Rebecca Leu

Northeastern University

Assignment 2

R functions: conducting hypothesis tests

PART A

In this assignment we will be exploring a few different data sets and running hypothesis tests of the data using R. To start we are loading in the “MASS” package into R as all of the data sets we are using in this assignment are in the MASS package.

> install.packages("MASS")

Error in install.packages : Updating loaded packages

> library(MASS)

PART B

In part B we are working with the “chem” dataset. This dataset records the amount of copper in whole meal flour in parts per million. There are 24 observations in the data. The question we are looking to answer was, “Is the production company producing whole meal flour with greater than 1 part per million copper in it?”

To answer this question, we will be using R to run a one sample t-test. First we load the data into R. The code to run this type of t-test includes specifying the data set to pull from (“chem”), mu=1 is the null hypothesis which is the mean is one. The alternative hypothesis is that the copper is greater than 1 part per million. The results in R are below

> chem

[1] 2.90 3.10 3.40 3.40 3.70 3.70 2.80 2.50 2.40 2.40 2.70

[12] 2.20 5.28 3.37 3.03 3.03 28.95 3.77 3.40 2.20 3.50 3.60

[23] 3.70 3.70

> data(chem)

> t.test(chem, mu=1, alternative = "greater")

One Sample t-test

data: chem

t = 3.0337, df = 23, p-value = 0.002952

alternative hypothesis: true mean is greater than 1

95 percent confidence interval:

2.427162 Inf

sample estimates:

mean of x

4.280417

The test tells us that the p-value is less than .05 which tells us that the result are very significant. The confidence interval is 2.427162 and infinity. This tells us that the mean is greater than one 95% of the time. In conclusion, the flour production company produces whole meal flour with greater than 1 part per million copper in it.

PART C

Next we will be looking at the cats data set. The cats dataset has three variables, sex, body weight, and heart weight. For this example, we will only be looking at sex and bodyweight. The question we are looking to answer with a hypothesis test was “do male and female cat samples have the same body weight?” To do this we will be running a two sample t test. To begin, we first must segment out the female and male observations. To this this we create subset variables male and female using the below code.

> data(cats)

> male <-subset(cats, subset=(cats$Sex=='M'))

> female <-subset(cats, subset=(cats$Sex=='F'))

Once we have broken up the data between the sex of the cat we can run the t test of the samples.

> t.test(male$Bwt, female$Bwt)

Welch Two Sample t-test

data: male$Bwt and female$Bwt

t = 8.7095, df = 136.84, p-value = 8.831e-15

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.4177242 0.6631268

sample estimates:

mean of x mean of y

2.900000 2.359574

From this we can conclude that the difference of the means is not equal to 0 since the confidence interval does not include 0. We can also conclude that this data is very significant since the p value is a very small percentage. We have rejected the null hypothesis that there is not a difference in body weight sizes between male and female cats. In conclusion, male and female cats in this sample do not have the same bodyweight.

PART D

For this section we will be using the shoes dataset to answer the question “did material A wear better than material B?”

The shoes data takes 10 boys and gives them shoe A and shoe B, one for each foot, and measures how each material wears. Since each boy has one of each shoe, this pairs the data. To answer our question, we will be running a paired t-test. I began by first looking at the mean of each dataset to get an idea of what each materials wear on its own.

> shoes

$A

[1] 13.2 8.2 10.9 14.3 10.7 6.6 9.5 10.8 8.8 13.3

$B

[1] 14.0 8.8 11.2 14.2 11.8 6.4 9.8 11.3 9.3 13.6

> mean(shoes$A)

[1] 10.63

> mean(shoes$B)

[1] 11.04

In the t test code we need to specify the dataset being used, that we are comparing A and B and that the data is paired. Lastly specifying the alternative hypothesis that the difference in means is less than 0.

