Exercise 5.2

a) To explore the distribution of the cloud seeded data graphically we used some different methods. First we looked at the symmetry and plotted a histogram and a box plot of the data, as you can see in the box plot, it’s not the same on each side of the median so there could be no symmetry. In the symplot, the data is roughly on the median but with larger values the points diverge from this line. Thus, we can’t know for certain that the data is symmetrical. Also, since the data is small there’s no certainty to assume whether the data is symmetric or not. After looking at a few QQ-plots, it’s still not clear what the distribution could be, this is also due the small data. Therefore, this data does not give a clear view of what the distribution could be.



The numerically data of Cloud seeded has a sample size of 26, a mean of 441.98, median of 221.60, standard deviation of 650.787 and a variance of 423523.9. The minimum value in this data is 4.1 and the maximum value is 2745.6. The IQR = 307.9.

b) Accuracy is the standard deviation of the data, so accuracy is 650.787.

c) The bootstrap method we used is the empirical bootstrap, when you are not certain of the distribution it is safer to use the empirical bootstrap method and the empirical bootstrap method makes no assumptions about how your observations are distributed.

Pn= (X1, . . . ,Xn)~clouds.txt$seeded of a unknown distribution and the Tn = accuracy(x) as the statistics. The standard deviation we got from this sample is 177.93.



d) Accuracy is the MAD of the data, so accuracy is 229.9513.

Pn= (X1, . . . ,Xn)~clouds.txt$seeded of a unknown distribution and the Tn = accuracy(x) as the statistics. The standard deviation we got from this sample is 71.371.



e) We prefer the … for the accuracy because …

f) Since the data is small and looking at the graphical data we don’t assume that there is normality, so we don’t want to use t-test. We want to use a non parametric test, because these test make no assumption about the underlying distribution. The signed rank test assumes that the data is continues and symmetric around the median. In exercise a, the data doesn’t look symmetric so we won’t use the signed rank test, however the data is small so it’s sceptical that the data is not symmetric. For the signed test, it only makes the assumption that the median is unique. For small data is better to trust the p-value of the signed test.

g) first we look at the data if there is a value that is equal to 119.0, if this is true we remove it and do a conditionally test on the data. Since there is a we use the following test.

Sign test

The **test statistic** is

The distribution of the test statistic under is

The test score for the sample is 16 and number of trials is 25. Using the corresponding binomial test (see R code), this gives an **p-value** of 0.9461. For the confidence interval of the sign test we have to make the interval manually, because R does not provide us the correct interval.

is rejected if



T = 9, then 302.8, 334.1, … exceed m\_0, so m\_0 < 302.8. T = 18, then 115.3, 118.3, … exceed m\_0, but 92.4 does not, so 92.4 < m\_0. Now we have to look if 92.4 and 302.8 are in the confidence interval and this has to be checked separately, in a conditional test with n = 24.

For H\_0 : m = 92.4 and T=18, check P(T>=18) = 1- P(T<=17):

1-pbinom(18-1,24,0.5)

[1] 0.01132792

For H\_0 : m = 302.8 and T=9, check P(T>=9) = 1- P(T<=8):

> 1-pbinom(9-1,24,0.5)

[1] 0.9242052

The 95-percent confidence interval of this sign test is (92.4, 302.8].

**Conclusion:** Since this p-value is greater than the significance level, we do not reject the null hypothesis. Therefore, we conclude that the median from the underlying distribution is (with significant probability) less then 119.

h) **Sign test**

The **test statistic** is

The distribution of the test statistic under is .

The test score for the sample is 16 and number of trials is 25. Using the corresponding binomial test (see R code), this gives an **p-value** of 0.9461. For the confidence interval of the sign test we have to make the interval manually, because R does not provide us the correct interval.

is rejected if



T = 6, then 489.1, 703.4, … exceed m\_0, so m\_0 < 489.1. T = 21, then 32.7, 40.6, … exceed m\_0, but 31.4 does not, so 31.4 < m\_0. Now we have to look if 31.4 and 489.1 are in the confidence interval and this has to be checked separately, in a conditional test with n = 24.

For H\_0 : m = 31.4 and T=21, check P(T>=21) = 1- P(T<=20):

1-pbinom(21-1,24,0.5)

[1] 0.0001385808

For H\_0 : m = 302.8 and T=9, check P(T>=9) = 1- P(T<=8):

1-pbinom(6-1,24,0.5)

[1] 0.9966946

The 99-percent confidence interval of this sign test is (31.4, 489.1].

**Conclusion:** Since this p-value is greater than the significance level, we do not reject the null hypothesis. Therefore, we conclude that the median from the underlying distribution is (with significant probability) less then 119.

**Signed rank test**

The **test statistic** is ,

The distribution of the test statistic under is ,

The signed rank test gives V = 263 for the sample. Using the corresponding Wilcoxon signed rank test (see R code), this gives an **p-value** of 0.9967 and the 99-percent confidence interval (-, 704).

**Conclusion:** Since this p-value is greater than the significance level, we do not reject the null hypothesis. Therefore, we conclude that the median from the underlying distribution is (with significant probability) less then 119.

**t-test**

The **test statistic** is ,

The distribution of the test statistic under is

The t-test for the sample gives t = 2.5306 and df = 25. Using the corresponding t-test (see R code), this gives an **p-value** of 0.991 and the 99-percent confidence interval (-, 784.265).

**Conclusion:** Since this p-value is greater than the significance level, we do not reject the null hypothesis. Therefore, we conclude that the median from the underlying distribution is (with significant probability) less then 119.

i) As in exercise f, we prefer the signed test. The data is small and there is no certainty for the distribution. Also the interval for the signed test is narrower so there is more precision of the true mean of the data.