SDA- Assignment 5

Jose Chacon (2699643)

5.1: Statistics Resit Grades

For the first exercise, we are given a sample with the grades of the resit for a statistic course. We assume that the sample is a random sample that represents the resit grades in any other year. We have tested several hypotheses using different non-parametric tests. we have tested the following hypothesis:

H0 : $\mu \geq 6.0$	Ha : $\mu < 6.0$ at level $\alpha = 0.10$
H0 : $\mu = 5.5$	Ha : $\mu \neq 5.5$ at level: $\alpha = 0.01$
H0 : $p \leq 0.40$	Ha : $p > 0.40$ at level: $\alpha = 0.05$

The results of the tests as well as their corresponding p-value are included in the *.RData file*. Therefore, they will not be included or explained in this report.

5.2: Cloud Seeding

For the second exercise, we were given two samples from the cloud seeding experiment in 1975. In each sample, we get the precipitation values of clouds when seeded and when not (unseeded). We will refer to the sample with the seeded clouds precipitation as the seeded sample and the other one as the unseeded sample. We have performed several procedures and tests in both samples to gain insights and understand their relation.

a) Graphical and Numerical Summaries of the Sample

First, we started with a numerical investigation of the samples. We can see that the seeded sample has higher values in every numerical statistic (min, median, mean, etc). This is most obvious in the range of each sample: the seeded with a range of [4.10, 2745.60] while the unseeded sample has a range of [0.01, 1202.60]. The mean of the unseeded sample is just about 164 while for the seeded sample is 442. Finally, we have computed the correlation between samples, which results in 0.15 The complete table with statistics can be found in the appendix.

On the other hand, we have also created some visual representations of the samples. We have created a histogram and a boxplot per sample as well as a QQ-plot of both samples. The histograms are presented shortly while the boxplots and the QQ-plot can be found in the appendix.

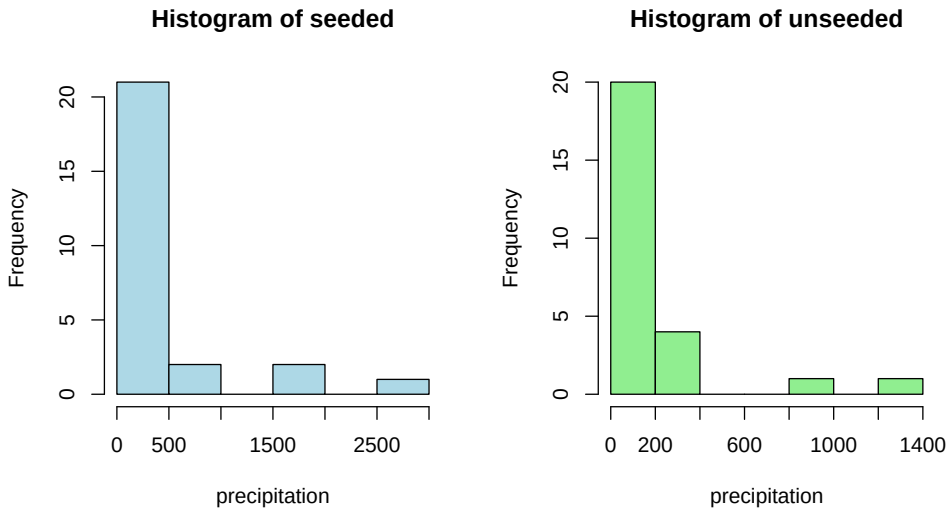

Figure 1: Histograms of the seeded and unseeded samples

From the series of visualizations, we realized that the shape of both the histograms and the boxplots are very similar. However, as mentioned before, the values in the seeded sample are much larger than in the unseeded sample, and this is also visible in the plots. On the other hand, it is very difficult to conclude something from the QQ-plot, because even though a straight line can be fitted; there are still some deviations.

b) Sample Standard Deviation

We have computed the sample standard deviation of both samples. The unseeded sample has a standard deviation of 278.45, while the seeded has 650.79.

