Statistical Data Analysis
Assignment 5

Matt Kuodis & James Zoryk.
2640428 : 2663347
Group: 36
Due Date: 19.04.2022.

Question 5.1.

In this exercise we test certain simple hypotheses about the resit grades of a statistics course for the Computer Science students. We first plot and summarize the data, and plot a vertical blue line to denote the median value we computed with the built in function.

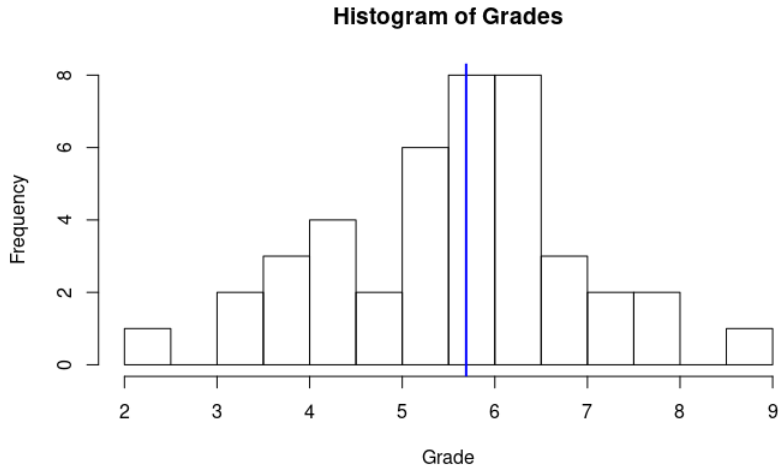

Figure 1: Histogram of the grades data, and the computed median value given by the vertical blue line.

Part a.

Initially we test the hypotheses $H_0 : m \geq 6.2$ against $H_\alpha : m < 6.2$ at level $\alpha = 1\%$, where m is the median. We use the built in binomial test with the parameters given by taking the number of grades greater than 6.2, and the total amount of grades minus the amount of grades equal to 6.2. We set the confidence level to 99% and set the alternative hypothesis to "less than", since we want to reject the null hypothesis for small values of m. The resulting p-value is 0.001443624, which is far below the desired value. This also seems fairly evident when looking at the histogram, and thus we reject the null hypothesis.

Part b.

The second set of hypotheses we test are $H_0 : m = 6$ against $H_a : m \neq 6$ at level $\alpha = 5\%$. We use the same method as before, however now we note that this is a two sided test, which is rejected for both large and small test values. We use as parameters the amount of grades larger than 6, and the total amount minus the amount of grades exactly equal to 6. Since this is a two sided test, it follows that the amount of grades less than 6 will be equal to the absolute value of the difference of the two parameters. We set the alternative hypothesis to two sided, for the aforementioned reasons, and we set the confidence level to 95%. This test returns a p-value of 0.2110236, which is larger than the level $\alpha = 0.05$ and thus we do not reject the null hypothesis. Our computed median value is 5.69 which seems to be close enough for us to lack the certainty to reject the null.

Part c.

In this subquestion, we test the probability of getting a grade equal to, or greater than, a 5.5. Our set of hypotheses is $H_0 : p \leq 45\%$ against $H_a : p > 45\%$ at level $\alpha = 10\%$. Here we use the built in binomial test with the parameters of grades greater than or equal to 5.5, the total amount of grades, the probability (which we have at 0.45 as per our hypotheses), and we set our alternative hypothesis to "greater than", since we reject the null hypothesis if the probability of gaining a grade above or equal to 5.5 is larger than 45%. We set the confidence level to 90%, and run the test. This gives us a p-value of 0.04151232, which is less than our desired value, and thus we again reject the null hypothesis, meaning that the probability of receiving a grade equal to or above a 5.5 is greater than 45%.

Question 5.2.

Part a.

In this exercise we are tasked with numerically and graphically analyzing the 'unseeded' cloud data from 'cloud.txt'. We begin by computing a numerical summary of the data which can be found in Table 2. Furthermore we note that the data has a couple of extreme outliers, so we also consider the data with values greater than 600 removed. The evidence for this comes from the boxplot of the data which is given in Figure 2.

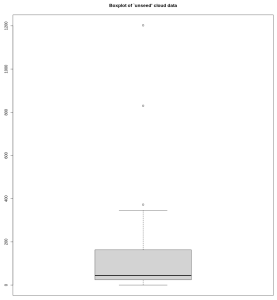

Figure 2: Boxplot of the data.

