Mustafa Abdullayev

Hypothesis Testing with R

AY6015 – Intermediate Analytics

Term: Winter 2020 - A

Instructor: Joe Manseau

01/20/2020

# Introduction

This is a Microsoft Word Report accompanying R Script. It contains my R code, outputs, my comments, and findings. In my analysis, I used 4 datasets: chem, cats, shoes, and bacteria. Explanations for all datasets can be found [here](https://cran.r-project.org/web/packages/MASS/MASS.pdf). These are built-in datasets in R. Data was cleaned, organized and ready for analysis. My main aim was to utilize R and its statistical methods to carry out hypotheses testing. I mainly used various statistical tests, (i.e T-test) in order to prove or disprove various hypotheses. Since I also provided R script with all the codes and comments, I removed some of the codes and comments from my report (such as package loading). It is due to keep my report brief, succinct and to the point.

# 

# Hypotheses Testing

I divided my report into 5 parts. They start from B since Part A is about package loading and other technical things, which are not necessary to add. I used 3 decimal notations in my report but every exact number is in the output part. For all tests, I chose my significance level as 0.05 (95% confidence). I noted the null hypothesis as H0, and alternative with HA. H0(null) is my default position. I assume this to be true unless we have evidence to the contrary. Also, I explicitly stated my hypotheses in the output part as comments. This will allow me to put technical things on the output part alongside my code. Moreover, In the explanation part, I usually used plain language in order to explain my findings and better engage with the non-technical audience.

## Part B

Chem dataset is about the determinations of copper in wholemeal flour, in parts per million. I will use this dataset in order to test the following hypothesis: “is the flour production company producing wholemeal flour with greater than 1 part per million copper in it ?”. In order to test this hypothesis, I am going to carry out a One-Sample T-test. My test is a right-sided test since my alternative hypothesis is concerned if the part amount is greater than the default amount

The P-value for this test (details below) is 0.003, which is extremely small (smaller than the significance level). So, getting the current result, under the assumption that the null hypothesis is true, is extremely unlikely. Therefore, my answer is yes and I can reject the null hypothesis, stating that the copper amount is not bigger than 1 part per million, with 95% confidence. So there is evidence that this company produces wholemeal flour with greater than 1 part per million copper in it.

*# H0: copper parts per million <= 1 part*  
*# HA: copper parts per million > 1 part*   
*# I will use 0.05 as a significance of my test.*  
*# In my test, my mu value is 1 and my test is right-sided.i.e greater*  
  
**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 is 4.280417  
*### As a result, my p-value is 0.002952. So since is it smaller than*  
*### 0.05 (significance level), it means it is highly unlikely that*   
*### test result is due to luck. So I reject the null hypothesis*   
*### staying that copper parts per million is smaller than 1. Answer is Yes :*  
*### Flour production company produces wholemeal flour which has*  
*### more than 1 part per million copper in it.*

## 

## Part C

Cats dataset contains the heart and body weights of samples of male and female cats. I will use this dataset in order to test the following hypothesis: “do male and females cat samples have the same bodyweight ?” In order to test this hypothesis, I am going to carry out a Two-Sample T-test, with a 0.05 significance level. My test is a two-sided test since there is no direction in the alternative hypothesis. I am only concerned about if mean values are the same or not.

The P-value for this test is extremely small (smaller than the significance level). So, getting that result, under the assumption that the null hypothesis is true, is extremely unlikely. Therefore, my answer is yes. I can reject the null hypothesis, which states male and female cat samples have the same body weight, with 95% confidence. So there is evidence that the bodyweight of male and female sample cats is not the same.

*# H0 : means are equal : m1 = m2*  
*# HA : means are not equal : m1 != m2*  
*# alpha = 0.05 (95% confidence)*  
*# my test is two sided*  
  
male <- cats[**which**(cats**$**Sex**==**"M"),]**$**Bwt  
female <- cats[**which**(cats**$**Sex**==**"F"),]**$**Bwt  
  
**t.test**(male,female,alternative = "two.sided")  
## Welch Two Sample t-test  
  
## data: male and female  
## 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 are 2.900000 and 2.359574  
*### my p-value is extremely small, so I can reject the null hypothesis :*  
*### I reject the null hypothesis, stating that male and female*  
*### cats have the same bodyweight, with 95% confidence*

## 

## Part D

Shoes dataset is a list of two vectors about shoe materials A and B. I will use this dataset in order to test the following hypothesis: “did material A wear better than material B?”. In order to test this hypothesis, I am going to carry out a Paired T-test, with a 0.05 significance level. I will use a new parameter called the difference. I calculate the difference to be a difference between material A and B (A - B for all values in vectors). A positive value means A is better. So my alternative hypothesis which says material A is better than material B means that the difference is higher than 0. My test is a right-sided test since my alternative hypothesis is concerned if the difference is greater than 0.

The P-value for this test (details below) is 0.996, which is extremely high (higher than the significance level). So, getting that result, under the assumption that the null hypothesis is true, is extremely likely. Therefore, my answer is No. I am unable to reject the null hypothesis, stating that the true difference in means is not bigger than 0. That, in turn, means that I must stick with my default choice. So there is no evidence that material A wears better than material B.