c) Bootstrap Estimate

With help of the empirical bootstrap method, we have computed an estimate of the standard deviation of the estimator of accuracy (the standard deviation). For the bootstrap method, we have used B=2000, so it is sufficiently large and is not heavily influenced by randomness. We have obtained a bootstrap sample with 2000 estimates of the standard deviation. Then we computed the standard deviation of the sample. The bootstrap estimates of the standard deviation of the estimator of accuracy for each sample are the following:

1. seeded: 167.19
2. unseeded: 84.03

d) Sample MAD and Bootstrap MAD

First, we computed the MAD for each of the samples. The MAD for the seeded sample is 229.95 and for the unseeded sample is 56.78. We also used the empirical bootstrap method to create a sample with 2000 observations of the MAD. Just as before, we computed the standard deviation of the estimator of accuracy (in this case, the MAD). The results of the bootstrap estimates are the following:

1. seeded: 71.32
2. unseeded: 33.87

e) Preferable estimator for the accuracy

When comparing the estimators of accuracy (the standard deviation and the MAD), we conclude that the MAD is a preferable estimator. There are some reasons for this choice. First, the MAD is less sensitive to outliers; and there are some heavy outliers in the samples. Second, the MAD usually performs better when the data is non-normal, which is the case in both cases. There was no sign of normality when graphically investigating the data. Therefore, a more robust measure of spread is preferred.

f) Test for testing the location

When testing for the location of the precipitation values of the seeded clouds, we would prefer to perform a sign test. Of all the other options, the sign test is the best for this situation. The sign test is a non-parametric test that does not assume a distribution and it is robust to outliers, which we already know we have in the data.

Why are the t-test and the signed-rank test, not good options? First, the t-test is a parametric test that assumes normality, which we have already stated that is not the case. Because the histogram is heavily skewed, we know that the data is not normally distributed. Second, the signed rank test assumes that there is symmetry in the data to be able to use the ranks as the test statistics. Unfortunately, in this case, the data is asymmetric (heavily skewed).

g) Test with significance level $\alpha = 0.05$

We have used the sign test to test whether the location of the precipitation value of the seeded clouds is less than 40 at a significance level $\alpha = 0.05$. By using the sign test, we obtained a p-value of 0.0025, which leads to the rejection of the null hypothesis. Therefore, we conclude that the location should be less than 40.

h) Two-sided 95% Confidence Interval

Finally, we have computed the two-sided 95% Confidence interval for the location of the precipitation value of the seeded clouds based on the sign test, the signed-rank test, and the t-test. The results of these confidence intervals will be presented in the .RData file, so they won't be displayed or explained in this report.

i) Most valuable CI from h)

Just as before, the preferable test to base the confidence interval on would be the sign test. The confident interval is the most narrow of the 3 sets, which means there is a better and more accurate approximation to the "real" value. We again have to mention that this test performs well when there are outliers, which is the case. In conclusion, the most valuable CI is based on the sign test; the second most important would be the signed test and the worst option would be the t-test.

5.3: Newcomb's Experiment

A) Investigation of Differences between Sets of Observations

In this first part, we would like to investigate whether there is a difference between the first 20 and the last 46 observations. We have done it by first plotting different graphs of the samples, by determining an estimate of the difference and by performing tests at level $\alpha = 0.05$.