Data	Min	1st Qu	Median	Mean	3rd Qu	Max
'unseeded' all	0.01	24.82	44.20	164.56	159.20	1202.60
'unseeded' <600	0.01	23.73	38.85	93.58	108.20	372.40

Table 1: Numerical summary of 'unseeded' data with removed possible outliers.

From the Table 2, in both cases, we see that the mean is far greater than the median, implying that the data has a positive skew. This can be clearly seen in the histogram plot of the data which is given in Figure 3.

There is more evidence to support this if we also look at the 'symplot' which can be seen in Figure 4.

The next stage in our analysis is to determine which possible location and scale family this data may have arisen from. To do this we shall consider the 'QQ-plot' against distributions that have a positive skewness such as the 'exponential', 'chi-squared' and 'normal' distributions. Figure 5, show the results of the QQ-plots.

Part b.

For this exercise we compute the Standard Deviation (SD) of the 'unseeded data' which is computed as 278.4462.

Part c.

For this exercise we conduct a bootstrap of the 'unseeded data' of the Standard Deviation to determine the accuracy of the value of the test statistic obtained in part b. We compute the bootstrap with $B = 5000$, the results of which are plotted in the histogram in figure 6.

The bootstrap Standard Deviation is computed as 82.58512.

Part d.

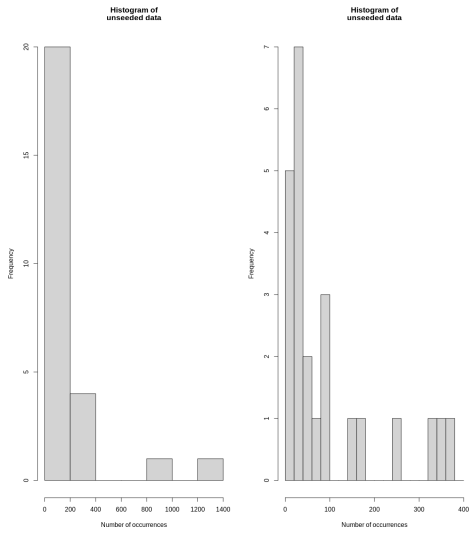

Figure 3: Histogram of both the full data and the reduced data set.

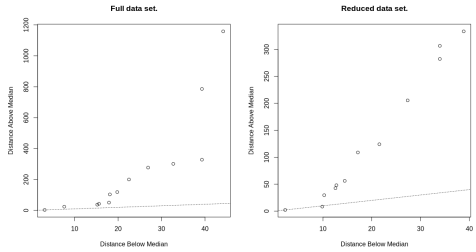

Figure 4: Histogram of both the full data and the reduced data set.

In this task we are asked to repeat our processes from parts [b] and [c], but now using the Median Absolute Deviation ('MAD') value instead. The 'MAD' of the 'unseeded data' is computed as 56.78358 Now we compute the bootstrap with the results plotted in figure 7 and the value of the SD of the MAD bootstrap is calculated as 32.05318.

Part e.

Now we compare the result obtained in part [b] and [c] against part [d], with regards to the accuracy of the measurements. We first note that the difference between the two is that

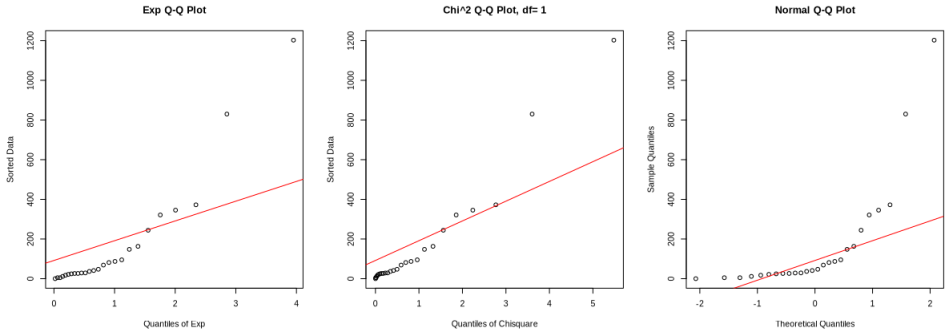

Figure 5: Various QQ-plots.

