Northeastern University

R Practice 2 ALY6015

R practice week 2 - Module 2 Assignment — Chi-Square Testing and ANOVA

by Mohammad Hossein Movahedi

movahed.m@northeastern.edu

27 April 2022

Table Of content

Introduction

This R practice mainly focuses on Chi-Square Testing and ANOVA testing, and during this assignment, I will use `baseball.csv` and `crop_data.csv` datasets to solve various problems.

Analysis

This part will go through the steps described in the assignment and solve problems with Chi-Square Testing and ANOVA testing methods.

Section 11-1

in this section, I will perform the following steps

- State the hypotheses and identify the claim.
- Find the critical value.
- Compute the test value.
- Make the decision.

- Summarize the results.

Question 6: Blood Types

> A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population. The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%. He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood. At α = 0.10, can it be concluded that the distribution is the same as that of the general population?

First, I formalize the question

Blood type	General population	observation out of 50
A	20%	12
B	28%	8
O	36%	24
AB	16%	6

then I state the hypotheses, identify the claim and define the critical value

$$H_0 : \text{the sample data comes from the general distribution}$$
$$H_1 : \text{the sample data comes a different general distribution}$$
$$\begin{cases} H_0 : P_0 = P_1 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where P is the distribution of sample}$$

then I run the chi test on R to get the results

```
#Question 6: Blood Types
#assigning alpha
alpha <- 0.1
#vector of observation
c <- c(12,8,24,6)
#vector of general distribution
p <- c(0.2,0.28,0.36,0.16)

#running chi test
r <- chisq.test(x=c,p=p)
r$statistic
r$parameter
r$p.value
```

```
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the results, the p-value of the test is `0.1403575`, the Chi test value is `5.471429`, and at last, the degree of freedom is 3 . By comparing the p-value and alpha, we conclude that we can't reject the null hypothesis.

Question 8: On-Time Performance by Airlines

> According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

Action	% of Time
On-time	70.8
National Aviation System delay	8.2
Aircraft arriving late	9
Other (because of weather and other conditions)	12

> Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of weather, ten because of a National Aviation System delay, and the rest because of arriving late. At α = 0.05, do these results differ from the government's statistics?

This question is similar to the previous question, so I follow the same steps.

I state the hypotheses, identify the claim and define the critical value

$$H_0 : \text{the sample data comes from the government's statistics' distribution}$$
$$H_1 : \text{the sample data comes a different distribution}$$
$$\begin{cases} H_0 : P_0 = P_1 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } P_0 \text{ is the distribution of sample and } P_1 \text{ is the distribution of statistics}$$

now I use R to solve the problem.

```
#assigning alpha
alpha <- 0.05
#vector of observation
c <- c(125,10,25,40)
#vector of general distribution
p <- c(0.708,0.082,0.09,0.12)

#running chi test
```

```
r <- chisq.test(x=c,p=p)
r$statistic
r$parameter
r$p.value

#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the results, the p-value of the test is `0.0004762587`, the Chi test value is `17.8325`, and at last, the degree of freedom is 3 . By comparing the p-value and alpha, we conclude that we reject the null hypothesis.

Section 11-2

In this section, I will follow the following steps

1. State the hypotheses and identify the claim.
2. Find the critical value.
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

Question 8:Ethnicity and Movie Admissions

> Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

	Caucasian	Hispanic	African American	Other
2013	724	335	174	107
2014	370	292	152	140

First, I state the hypothesis.

$$H_0 : \text{movie attendance by year was independent upon ethnicity}$$
$$H_1 : \text{movie attendance by year was dependent upon ethnicity}$$
$$\begin{cases} H_0 : X_0^2 = X_1^2 \\ H_1 : \text{otherwise} \end{cases}$$

