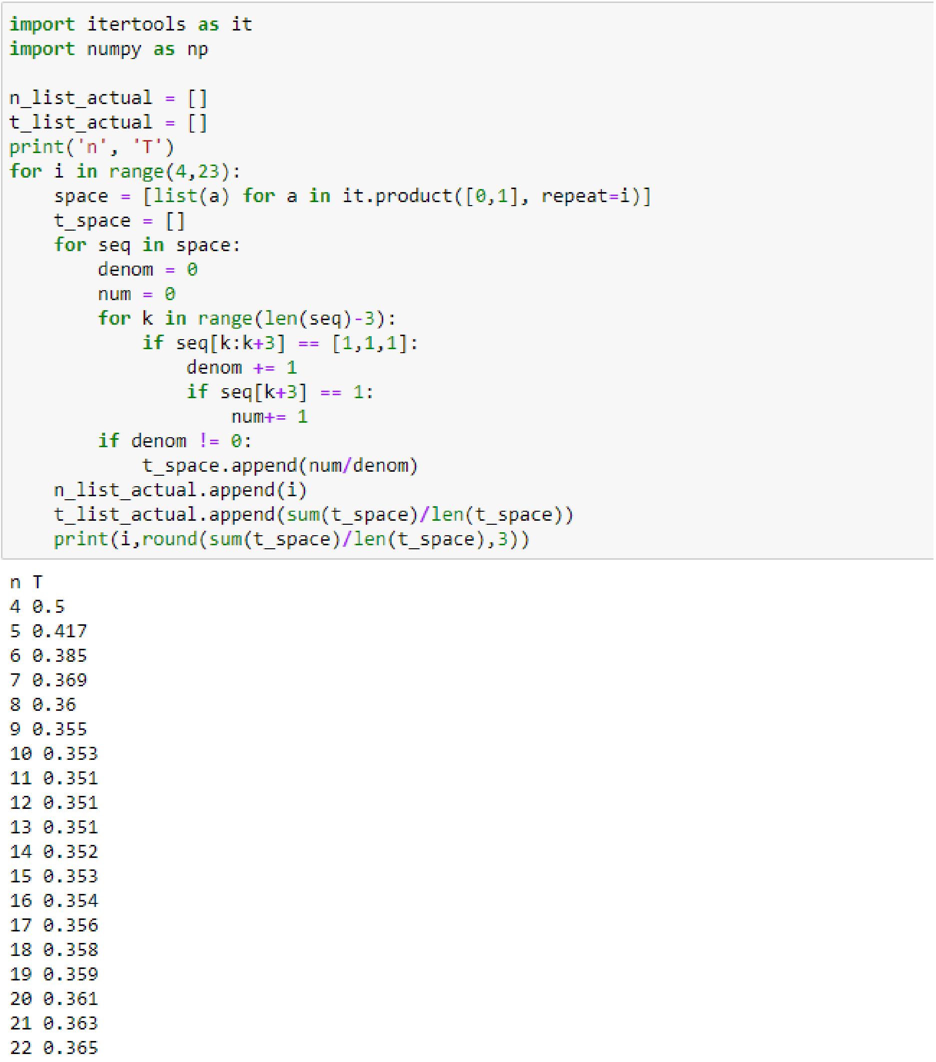
Assuming $S \neq \emptyset$ we compute the statistic T given by,

$$T = \frac{\sum_{j \in S} X_j}{|S|} \tag{11}$$

That is, T is the proportion of shots made in situations when the previous three shots were made. Several studies have incorrectly assumed that if $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(p)$ for some $p \in (0, 1)$ then T is an unbiased estimator of p (i.e., they assume that if the shots are i.i.d. coin flips, then $E[T|S \neq \emptyset] = p$).

(a) Assuming $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(1/2)$, compute $E[T|S \neq \emptyset]$ for $n = 6, ..., 10$ to 3 digits of precision. You are expected to solve this computationally. Make sure to include your code. [Hint: Use *itertools.product* in Python to enumerate all possible outcomes for each n. The solutions for $n = 4, 5$ are $0.5, 0.417$, respectively.]