Graphical Summaries

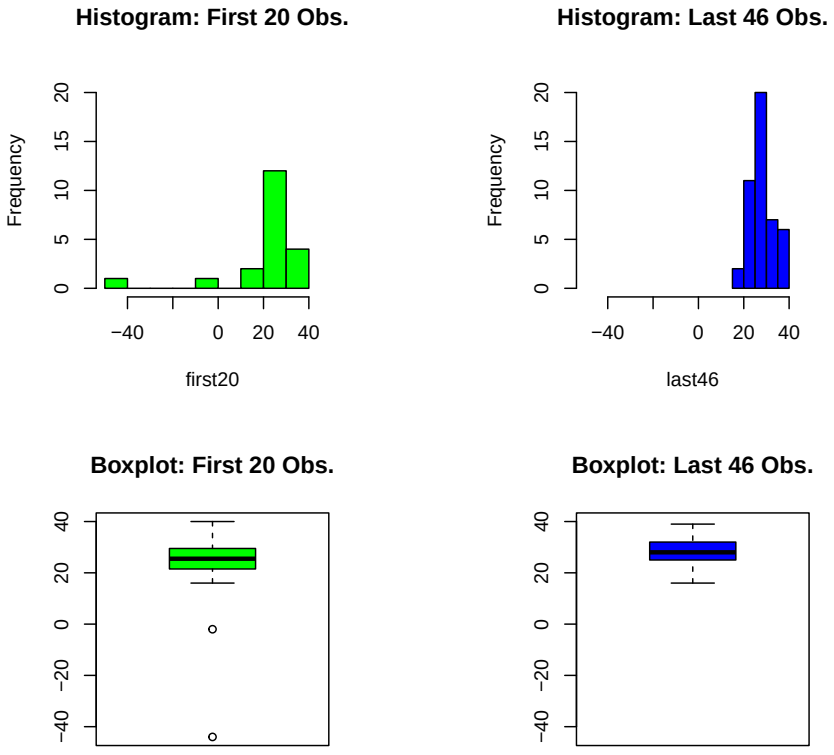

Figure 2: Graphical Summaries of the First 20 and Last 46 Obs.

Description of the findings of the figure above: First, we see that both histograms are bell-shaped (probably an indication of normality). We have found at least two outliers in the first 20 observations. This is visible in the boxplot (the points outside the whiskers) as well as in the histogram. It is also clear that the mean and median of the first 20 are lower than the mean and median of the last 46. This is due to the outliers in the first 20 observations. Lastly, we can also see that the mean and median in the last 46 obs. are very close to each other(mean of 28.15 and median of 28), while in the first 20, there is a difference (mean of 21.75 and median of 25.5).

To investigate deeper the underlying distribution of the samples, we plotted the QQ plots of the sample against known distribution. The most important finding is that both samples have positive results when plotting the QQ plot against the normal distribution. We additionally plotted the QQ-line is almost a perfect fit for these QQ plots. This suggests normality in both samples (although there are some heavy outliers). The QQ plots are shown below.

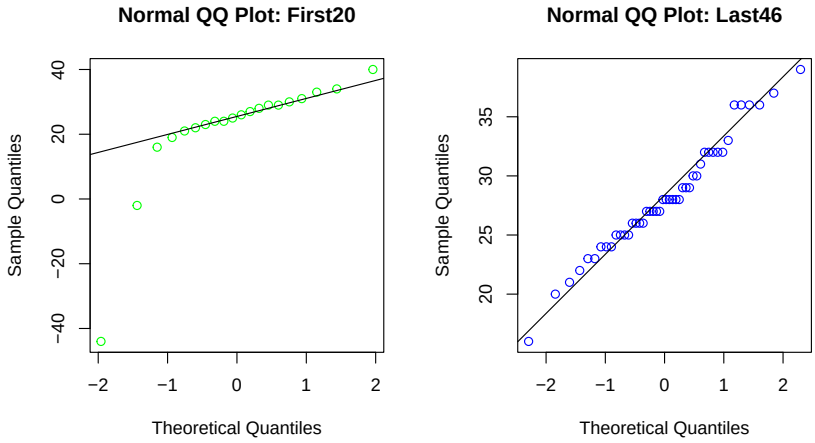

Figure 3: QQ Plots against the Normal Distribution

Estimate of the Difference and Tests at level $\alpha = 0.05$

As the estimate of the difference between the samples, we have chosen the sample median difference. This is the best option when there are some outliers in the data. It provides a good estimator of the usual difference between the two populations.

We have chosen to use the following tests: Two Samples Kolmogorov Smirnov Test and the Mann-Whitney U Test (also known as the Wilcoxon rank-sum test),

First, we used the KS test because it does not make any assumption of shape or parameters, which means that does not assume a distribution. Only when the test statistic is larger than the critical value, the null hypothesis is rejected. This indicates that both samples do not come from the same distribution. However, when applying this test to the first 20 obs. and the last 46 obs., we failed to reject the null hypothesis (we got a p-value of 0.18). This means that is very probable that the same come from the same distribution.