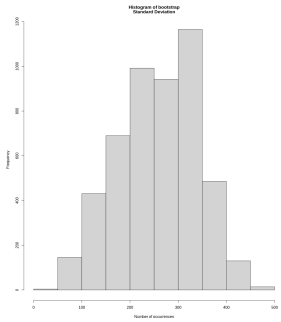

Figure 6: Histogram of the bootstrap of the Standard Deviation (SD).

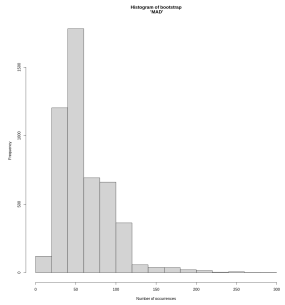

Figure 7: Histogram of the bootstrap of the MAD.

the SD is the square of the difference

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2},$$

whereas the MAD is only looking at the absolute difference

$$\text{MAD} = \text{median}(|x_i - \overline{X}|).$$

Therefore any outliers in the data will create a higher dispersion when using SD compared to MAD. Furthermore, we see that the SD of the bootstrap MAD is smaller than that of the regular SD ($56.78 < 82.85$), this indicates that it is a more precise result to use. There is more evidence of this if we compare the histogram in Figure 6 and 7, as we see that the MAD histogram has a higher density than the SD. However the right-hand tail of the MAD may be an issue. To conclude, we believe that in this case the MAD will make for the better choice for the reasons stated above.

Part f.

We shall consider the t-test, signed test and the signed rank test, to see which one is the most ideal for the unseeded data. Let us first consider the signed rank test, we note that the test assumes that the data is continuous and symmetric around the median. This is clearly not the case if we refer back to the analysis done in part [a]. Therefore we can dismiss the signed rank test for this application.

Next we shall consider the t-test, which by definition assumes under the H_0 that the distribution follows the normal distribution. Again, from part [a], namely the qq-plots in Figure 5, we see that that is not the case. Therefore the t-test is not the best suited for this.

Lastly, let us consider the signed test which assumes that the data has a unique median for each of the observation. The test makes no assumption about the data and can be used on unsymmetrical data, which we have in this case. Moreover, the sign test has the weaker assumption of the above stated tests.

Therefore we conclude that the signed test would be the preferred test to use with the presented data.

Part g.

For this exercise we conducted the following test

$$H_0 : m < 40$$
$$H_1 : m \geq 40.$$

We have the following test statistc

$$T = \#(X_i > 40).$$

Here we should note the test statistic T has a binomial distributed under the H_0 with the parameters $n = 26$ and $p = 1/2$. Here we conduct the test with significance level $\alpha = 0.05$. We find that there are 14 values greater than 40. Conducing the 'binom.test' that is available in RStudio we obtained the p-value $= 0.4225$. Since p-value $> \alpha$, we conclude that there is not enough evidence to reject H_0.

Part h.

In this exercise we are tasked with computing two-sided 95% confidence intervals of the unseeded cloud data for the sign test, the signed rank test and the t-test. Using the built-in function of RStudio, namely 'wilcox.test', 't.test' and computing the sign test ourselves, we yield the results in Table 2.

Test	Lower Interval	Upper Interval	P-value
Sign test:	26.3	147.8	0.845
Sign Rank test	36.69997	187.24999	0.07334
t- test	52.09509	277.02876	0.03133

Table 2: Confidence and p-values of the three tests.

Part i. The confidence interval we prefer to use in part [h] will be the signed test for the following reasons. The first reason is that the interval is the smallest out of all the test, $121.5 < 150.55 < 224.93$. Therefore the confidence interval is more precise for the sign test, compared to the other tests. Another reason for the preference of using the signed test in this case is that the sign test uses fewer assumptions than the sign rank and t tests.

Question 5.3.

Part a.

Here we compare the first 20 data points to the remaining data points of the given data set. We first compare the data graphically using histograms and boxplots.