Where X_0^2 is the expected Chi distribution and X_1^2 is Chi distribution of statistics

```
#section 11-2
#question 8
#assigning alpha
```

```
alpha <- 0.05
#vector of observation
c2013 <- c(724,335,174,107)
c2014 <- c(370,292,152,140)
#stating n of rows
rows = 2
#creating a matrix of rows
mt <- matrix(c(c2013,c2014),nrow = rows,byrow  = T)
rownames(mt) = c(2013,2014)
colnames(mt) = c("Caucasian","Hispanic","African American","Other")

#doing the test
r <- chisq.test(mt)
r$p.value
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the results, the p-value of the test is `5.477507e-13`, the Chi test value is `60.144`, and at last, the degree of freedom is 3 . By comparing the p-value and alpha, we conclude that we reject the null hypothesis. Therefore we can't say that movie attendance by year was independent of ethnicity.

Question 10: Women in the Military

> This table lists the numbers of officers and enlisted personnel for women in the military. At α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

Action	Officers	Enlisted
Army	10,791	62,491
Navy	7,816	42,750
Marine Corps	932	9,525
Air Force	11,819	54,344

First, I state the hypothesis.

$$H_0 : \text{no relationship exists between rank and branch of the Armed Forces}$$
$$H_1 : \text{a relationship exists between rank and branch of the Armed Forces}$$
$$\begin{cases} H_0 : X_0^2 = X_1^2 \\ H_1 : \text{otherwise} \end{cases}$$

Where X_0^2 is the expected Chi distribution and X_1^2 is Chi distribution of statistics

then I conduct the Chi test

```
#question 10
#assigning alpha
alpha <- 0.05
#vector of observation
```

```
Army<-c(10791,62491)
Navy<-c(7816,42750)
Marine_Corps<-c(932,9525)
Air_Force<-c(11819,54344)
#stating n of rows
rows = 4
#creating a matrix of rows
mt <- matrix(c(Army,Navy,Marine_Corps,Air_Force),nrow = rows,byrow  = T)
rownames(mt) = c("Army","Navy", "Marine_Corps","Air_Force")
colnames(mt) = c("Officers","Enlisted")

#doing the test
r <- chisq.test(mt)
r$p.value
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the results, the p-value of the test is `1.726418e-141`, the Chi test value is `654.27`, and at last, the degree of freedom is 3 . By comparing the p-value and alpha, we conclude that we reject the null hypothesis. Therefore we can't say that no relationship exists between rank and branch of the Armed Forces.

Section 12-1

This section assumes that all variables are normally distributed, that the samples are independent, that the population variances are equal, and that the samples are simple random samples, one from each population.

Question 8: Sodium Contents of Foods

> The sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

Condiments	Cereals	Desserts
270	260	100
130	220	180
230	290	250
180	290	250
80	200	300
70	320	360
200	140	300
		160

First, I state the hypothesis.

$$H_0 : \text{there is no difference in mean sodium amounts}$$
$$H_1 : \text{there is a difference in mean sodium amounts}$$
$$\begin{cases} H_0 : \overline{X}_0 = \overline{X}_1, \overline{X}_1 = \overline{X}_2, \overline{X}_2 = \overline{X}_0 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } \overline{X} \text{ is the mean of sodium in each product}$$

Then I ran the T tests for each pairs

```
#question 8
#assigning alpha
alpha <- 0.05
#vector of observation
Condiments<-c(270,
              130,
              230,
              180,
              80,
              70,
              200)
Cereals<-c(260,
           220,
           290,
           290,
           200,
           320,
           140)
Desserts<-c(100,
            180,
            250,
            250,
            300,
            360,
            300,
            160)
#running t.test for each pairs
r <- t.test(Cereals,Condiments)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
r <- t.test(Cereals,Desserts)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
r <- t.test(Condiments,Desserts)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
```

According to the results, non of the tests rejected that there is a significant difference. Therefore, we can't reject the null hypothesis that there is no difference in means.

Section 12-2

In this section, I finally do something new. I perform a complete one-way ANOVA. If the null hypothesis is rejected, I use either the Scheffé or Tukey test to see if there is a significant difference in the pairs of means. I assume all assumptions are met.

Question 10: Sales for Leading Companies

> The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means?