We calculated the theoretical values of T for $n \in \{1, ..., 10\}$ using the following code. In the next portion, we display a graph of the results.



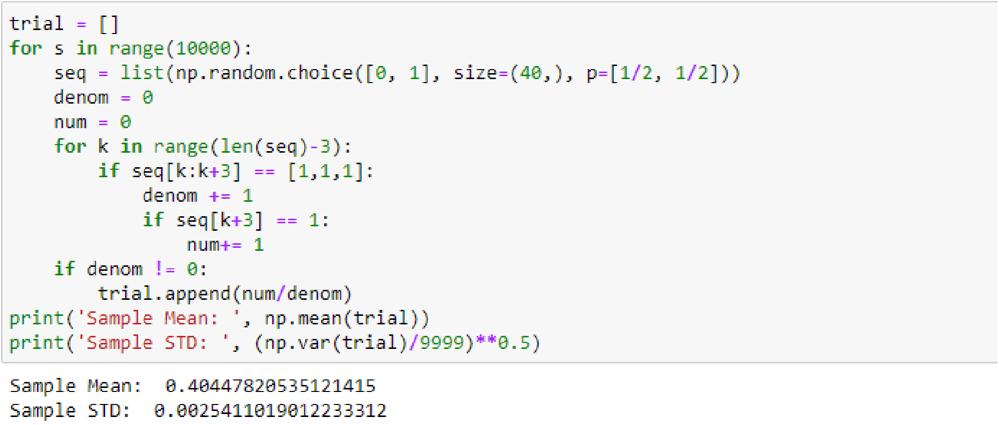
```
import itertools as it
import numpy as np

n_list_actual = []
t_list_actual = []
print('n', 'T')
for i in range(4,23):
    space = [list(a) for a in it.product([0,1], repeat=i)]
    t_space = []
    for seq in space:
        denom = 0
        num = 0
        for k in range(len(seq)-3):
            if seq[k:k+3] == [1,1,1]:
                denom += 1
                if seq[k+3] == 1:
                    num+= 1
        if denom != 0:
            t_space.append(num/denom)
    n_list_actual.append(i)
    t_list_actual.append(sum(t_space)/len(t_space))
    print(i,round(sum(t_space)/len(t_space),3))
```

```
n T
4 0.5
5 0.417
6 0.385
7 0.369
8 0.36
9 0.355
10 0.353
11 0.351
12 0.351
13 0.351
14 0.352
15 0.353
16 0.354
17 0.356
18 0.358
19 0.359
20 0.361
21 0.363
22 0.365
```

We calculated values past the threshold of $n = 10$ to show that the trend of T is not always decreasing.

(b) **Estimate the answer to the previous part for $n = 40$ using 10,000 simulations. Report both your estimated answer, and an approximate normal-based 95.4% confidence interval for the true answer. Remember to also include your code. [Hint: To test your code, simulate for $n = 4, ..., 10$ and compare your answers with the previous part.]**

We approximated the value of T for $n = 40$ using 10,000 simulations. We produced the results as follows,