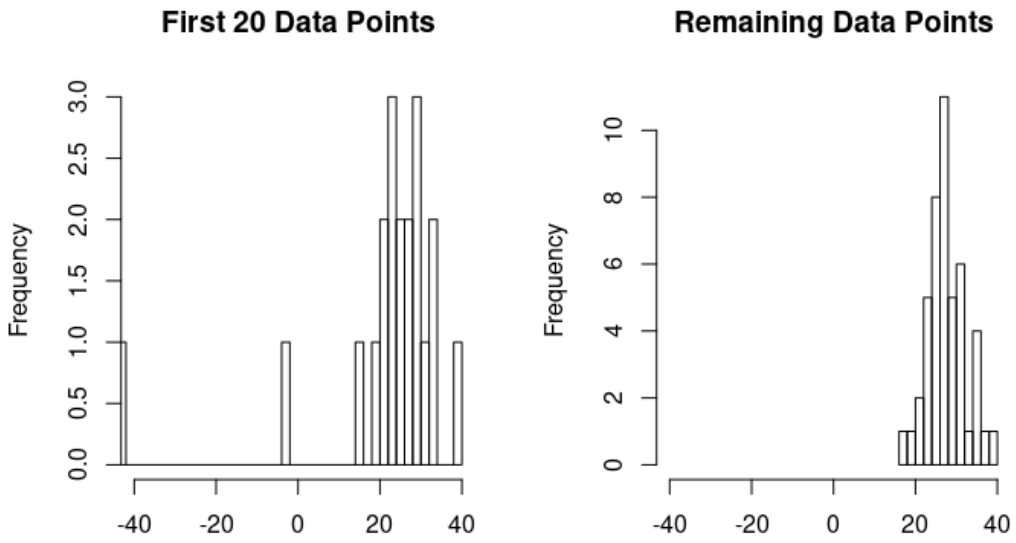

Figure 8: Histograms of the first 20, and the remaining 46 data points.

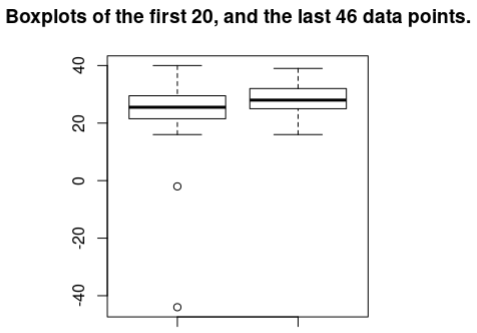

Figure 9: Boxplots of the first 20 data points (left), and the remaining 46 data points (right).

We note that the data of the first 20 points is much more spread out, including some quite noteworthy outliers in the data. The dataset itself seems to be bimodal, whereas the remaining 46 data points seem to be very normally distributed. The outliers and spread of the data are also well visualized in the boxplots.

Since the data seems to be, at least on a surface level, normally distributed, we made QQ-plots where we have looked at how well the data fits the normal distribution.

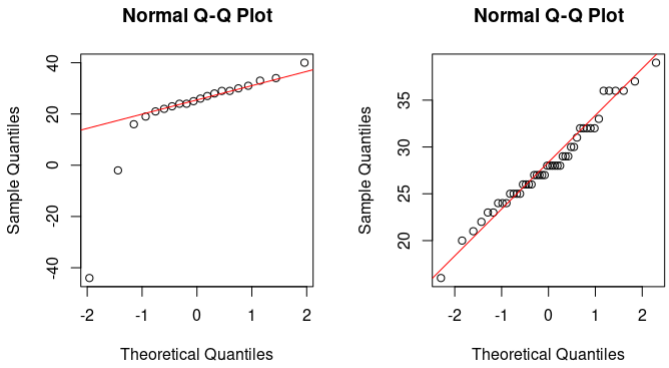

Figure 10: QQ-Plots of the first 20, and the remaining 46 data points.

Furthermore, we investigate the differences between the medians of bootstrapped samples of the data. If the data sets were similar, this would reduce the impact of the outliers. The differences in the medians of the bootstrapped samples are shown in the histogram below.

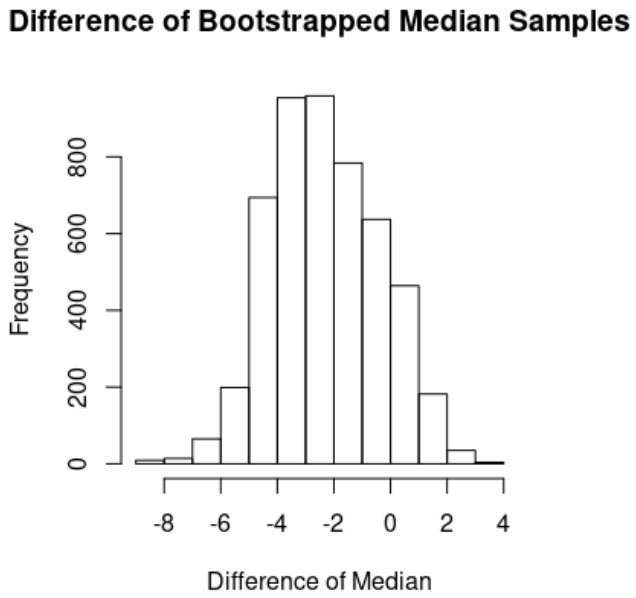

Figure 11: Differences between bootstrapped samples from the first 20, and the remaining 46 data points.