Cereal	Chocolate Candy	Coffee
578	311	261
320	106	185
264	109	302
249	125	689
237	173	0

First, I state the hypothesis.

$$H_0 : \text{there is no significant difference in the means}$$
$$H_1 : \text{there is a significant difference in the means}$$
$$\begin{cases} H_0 : \overline{X}_0 = \overline{X}_1 = \overline{X}_2 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } \overline{X} \text{ is the mean of each of each product}$$

Then I ran the one-way ANOVA

```
#question  10
#assigning alpha
alpha <- 0.05
#input observations
Cereal<-data.frame('data'=c(578,
                            320,
                            264,
                            249,
                            237),'Food'="Cereal")
Chocolate<-data.frame('data'=c(311,
                                  106,
                                  109,
                                  125,
                                  173),'Food'="Chocolate")
Coffee<-data.frame('data'=c(261,
                            185,
                            302,
                            689),'Food'="Coffee")
data <- rbind(Cereal,Chocolate,Coffee)
data$Food<-as.factor(data$Food)
#running the test
anova<-aov(data~Food,data = data)
summary(anova)
#extracting P-value
p<- summary(anova)
pv<-p[[1]][[1,'Pr(>F)']]
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the one-way ANOVA test results, the null hypothesis is not rejected, and there is no significant difference between means of populations.

Question 12: Per-Pupil Expenditures

> The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using α = 0.05, can you conclude that there is a difference in means?

Eastern third	Middle third	Western third
4946	6149	5282
5953	7451	8605
6202	6000	6528
7243	6479	6911
6113	0	0

First, I state the hypothesis.

$$H_0 : \text{there is no significant difference in the means}$$
$$H_1 : \text{there is a significant difference in the means}$$
$$\begin{cases} H_0 : \overline{X}_0 = \overline{X}_1 = \overline{X}_2 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } \overline{X} \text{ is the mean of each of each product}$$

Then I ran the test

```
#question 12
#assigning alpha
alpha <- 0.05
#input observations
inputdata<-data.frame('data'=c(4946,
                               5953,
                               6202,
                               7243,
                               6113,
                               6149,
                               7451,
                               6000,
                               6479,
                               5282,
                               8605,
                               6528,
                               6911),lable =c('Eastern','Eastern','Eastern',
                                              'Eastern','Eastern','Middle',
                                              'Middle','Middle','Middle',
                                              'Western','Western','Western',
                                              'Western'))
inputdata$lable<-as.factor(inputdata$lable)
```

```
#running the test
anova<-aov(data~lable,data = inputdata)
summary(anova)
#extracting P-value
p<- summary(anova)
pv<-p[[1]][[1,'Pr(>F)']]
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

According to the one-way ANOVA test results, the null hypothesis is not rejected, and there is no significant difference between means of populations.

Section 12-3

the steps are :

a. State the hypotheses.
b. Find the critical value for each F test.
c. Complete the summary table and find the test value.
d. Make the decision.
e. Summarize the results. (Draw a graph of the cell means if necessary.)

question 10: Increasing Plant Growth

> A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a “Grow-light” in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes.

	Grow-light 1	Grow-light 2
Plant food A	9.2, 9.4, 8.9	8.5, 9.2, 8.9
Plant food B	7.1, 7.2, 8.5	5.5, 5.8, 7.6

Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use α = 0.05.

First, I state the hypothesis.

$$H_0 : \text{there is no difference in mean growth with respect to light or plant food}$$
$$H_1 : \text{there is difference in mean growth with respect to light or plant food}$$
$$\begin{cases} H_0 : \begin{cases} \overline{X}_0 = \overline{X}_1 \\ \overline{P}_0 = \overline{P}_1 \end{cases} \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } \overline{X} \text{ and } \overline{P} \text{ is the mean of growth based on each factor}$$

Then I proceed with coding part

```
#section 12-3
#question 10
m11 <- data.frame('data'=c(9.2, 9.4, 8.9),'grow' = 1,plant='a')
m12 <- data.frame('data'=c(8.5, 9.2, 8.9),'grow' = 2,plant='a')
m21 <- data.frame('data'=c(7.1, 7.2, 8.5),'grow' = 1,plant='b')
m22 <- data.frame('data'=c(5.5, 5.8, 7.6),'grow' = 2,plant='b')
indata<-rbind(m11,m12,m21,m22)
indata$grow <- as.factor(indata$grow)
indata$plant <- as.factor(indata$plant)
anova<-aov(data~grow+plant,data=indata)
summary(anova)
```

the results are following

```
> summary(anova)
            Df Sum Sq Mean Sq F value   Pr(>F)
grow         1  1.920   1.920    3.51 0.093781 .
plant        1 12.813  12.813   23.42 0.000921 ***
Residuals    9  4.923   0.547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the results and the test alpha, grow light is a significant factor while plant type wasn’t very substantial.

Baseball dataset part

In this part, I use the provided dataset to do a chi test on the number of wins in each decade the

First I create the wins table