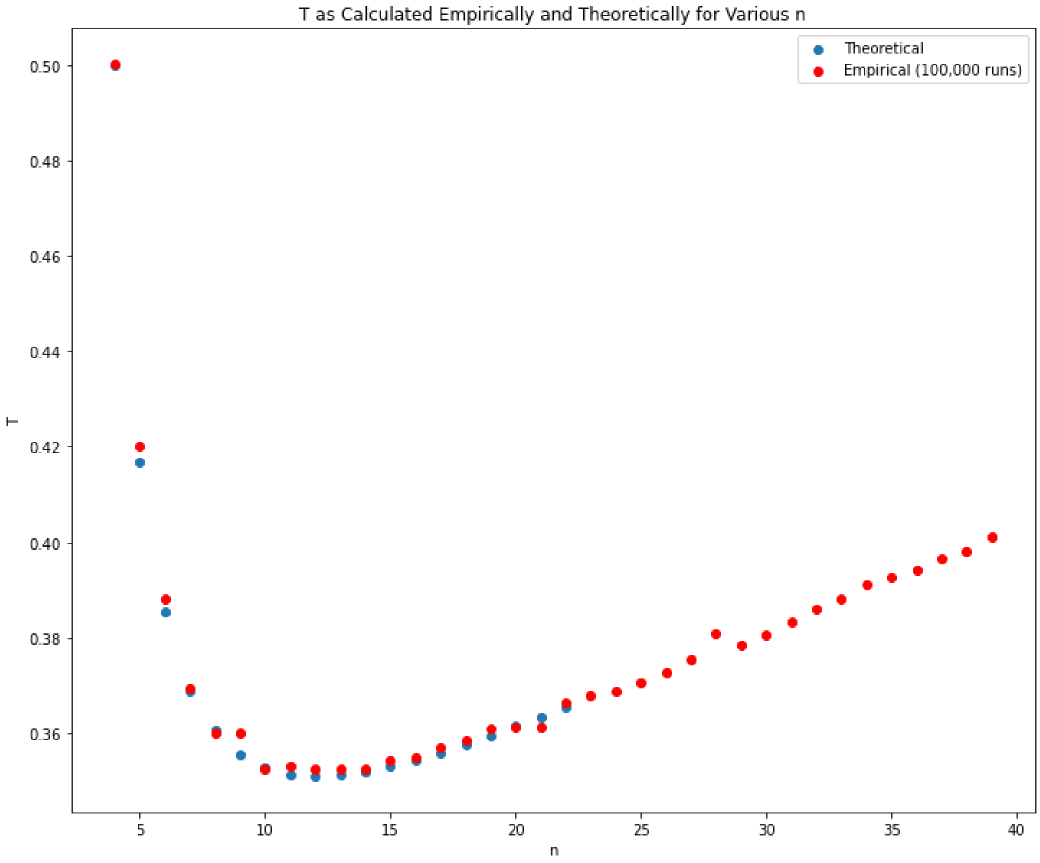
```
trial = []
for s in range(10000):
    seq = list(np.random.choice([0, 1], size=(40,), p=[1/2, 1/2]))
    denom = 0
    num = 0
    for k in range(len(seq)-3):
        if seq[k:k+3] == [1,1,1]:
            denom += 1
            if seq[k+3] == 1:
                num+= 1
    if denom != 0:
        trial.append(num/denom)
print('Sample Mean: ', np.mean(trial))
print('Sample STD: ', (np.var(trial)/9999)**0.5)
```

```
Sample Mean:  0.40447820535121415
Sample STD:  0.0025411019012233312
```

From these results, we can construct a 95.4% confidence interval for T as follows, using the sample mean and sample standard deviation, which we refer to as $\bar{X}$ and $\sigma_{\bar{X}}$, then,

$$[\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}] = [0.3994, 0.4096] \tag{12}$$

Furthermore, as a reassurance, we plotted the empirical and theoretical results for $n \in \{1, ..., 40\}$, though we used $100,000$ trials instead. The results are shown in the following plot. It is clear that the emipirical and theoretical results match up nicely.



3. **(Modeling and Confidence Intervals) Suppose an undergraduate course has 300 students. 40 students are selected at random (without replacement) and asked through email whether they are interested in tutoring (responses are "yes" or "no"). Our goal is to use the data from the 40 students to estimate the proportion of students in the course that want tutoring.**

   (a) **Propose a statistical model for the data.**

   We choose to model a response i as X_i, which is either 1 if the respondent says yes, and 0 if the respondent says no. Then $X_1, ..., X_{40} \overset{iid}{\sim} Bernoulli(\theta)$, and we have an estimator for the total proportion of students who want tutoring,

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{13}$$

(b) Show that the sample mean is an unbiased estimator for the proportion θ of the 300 students that are interested in a tutor.

T as defined above is the sample mean, and to show that T is an unbiased estimator of θ we must show that $E[T] = \theta$,

$$E[T] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta \tag{14}$$

(c) Compute the standard error of the sample mean in terms of θ. [Hint: Compute the covariance between two responses.]