We also create a numerical summary of this difference sample, which gives us the following table.

Data	Min	1st Qu	Median	Mean	3rd Qu	Max
Bootstrap Median Differences	-9.000	-3.500	-2.000	-2.014	-0.500	4.000

Table 3: Numerical summary of the bootstrapped median differences.

We note that there is a substantial difference between the medians of the bootstrapped samples. On average the first 20 data points are off by -2 (median) and -1.889 (mean), a pretty substantial deviation.

Finally we perform both the Wilcoxon Two-Sample test, and the Kolmogorov-Smirnov test to check whether the two samples have a similar distribution. In either case we have the hypotheses $H_0 : F = G$, against $H_a : F \neq G$ where F is the distribution of the first 20 data points, and G is the distribution of the last 46 points. For the Kolmogorov-Smirnov test we get the test value $D = 0.25435$, and p-value $= 0.328$. For the Wilcoxon test we get $W = 348$, and p-value $= 0.1189$. In either case we cannot reject the null-hypothesis that the two data sets share a common distribution.

Part b.

Since we can use the modern measurement of 24.8332 for light to travel the 3721 metres, we can consider using the median as a location estimator. The reason for this choice is that the median is a robust estimator that splits the data into two equal half's and can be thought of as the 'middle' value of the data. Therefore the median is less affected by any outliers we may have in the data set, given there could be many, considering the accuracy of measuring devices of that time period. With this in mind we construct the following Wilcoxon test to determine the 95% confidence interval ($\alpha = 0.05$)

$$H_0 : m = 24.83,\ H_1 : m \neq 24.83.$$

We compute the confidence interval using the built-in Wilcoxon test, which gives us a 95% confidence interval between $[26.00003, 28.50007]$, the test statistic $V = 1672$, and the p-value $=0.00029$. Here we clearly see that 24.83 is not in the interval and that p-value is smaller than α. Therefore, there is evidence to reject H_0.

The deviation from the computed median and the modern measurement could potentially come from unreliable technology used at the time. This is also more apparent if we look at the deviation in the medians between the first 20 and the last 46 data points as done earlier. It's possible that the technology at the time was unreliable, or reliable only for a certain amount of time, and thus the final result ends up deviating from our modern day estimations.