Second and most important, we used the Mann-Whitney test. This test compares the medians of the two samples to determine if they come from the same population. This test is more relevant in this case because it uses the median (just as we used the median as the estimate of the difference in the samples). The results of this test are in line with the KS test. We failed to reject the hypothesis that both samples come from the sample distribution. In this case, the p-value was 0.12. Then, we conclude that is most likely that both samples come from the same distribution (probably the normal distribution.)

b) 90% Confidence Interval

First, as a location estimator, we would choose the median of the sample. This is because we need a robust estimator. It is less sensitive to outliers and reflects better a sample like the one we have. We used the Wilcoxon test to create the confidence interval, which was (26.5, 28.42). When comparing it with the true value (after the transformation (x-24.8)*1000), we realized that the values are not in the Confidence interval but it is close. It is a good approximation for an experiment in 1882.

Appendix

5.1 Supporting tables and Plots

stats	seeded.clouds	unseeded.clouds
Min.	4.10	0.01
1st Qu.	98.12	24.82
Median	221.60	44.20
Mean	441.98	164.56
3rd Qu.	406.02	159.20
Max.	2745.60	1202.60

Table 1: Numerical Summary for the seeded and unseeded samples

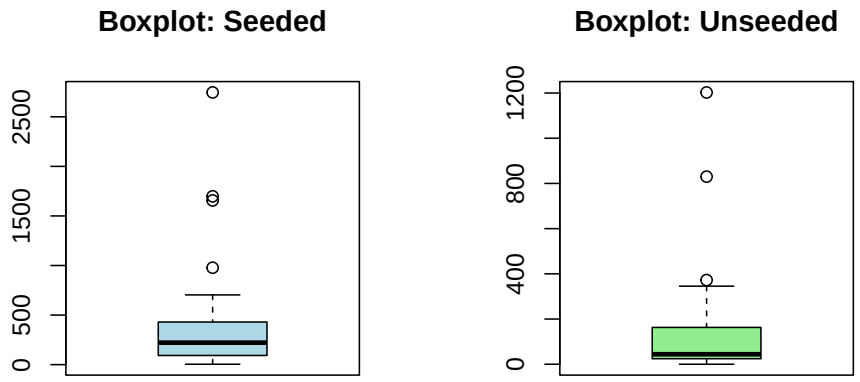

Figure 4: Boxplots of the seeded and unseeded samples

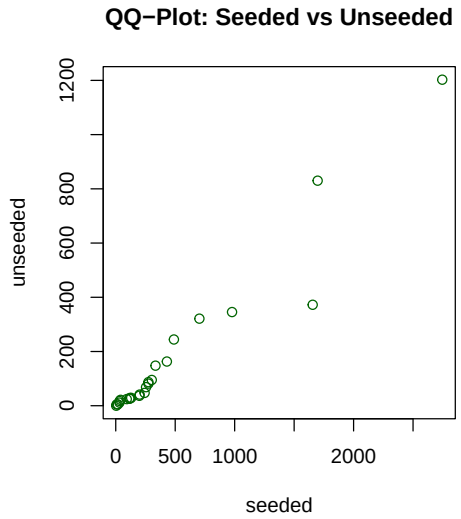

Figure 5: QQ-Plot of the seeded and unseeded samples

Code 5.1

```
sample51 = scan('statgrades.txt')
mean(sample51) #5.536
median(sample51) #5.69

reject_hypothesis = function(pvalue, alpha) {
```

```
  if (pvalue <= alpha) {
    return(TRUE)  # Reject the null hypothesis
  } else {
    return(FALSE)  # Fail to reject the null hypothesis
  }
}

E1.a.pval = wilcox.test(sample51, mu=6, alternative='less')[[3]] #0.02
E1.a.reject = reject_hypothesis(E1.a.pval, 0.1) #TRUE

E1.b.pval = wilcox.test(sample51, mu=5.5)[[3]] #p-val=.65
E1.b.reject = reject_hypothesis(E1.b.pval, 0.01) #False

E1.c.pval = binom.test(sum(sample51>=5.5), n=length(sample51), p=0.40, alternative='g')[[3]]
E1.c.reject = reject_hypothesis(E1.c.pval , 0.05) #TRUE
```