If we sample without replacement, we cannot make the assumption of independence and we have $X_1, ..., X_{40} \sim Bernoulli(\theta)$. Then,

$$\begin{aligned} Var(T) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + 2\sum_{i\neq j}^{n,n} Cov(X_i, X_j)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \theta(1-\theta) + 2\sum_{i\neq j}^{n,n} Cov(X_i, X_j)\right) \end{aligned} \tag{15}$$

Now we need to evaluate the covariance term,

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] \tag{16}$$

$$E[X_i X_j] = \theta(\theta - \frac{1}{300}) \tag{17}$$

$$E[X_i]E[X_j] = \theta^2 \tag{18}$$

$$Cov(X_i, X_j) = -\frac{\theta}{300} \tag{19}$$

Back to the variance equation, we get,

$$Var(T) = \frac{\theta(1-\theta)}{n} - \frac{2}{n^2}\frac{(n^2-n)\theta}{300} \tag{20}$$

(d) How does your answer to the previous part change if you sample with replacement?

If we sample with replacement, the problem actually becomes much easier. We can consider $X_1, ..., X_{40} \overset{iid}{\sim} Bernoulli(\theta)$, and thus,

$$\begin{aligned} Var(T) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\theta(1-\theta) = \frac{\theta(1-\theta)}{n} \end{aligned} \tag{21}$$

So then it follows readily that the standard error is given by,

$$\sigma_\theta = \sqrt{\frac{\theta(1-\theta)}{n}} \tag{22}$$

(e) Give an approximate normal-based 95% confidence interval for θ assuming 10 students said yes, 30 said no, and the samples were taken with replacement. Your answer should not depend on θ.

We have that T is our estimator of θ, and the standard error of θ, so,

$$T = \frac{10}{40} = 0.25 \tag{23}$$

$$\sigma_\theta = \sqrt{\frac{0.25(0.75)}{40}} = 0.0685 \tag{24}$$

The approximately 95% confidence interval becomes,

$$[T - 1.96\sigma_\theta, T + 1.96\sigma_\theta] = [0..1157, 0.3843] \tag{25}$$

(f) **Suppose only 30 of the 40 students replied to your email, and we proceed using the 30 replies. Explain how this could effect our estimate (other than the fact that our dataset is smaller). [Hint: Non-response bias.]**

We need to consider the fact that those who are uninterested in tutoring are probably more likely not to respond to the email, because they have no need of the service, and thus no need to participate in the survey. This is classic case of non-response bias: the individuals not responding have positions that differ greatly from those that do respond.

4. **(Confidence Intervals and Models) You have been given a file *rain.txt* containing rainfall data from 52 clouds. Half of the 52 clouds were chosen at random, and treated with a chemical to increase precipitation. We assume all of the clouds are independent. To model the effect of the treatment, we consider two options:**

- **(Additive Treatment Effect) Let X represent the pre-treatment rainfall of a cloud, and $X^{‘}$ the post-treatment rainfall. Then**

$$X^{‘} = X + \eta$$

for some $\eta \in \mathbb{R}$.