> t.test(shoes$A, shoes$B, paired = TRUE, alternative = "less")

Paired t-test

data: shoes$A and shoes$B

t = -3.3489, df = 9, p-value = 0.004269

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.1855736

sample estimates:

mean of the differences

-0.41

The null hypothesis in this test would be that the true difference in means is equal to 0, meaning material A has less wear than material B. The alternative hypothesis is the true difference in means is less than 0 meaning that B has more wear then A. In conclusion we would confirm that material A does wear better than material B. This is because we find that the difference in means in less than 0 and that number is included in the confidence interval.

PART E

The next hypothesis test will test equal or given proportions. The bacterial dataset looks at placebo and active drug participants to see if the drug changed the presents of bacteria. The question we are looking to answer is, “did the drug treatment have a significant effect of the presence of the bacteria compared with the placebo?” To determine the answer to this we will be running a two-sample test for the equality of proportions. To do this we be using just the yes or no column and the active or placebo column on the table in the prop.test.

> data(bacteria)

> prop.test(table(bacteria$y, bacteria$ap)

2-sample test for equality of proportions with continuity

correction

data: table(bacteria$y, bacteria$ap)

X-squared = 4.6109, df = 1, p-value = 0.03177

alternative hypothesis: two.sided

95 percent confidence interval:

0.02813119 0.36288182

sample estimates:

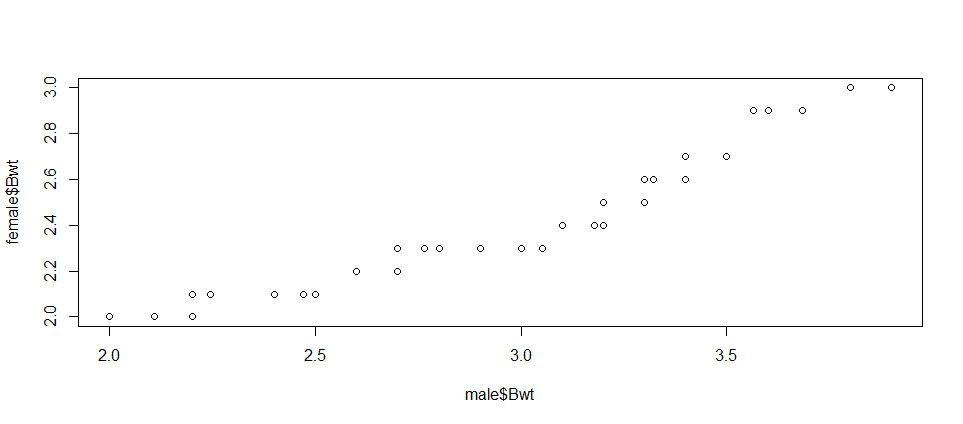
prop 1 prop 2

0.7209302 0.5254237

The prop test concludes that since the p-value is less than .05 we can conclude that the proportion of bacteria is significantly different in the two groups, so yes, the drug treatment did have an effect on the presence of bacteria compared to the placebo.

PART F

In the final part of the assignment we are going back to the cats data used in part C to use an F test for the variance of body weight in male and female cats. The begin we first test that the data follows a normal curve. I did this visually by plotting a qq normality plot. The data follows a normal line. From here we continue with our test.



> data(cats)

> male <-subset(cats, subset=(cats$Sex=='M'))

> female <-subset(cats, subset=(cats$Sex=='F'))

> qqplot(male$Bwt, female$Bwt)

> var.test(male$Bwt,female$Bwt, data=cats)

F test to compare two variances

data: male$Bwt and female$Bwt

F = 2.9112, num df = 96, denom df = 46, p-value = 0.0001157

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

1.723106 4.703057

sample estimates:

ratio of variances

2.911196

The F test is p=.0001157 which is less than the significance level of .05. In conclusion there is a significant difference between the two variances. This tells us that there is a difference between the spread of the weights for male and female cats. Looking at the data, it appears that the males have a larger spread then the females to which would match our conclusion from this F test.

References

F-Test: Compare Two Variances in R. (n.d.). Retrieved April 19, 2020, from http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r#compute-f-test-in-r

Spector, P. (n.d.). Department of Statistics. Retrieved April 19, 2020, from https://statistics.berkeley.edu/computing/r-t-tests

stats. (n.d.). Retrieved April 19, 2020, from https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/prop.test