R code

```
# Pull data from disk
setwd("~/Maths/SDA/Assignment_5")
# Grab the data
source("functions_Ch3.txt")
source("functions_Ch5.txt")

grades_data <- scan("statgrades.txt")

# ex1
hist(grades_data, breaks=20)
abline(v = median(grades_data), col = "blue", lw = 2)
summary(grades_data)

#a.

larger_grade = sum(grades_data>6.2)
smaller_grade = sum(grades_data<6.2)
equal_grade = sum(grades_data==6.2)
length_grades = length(grades_data)

test_value = binom.test(larger_grade, length_grades-equal_grade, alt="l", conf
    .level = 0.99)
test_value
test_value["p.value"]

#b.

larger_grade2 = sum(grades_data>6)
equal_grade2 = sum(grades_data==6)

test_value2 = binom.test(larger_grade2, length_grades-equal_grade2, alt="two.
    sided", conf.level = 0.95)
test_value2
test_value2["p.value"]

#c.

greater_grade3 = sum(grades_data>=5.5)
greater_grade3

test_value3 = binom.test(greater_grade3, length_grades, p = 0.45, alt="g",
    conf.level = 0.9)
test_value3
test_value3["p.value"]

##########################################
#ex2
data <- read.table("clouds.txt")
# split data
data_unseed = unlist(data['unseeded.clouds'])
data_seed   = unlist(data['seeded.clouds'])

summary(data_unseed)
summary(data_unseed[data_unseed < 600])
par(mfrow=c(1,1))

boxplot(data_unseed,
        main="Boxplot of 'unseed' cloud data")

par(mfrow=c(1,2))
hist(data_unseed,
     main=c("Histogram of", "unseeded data"),
```

```
      xlab = "Number of occurrences")
hist(data_unseed[data_unseed < 600], seq(0,400,20),
      main=c("Histogram of", "unseeded data"),
      xlab = "Number of occurrences")

par(pty='s')
symplot(data_unseed,
        main="Full data set.")
symplot(data_unseed[data_unseed < 600],
        main="Reduced data set.")

par(mfrow=c(1,3))
qqexp(data_unseed)
qqline(data_unseed, col="red")
qqchisq(data_unseed, df=1)
qqline(data_unseed, col="red")
qqnorm(data_unseed)
qqline(data_unseed, col="red")


#b
ex2_b_sd = sd(data_unseed)
#part c
# bootstrap
 bs_sd_unseed <- bootstrap(data_unseed, sd, B=5000)
 par(mfrow=c(1,1),pty='m')
hist(bs_sd_unseed,
      main=c("Histogram of bootstrap", "Standard Deviation"),
      xlab = "Number of occurrences")
sd(bs_sd_unseed)



#part d
# bootstrap with mad
ex2_d_mad=mad(data_unseed)
bs_mad_unseed <- bootstrap(data_unseed,mad,B=5000)
hist(bs_mad_unseed,
      main=c("Histogram of bootstrap", "  'MAD'"),
      xlab = "Number of occurrences")
sd(bs_mad_unseed)


#part g
sum(data_unseed>40)
length(data_unseed)
sum(data_unseed<=40)
binom.test(14,26, alt="g", conf.level = 0.95)

#part h
# signed rank test
wilcox.test(data_unseed, mu=40, conf.level = 0.95, conf.int = T)
# t test
t.test(data_unseed,mu=40,conf.level = 0.95)
#signed test
length(data_unseed)
sum(data_unseed<40)
sum(data_unseed==40)
binom.test(12,26,alt="two.side", conf.level = 0.95)

rbind(0:26, round(pbinom(0:26, size = 26, p = 0.5), 3))
rbind(0:26, round(1-pbinom((0:26)-1, size = 26, p = 0.5), 3))
# {9, 19}
```

```
sort(data_unseed)[9:19]
#  confidence intervals
# (26.3, 147.8)

############################################################
# ex 3
data <- scan("newcomb.txt")
f_data = data[1:20]
l_data = data[21: 66]

boxplot(f_data, l_data, main = "Boxplots of the first 20, and the last 46 data
     points.")

par(mfrow=c(1,2), pty="m")
hist(f_data, breaks = 50,
     main="First 20 Data Points", xlab = " ", xlim=c(-40, 40))
hist(l_data, breaks = 10,
     main="Remaining Data Points", xlab = " ", xlim=c(-40, 40))


par(mfrow=c(1,2), pty="s")
qqnorm(f_data)
qqline(f_data, col="red", xlim = c(-2, 2), ylim = c(-40, 40))
qqnorm(l_data)
qqline(l_data, col="red", xlim = c(-2, 2), ylim = c(-40, 40))

shapiro.test(f_data)
shapiro.test(l_data)

#bootstrap difference
bs_diff <- bootstrap(f_data,median, B=5000)- bootstrap(l_data,median, B=5000)
par(mfrow=c(1,1), pty="s")
hist(bs_diff, main="Difference of Bootstrapped Median Samples", xlab = "
    Difference of Median")
summary(bs_diff)


wilcox.test(f_data, l_data, alternative="two.sided")
ks.test(f_data, l_data)

# Part B.
wilcox.test(data, mu=24.8332, conf.int = T)



# rdata
mylist<- list(stud_no=c(2640428,2663347),
              "1_a_pval" =test_value["p.value"],
              "1_b_reject" = TRUE,
              "1_b_pval"= test_value2["p.value"],
              "1_b_reject"= FALSE,
              "1_c_pval"= test_value3["p.value"],
              "1_c_reject"= TRUE,
              "2_b_sd"= ex2_b_sd,
              "2_d_mad"= ex2_d_mad,
              "2_h_CI_sign"= c(26.3, 147.8),
              "2_h_CI_wilcox"= c(36.69997,187.24999),
              "2_h_CI_t"= c(52.09509,277.02876)
              )

# Save to disk the return of the list
save(mylist, file="~/Maths/SDA/Assignment_5/myfile5_36.RData")
```

Code/sda_hw5_code.R