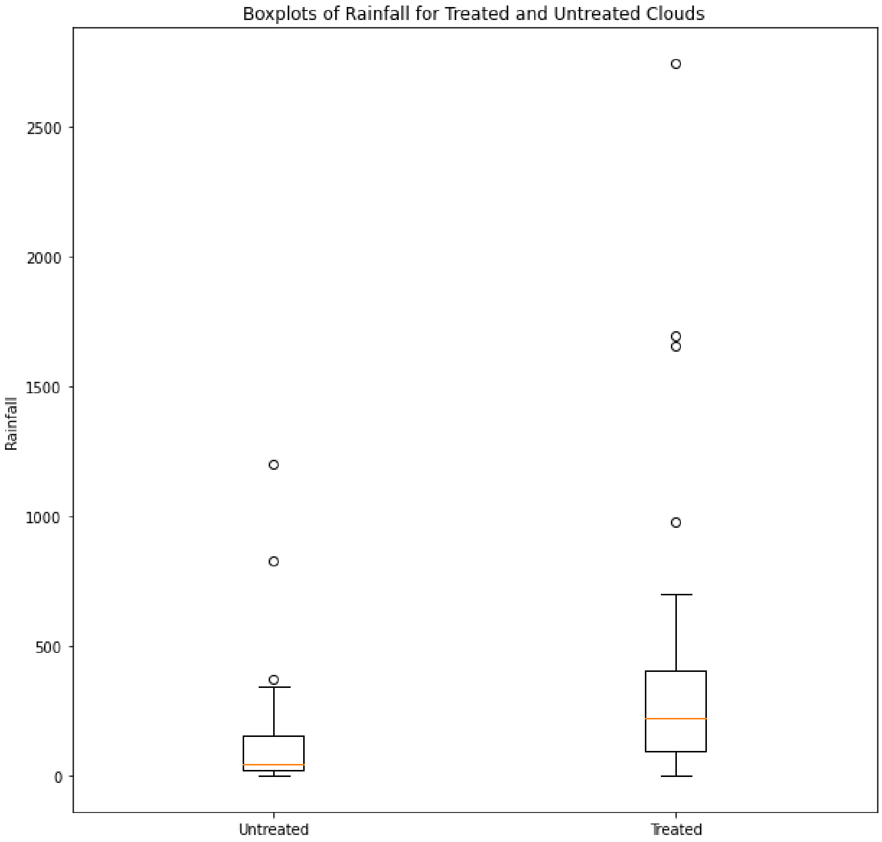
- **(Multiplicative Treatment Effect) Let X represent the pre-treatment rainfall of a cloud, and $X^{‘}$ the post-treatment rainfall. Then**