```
#the baseball part
install.packages('FSA')
install.packages('FSAdata')
install.packages('magrittr')
install.packages('dplyr')
install.packages('tidyr')
install.packages('plyr')
install.packages('tidyverse')
install.packages('outliers')
install.packages('ggplot2')
install.packages('lubridate')
install.packages('corrplot')

library(ggplot2)
library(outliers)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(tidyr)
library(dplyr)
```

```
library(tidyverse)
library(scales)
library(lubridate)
library(corrplot)
#read the data
bb<-read.csv('baseball.csv')

# Extract decade from year
bb$Decade <- bb$Year - (bb$Year %% 10)
# Create a wins table by summing the wins by decade
wins <- bb %>%
  group_by(Decade) %>%
  summarize(wins = sum(W)) %>%
  as.tibble()
```

then I state the hypothesis

$$H_0 : \text{the data is equally distributed}$$
$$H_1 : \text{the data is not equally distributed}$$
$$\begin{cases} H_0 : P_0 = P_1 \\ H_1 : \text{otherwise} \end{cases}$$
$$\text{Where } P_0 \text{ is the distribution of data and } P_1 \text{ is the uniform distribution}$$

then I do a Chi test on the results

```
#assigning alpha
alpha <- 0.05
p =rep(1/6,6)
r <-chisq.test(wins$wins,p =p)
r$statistic
r$p.value
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```

according to the test, the null hypothesis is strongly rejected, stating a difference in the number of wins in each decade.

Conclusion

In this assignment, I learned different ANOVA and Chi testing applications, and I used them on a real dataset.

Bibliography

Sthda.com. (2020). *Chi-square Goodness of Fit Test in R - Easy Guides - Wiki - STHDA*. [online] Available at: http://www.sthda.com/english/wiki/chi-square-goodness-of-fit-test-in-r [Accessed 2 May 2022].

chi (2019). *Test of uniform distribution using KS-test and chi square in R*. [online] Cross Validated. Available at: https://stats.stackexchange.com/questions/406406/test-of-uniform-distribution-using-ks-test-and-chi-square-in-r [Accessed 2 May 2022].