*# H0: material B is not worse than A: difference <= 0*  
*# HA: material A is better than B: difference > 0*   
*# alpha = 0.05 (95% confidence)*  
*# my test is one-sided*

A <- shoes**$**A  
B <- shoes**$**B  
  
**t.test**(A,B,paired = TRUE,alternative = "greater")   
## Paired t-test  
  
## data: A and B  
## t = -3.3489, df = 9, p-value = 0.9957  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.6344264 Inf  
## sample estimates:  
## mean of the differences -0.41  
*### my p-value is big (almost 1). So I am unable to reject null hypothesis*  
*### I have to accept default choice : Material B is not worse than material A* *, with 95% confidence*

## 

## Part E

Bacteria dataset is about the presence of the bacteria in two different groups after the drug treatment. I will use this dataset in order to test the following hypothesis: “did the drug treatment have a significant effect on the presence of the bacteria compared with the placebo ?”. In order to test this hypothesis, I am going to carry out a proportion test, with a 0.05 significance level. My test is a two-tailed since my alternative hypothesis states that there is an effect(no direction). It does not state that if the drug works (i.e p2 is smaller than p1). Instead, it states that if there is a difference in the proportions in 2 different groups (positive or negative effect).

The P-value for this test (details below) is 0.032, which is small (smaller than the significance level). So, getting that result, under the assumption that the null hypothesis is true, is unlikely. Therefore, my answer is yes. I can reject the null hypothesis, stating that the proportions of people with the bacteria in two different groups are the same, with 95% confidence. So there is a piece of evidence that drug has an effect.

I did not carry out a directional test. So I can not say that drug works in a positive way (i.e positive effect).

*# H0 : proportions are equal (no effect) : p1 = p2*  
*# H1 : proportions are not equal (yes effect) : p1 != p2*   
*#significance level = 0.05*  
*# this is two sided test (since question asks if there is an effect)*  
  
*# I only need first 2 columns to find placebo and active group members count*   
mydata <- **as.data.frame**(bacteria[1**:**2])  
  
*# Number of people in placebo group*  
size\_p\_all <- **nrow**(mydata[**which**(mydata**$**ap **==** "p"),])  
  
*# Number of people in active group*  
size\_a\_all <- **nrow**(mydata[**which**(mydata**$**ap **==** "a"),])  
  
*# Number of people in placebo group who has bacteria (success)*  
size\_p\_bacteria <- **nrow**(mydata[**which**(mydata**$**ap **==** "p" **&** mydata**$**y **==** "y"),])  
*# Number of people in active group who has bacteria(success)*  
size\_a\_bacteria <- **nrow**(mydata[**which**(mydata**$**ap **==** "a" **&** mydata**$**y **==** "y"),])  
  
**prop.test**(x = **c**(size\_a\_bacteria,size\_p\_bacteria),n = **c**(size\_a\_all,size\_p\_all))

2-sample test for equality of proportions with continuity correction  
  
## data: c(size\_a\_bacteria, size\_p\_bacteria) out of c(size\_a\_all, size\_p\_all)  
## X-squared = 4.6109, df = 1, p-value = 0.03177  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.23516294 -0.01483706  
## sample estimates:  
## prop 1 prop 2 0.750 0.875  
*### p-value of this test is 0.03177 which is smaller than 0.05 (significance level)*  
*### So, I can reject the null hypothesis. With 95% confidence,*  
*### drug treatment has a significant effect on the presence of the bacteria*

## 

## Part F

In this part, I will use the cat dataset to check that if there is a difference in the variation of male and female cat’s bodyweights. In order to test this hypothesis, I am going to carry out an F-test, with a 0.05 significance. My test is a two-sided test since there is no direction in the alternative hypothesis. We are only concerned if variances are the same or not.

The P-value for this test is extremely small (smaller than the significance level). So, getting that result, under the assumption that the null hypothesis is true, is extremely unlikely. Therefore, my answer is yes. I can reject the null hypothesis, which states male and female cat samples have the same variation in weight, with 95% confidence. So there is evidence that the variation of body weight of male and female sample cats is different.

*# H0 : variances are equal : v1 = v2*  
*# HA : variances are not equal : v1 != v2*   
*# alpha = 0.05 (95% confidence)*  
*# my test is two-sided*  
  
**var.test**(male,female,alternative = "two.sided")   
## F test to compare two variances  
   
## data: male and female  
## 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 is : 2.911196  
*### p-value is extremely small. So, I reject null hypothesis : ,*  
*### I am 95% confident that male and female cats' bodyweights do not have same variances*

# Conclusion

To conclude, I carried out various statistical tests with the aid of R. I utilized 4 different datasets namely cats, shoes, bacteria and chem. These were built-in datasets provided by R. My main aim was to utilize these datasets in order to prove/disprove some hypotheses. In order to achieve that I conducted various statistical tests (i.e F-test). The most interesting test was in Part E. In the problem, it does not ask if the drug had a positive effect. But one can easily assume so. But, this means I have to carry out the directional test, which was not the case.

# Reference

Ripley, B. (2019, December 20). Support Functions and Datasets for Venables and Ripley's MASS [R package MASS version 7.3-51.5]. Retrieved from [Package MASS](https://cran.r-project.org/web/packages/MASS)