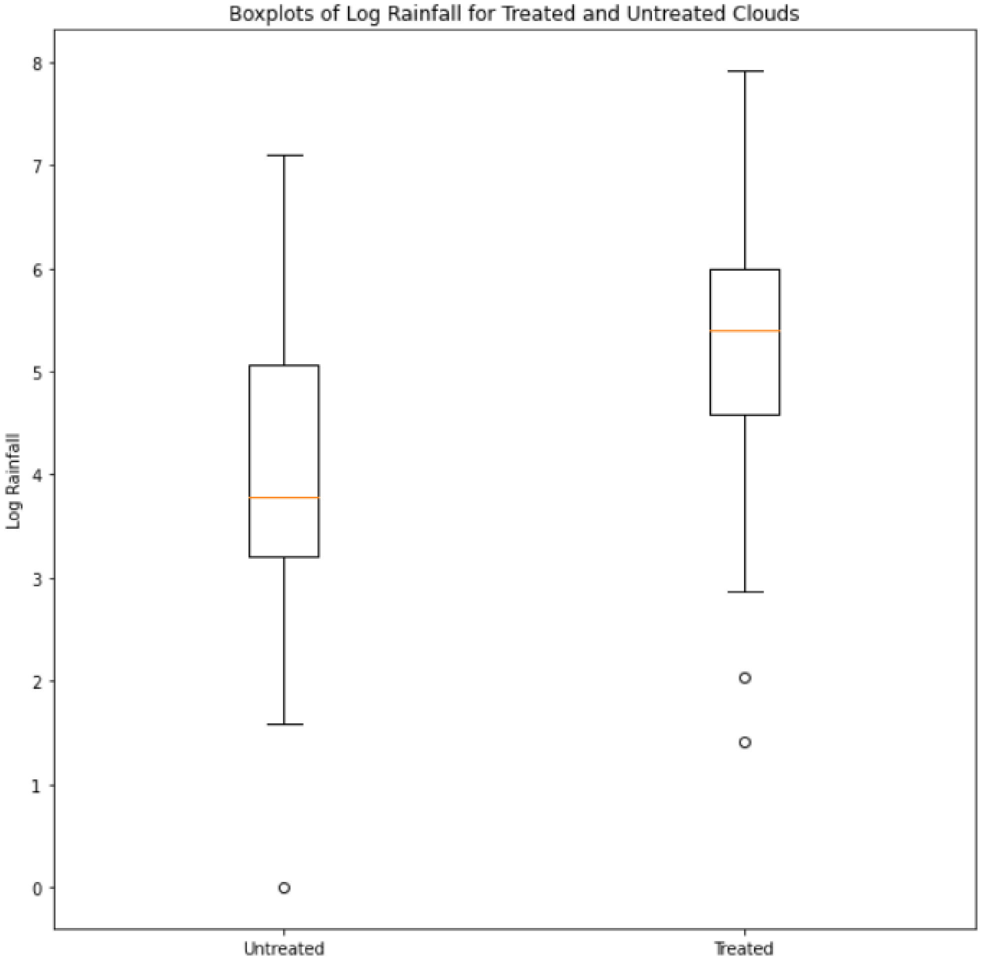
$$X^{‘} = e^{\delta} X$$

for some $\delta \in \mathbb{R}$.

(a) **Construct a box plot showing the untreated and treated rainfall data (2 columns in 1 plot). Then construct a second boxplot for the logarithms of the untreated and treated rainfall data.**

Below are the boxplots of the rainfall and log rainfall concerning the treated and untreated clouds:


Boxplots of Rainfall for Treated and Untreated Clouds
Rainfall
0
500
1000
1500
2000
2500
Untreated
Treated



(b) Assuming an additive treatment model, give an estimate, and an approximate normal-based 95% confidence interval for η. [Hint: What is the variance of your estimator?]

Since we have that $\eta = X^{`} - X$, it follows that a reasonable estimator is given by,

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i^{`} - X_i \tag{26}$$

Where index i serves to iterate us through treated and untreated clouds, taking pairwise differences. T is an unbiased estimator of η since

$$E[T] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^{`} - X_i\right] = E[X^{`} - X] = E[\eta] = \eta \tag{27}$$

And furthermore,

$$\begin{aligned} Var(T) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i^{`} - X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i^{`} - X_i) \\ &= \frac{Var(X^{`}) + Var(X)}{n} \end{aligned} \tag{28}$$

Then, we have that,

$$T = 277.396 \qquad Var(T) = 19270 \tag{29}$$

So in order to construct our confidence interval, we first use the fact that $\sigma_T = \sqrt{Var(T)} = 138.820$, and so our approximately 95% confidence interval for η is given by,

$$[T - 1.96\sigma_T, T + 1.96\sigma_T] = [5.309, 549.483] \tag{30}$$

(c) Assuming a multiplicative treatment model, give an estimate, and an approximate normal-based 95% confidence interval for δ. [Hint: Express $log(X^{`})$ in terms of $log(X)$.]

We can rewrite the multiplicative treatment model as $\delta = ln(X^{`}) - ln(X)$. Then, we consider the estimator,

$$T = \frac{1}{n}\sum_{i=1}^{n} ln(X_i^{`}) - ln(X_i) \tag{31}$$

Where index i serves to iterate us through treated and untreated clouds, taking pairwise differences. Then we simply calculate the sample standard deviations of these natural log differences, yielding,

$$T = 1.144 \qquad \sigma_{\bar{T}} = 0.0708 \tag{32}$$

Then we have an approximately 95% normal based confidence interval which results, and is given by,

$$[T - 1.96\sigma_{\bar{T}}, T + 1.96\sigma_{\bar{T}}] = [1.0024, 1.2856] \tag{33}$$

(d) Give an estimate, and an approximate normal-based 95% confidence interval for the mean rainfall of the untreated clouds. Explain why this should be shorter than our interval for η.

We can easily find the sample mean and sample standard deviations of the rainfall of the untreated clouds, which results in,

$$\bar{X} = 164.588 \tag{34}$$

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{278.426}{\sqrt{26}} = 54.604 \tag{35}$$

So the approximate normal-based 95% confidence interval is given by,

$$[\bar{X} - 1.96\sigma_{\bar{X}}, \bar{X} + 1.96\sigma_{\bar{X}}] = [57.565, 271.612] \tag{36}$$

It makes sense that the confidence interval for the mean rainfall of the untreated clouds is shorter than that of η because the variance of η is much greater, as it takes into account the variance of the *both* the treated and untreated clouds.