Appendix

```
print('Mohammad Hossein Movhaedi')
print('assignment 2')
install.packages('Hmisc')
library(Hmisc)
#Question 6: Blood Types
#assigning alpha
alpha <- 0.1
#vector of observation
c <- c(12,8,24,6)
#vector of general distribution
p <- c(0.2,0.28,0.36,0.16)

#running chi test
r <- chisq.test(x=c,p=p)
r$statistic
r$parameter
r$p.value

#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#question 8
#assigning alpha
alpha <- 0.05
#vector of observation
c <- c(125,10,25,40)
#vector of general distribution
p <- c(0.708,0.082,0.09,0.12)

#running chi test
r <- chisq.test(x=c,p=p)
r$statistic
r$parameter
r$p.value

#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")

#section 11-2
#question 8
#assigning alpha
alpha <- 0.05
#vector of observation
c2013 <- c(724,335,174,107)
c2014 <- c(370,292,152,140)
#stating n of rows
rows = 2
#creating a matrix of rows
mt <- matrix(c(c2013,c2014),nrow = rows,byrow  = T)
rownames(mt) = c(2013,2014)
colnames(mt) = c("Caucasian","Hispanic","African American","Other")

#doing the test
r <- chisq.test(mt)
```

```
r$p.value
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")

#question 10
#assigning alpha
alpha <- 0.05
#vector of observation
Army<-c(10791,62491)
Navy<-c(7816,42750)
Marine_Corps<-c(932,9525)
Air_Force<-c(11819,54344)
#stating n of rows
rows = 4
#creating a matrix of rows
mt <- matrix(c(Army,Navy,Marine_Corps,Air_Force),nrow = rows,byrow  = T)
rownames(mt) = c("Army","Navy", "Marine_Corps","Air_Force")
colnames(mt) = c("Officers","Enlisted")

#doing the test
r <- chisq.test(mt)
r$p.value
r
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#section 12-1
#question 8
#assigning alpha
alpha <- 0.05
#vector of observation
Condiments<-c(270,
              130,
              230,
              180,
              80,
              70,
              200)
Cereals<-c(260,
           220,
           290,
           290,
           200,
           320,
           140)
Desserts<-c(100,
            180,
            250,
            250,
            300,
            360,
            300,
            160)
#running t.test for each pairs
r <- t.test(Cereals,Condiments)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
r <- t.test(Cereals,Desserts)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
r <- t.test(Condiments,Desserts)
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#not rejected
#section 12-2
#question  10
```

```
#assigning alpha
alpha <- 0.05
#input observations
Cereal<-data.frame('data'=c(578,
                            320,
                            264,
                            249,
                            237),'Food'="Cereal")
Chocolate<-data.frame('data'=c(311,
                               106,
                               109,
                               125,
                               173),'Food'="Chocolate")
Coffee<-data.frame('data'=c(261,
                            185,
                            302,
                            689),'Food'="Coffee")
data <- rbind(Cereal,Chocolate,Coffee)
data$Food<-as.factor(data$Food)
#running the test
anova<-aov(data~Food,data = data)
summary(anova)
#extracting P-value
p<- summary(anova)
pv<-p[[1]][[1,'Pr(>F)']]
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#question 12
#assigning alpha
alpha <- 0.05
#input observations
inputdata<-data.frame('data'=c(4946,
                               5953,
                               6202,
                               7243,
                               6113,
                               6149,
                               7451,
                               6000,
                               6479,
                               5282,
                               8605,
                               6528,
                               6911),lable =c('Eastern','Eastern','Eastern',
                                              'Eastern','Eastern','Middle',
                                              'Middle','Middle','Middle',
                                              'Western','Western','Western',
                                              'Western'))
inputdata$lable<-as.factor(inputdata$lable)
#running the test
anova<-aov(data~lable,data = inputdata)
summary(anova)
#extracting P-value
p<- summary(anova)
pv<-p[[1]][[1,'Pr(>F)']]
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
#section 12-3
#question 10
m11 <- data.frame('data'=c(9.2, 9.4, 8.9),'grow' = 1,plant='a')
m12 <- data.frame('data'=c(8.5, 9.2, 8.9),'grow' = 2,plant='a')
m21 <- data.frame('data'=c(7.1, 7.2, 8.5),'grow' = 1,plant='b')
m22 <- data.frame('data'=c(5.5, 5.8, 7.6),'grow' = 2,plant='b')
indata<-rbind(m11,m12,m21,m22)
indata$grow <- as.factor(indata$grow)
indata$plant <- as.factor(indata$plant)
```

```
anova<-aov(data~grow+plant,data=indata)
summary(anova)

#the baseball part
install.packages('FSA')
install.packages('FSAdata')
install.packages('magrittr')
install.packages('dplyr')
install.packages('tidyr')
install.packages('plyr')
install.packages('tidyverse')
install.packages('outliers')
install.packages('ggplot2')
install.packages('lubridate')
install.packages('corrplot')

library(ggplot2)
library(outliers)
library(FSA)
library(FSAdata)
library(magrittr)
library(dplyr)
library(tidyr)
library(dplyr)
library(tidyverse)
library(scales)
library(lubridate)
library(corrplot)
#read the data
bb<-read.csv('baseball.csv')

# Extract decade from year
bb$Decade <- bb$Year - (bb$Year %% 10)
# Create a wins table by summing the wins by decade
wins <- bb %>%
  group_by(Decade) %>%
  summarize(wins = sum(W)) %>%
  as.tibble()
#assigning alpha
alpha <- 0.05
p =rep(1/6,6)
r <-chisq.test(wins$wins,p =p)
r$statistic
r$p.value
#cheaking the p-value
ifelse(r$p.value > alpha,"H0 is not rejected","H0 is rejected")
```