Code 5.2

```
source("functions_Ch5.txt")
sample52 = read.table('clouds.txt')
seeded = sample52$seeded.clouds
unseeded = sample52$unseeded.clouds

par(mfrow=c(1,2))
hist(seeded, col='lightblue', xlab='precipitation')
hist(unseeded, col='lightgreen', xlab='precipitation') #very similar
boxplot(seeded, main='Boxplot: Seeded', col='lightblue')
boxplot(unseeded, main='Boxplot: Unseeded', col='lightgreen') #very similar

par(mfrow=c(1,1), pty='s')
qqplot(seeded, unseeded, main='QQ-Plot: Seeded vs Unseeded', col='darkgreen')

summary(sample52)
cor(sample52)

#b) sample standard deviation
sd.unseeded= sd(unseeded) #278.45
sd.seeded =sd(seeded) #650.79

#c) bootstrap estimate of the standard deviation
set.seed(20202020)
bs.sd.seeded2 = sd(bootstrap(seeded, sd, 2000)) #167.19
bs.sd.unseeded2 = sd(bootstrap(unseeded, sd, 2000)) #84.03

mad.seeded = mad(seeded) #229.95
mad.unseeded = mad(unseeded) #56.78

bs.mad.seeded2 = sd(bootstrap(seeded, mad, 2000)) #71.32
bs.mad.unseeded2 = sd(bootstrap(unseeded, mad, 2000)) #33.87

binom.test(sum(seeded>40), length(seeded))[[3]] #0.0025

t.test(seeded, conf.level=0.95) #correct way to solve it
wilcox.test(seeded, mu=40, conf.int = TRUE, conf.level = 0.95) #[147.80, 505.35]

rbind(0:length(seeded), round(pbinom(0:length(seeded), size=length(seeded), p=0.5),3))
```

```
rbind(0:length(seeded), round(1-pbinom(0:length(seeded)-1, size=length(seeded), p=0.5),3))
CI_l = sort(seeded)[8]
CI_u = sort(seeded)[19] #
CI = c(CI_l, CI_u) #115.3 and 334.1

E2.h.CI.sign.upper.bound.included = FALSE
E2.h.CI.sign.lower.bound.included = TRUE

#1-P(T<=18)
1-pbinom(18, length(seeded)-1, 0.5)
pbinom(9, length(seeded)-1, 0.5)
```

Code 5.3

```
sample53 = scan('newcomb.txt')
first20 = sample53[1:20]
last46 = sample53[21:66]

par(mfrow=c(2,2))
hist(first20, xlim=c(-50,45), ylim=c(0,22), col='green', main='Histogram: First 20 Obs.')
hist(last46, xlim=c(-50,45), ylim=c(0,22), col='blue', main='Histogram: Last 46 Obs.')

ylim <- c(min(first20, last46), max(first20, last46))
boxplot(first20, col='green', ylim=ylim, main='Boxplot: First 20 Obs.')
boxplot(last46, col='blue', ylim=ylim, main='Boxplot: Last 46 Obs.')

par(mfrow=c(1,1), pty='s')
qqplot(first20, last46, col='darkgreen', main='QQ Plot First_20 vs Last_46') #not a straight li
par(mfrow=c(1,2), pty='s')
qqnorm(first20, col='green', main='Normal QQ Plot: First20')
qqline(first20)
qqnorm(last46, col ='blue', main='Normal QQ Plot: Last46')
qqline(last46)

mean(first20) #21.75
mean(last46) #28.15
median(first20) #25.5
median(last46) #28

diffMean = mean(last46) - mean(first20) #6.4
diffMedian = median(last46) - median(first20) #2.5

wilcox.test(first20, last46, conf.level = .9) #p-val=0.12
ks.test(first20, last46) #p-val=0.18

mean(sample53) #26.21
median(sample53) #27

t.test(sample53, conf.level=0.9) #(24.01, 28.42)
wilcox.test(sample53, conf.int = TRUE, conf.level = 0.9) #(26.5, 28.50)
true_value = (24.8332-24.8)*1000 #33.20
```