CHI-SQUARE TESTING AND ANOVA

ALY6015: Intermediate Analytics

DHRUV VIJAY GUJRATHI

Northeastern University

DHRUV VIJAY GUJRATHI, College of Professional Studies, Northeastern University, Boston, MA 02115.

This report is regarding chi-square testing and ANOVA.

Dhruv Gujrathi is now a student at Department of Analytics, Northeastern University

Contact: gujrathi.d@northeastrn.edu

(NUID: 001029464)

INTRODUCTION

In this assignment, we are required to solve real world problems using the chi-square test or the ANOVA test based on the type of variables given and the type of tests that are to be conducted. Chi square test is mainly used when the sample data or the given data matches the population. There are two types of chi-square tests, namely, chi-square test of good fit and chi-square test for independence. The chi-square test for independence is used to compare two variables to see if they are related or not.

The ANOVA test stands for Analysis Of Variance. It is used to evaluate the effect of two grouping variables on a response variable. The grouping variables are known as factors. The group of these factors are called levels and a combination of cells is known as a cell.
When the sample sizes within cells are equal, we have the so-called balanced design. In this case the standard two-way ANOVA test can be applied.
When the sample sizes within each level of the independent variables are not the same (case of unbalanced designs), the ANOVA test should be handled differently.

ANALYSIS

1. Section 11-1

In this question we are asked to determine if the blood types of patients admitted in a hospital is same as the ones in the population. We are given a general distribution of the blood types in the hospital. Using these values, we conduct the chi-square test and determine if the distribution is same as that of the general population.
For alpha=0.10 we do not reject null hypothesis because our p value is greater than 0.10 and the chi square value is less than the critical value. We conclude that the distribution of blood type of the sample is likely the same as the distribution of the blood type in the population.

```
#Question 1 Blood Types

#ho: The distribution of blood type in the sample is the same as the distribution of the blood type of the population
chisq.test(x=c(12,8,24,6),
         p=c(0.2,0.28,0.36,0.16))

##
##  Chi-squared test for given probabilities
##
## data:  c(12, 8, 24, 6)
## X-squared = 5.4714, df = 3, p-value = 0.1404

#critical value assuming alpha is 0.10
```

2. In this question we are asked to determine if the reasons for delay of an airline matches with the ones in the government database. We are given a general distribution of the reasons for delay of the airlines. Using these values, we conduct the chi-square test and determine if the distribution is same as that of the general population (government statistics).

   For alpha=0.05 we reject the hypothesis that the distribution of on time performance by the airline under steady do not differ from the government statistics

```
#Question 2 On Time Performance by Airlines

#ho: The distribution of on time performace by the airline of the company under the study do not differ from the government statistics

#value for national aviation system delay to be calculated as (200-125-40-10=25)

chisq.test(x=c(125,10,25,40),
           p=c(0.708,0.082,0.09,0.12))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(125, 10, 25, 40)
## X-squared = 17.832, df = 3, p-value = 0.0004763
```

3. Section 11-2

   In this question we are asked to determine if the ethnicity of a person determines that person’s admission to movies. We are given a general distribution of the relation between his/her admission to movies. Using these values, we conduct the chi-square test and determine if the distribution is same as that of the general population (yearly statistics). We reject the null hypothesis and conclude that the movie attendance is dependent on ethnicity. As p value is less than 0.05

```
#Question3 Ethnicity and Movie Admissions

#Ho=Movie attendance by year is independent of ethnicity
dat<-matrix(c(724,370,
              335,292,
              174,152,
              107,140),
            nrow=2)

chisq.test(dat)

##
##  Pearson's Chi-squared test
##
## data:  dat
## X-squared = 60.144, df = 3, p-value = 5.478e-13
```

4. In this question we have been asked to determine if the number of officers and enlisted personnel for women in military have an established relationship with rank and branch of armed forces. For alpha 0.05 we reject the null hypothesis because our p value is less than 0.05 and the chi square value is higher than the critical value. We conclude that that there is no relationship between the rank and branch of the armed forces.

```
#Question 4 Women in the Military
#HO : µ1 = µ2 = µ3

#Ho=There is no relationship between rank and branch of the armed forces
data<-matrix(c(10791,62491,
              7816,42750,
              932,9525,
              11819,54344),
              nrow=2)

chisq.test(data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 654.27, df = 3, p-value < 2.2e-16
```

```
qchisq(p=0.95,df=3)
```

```
## [1] 7.814728
```

5. Section 12-1

   In this question we have been asked to determine if the difference in mean of sodium quantities exists in condiments, cereals, and desserts. Using these values, we conduct the ANOVA test and determine if the distribution is same as that of the general population. We do not reject the null hypothesis as p value is greater than 0.05 and the F Stat is less than the critical value. There is not enough evidence to conclude that there is a difference in mean sodium amounts among the three types.

```
#Question 5 Sodium Contents of Foods
#HO : µ1 = µ2 = µ3
#the alpha is 0.01 so p needs to be less than 0.1 and

library(tidyverse)

library(Hmisc)

data2<-
tibble(y=c(270,130,230,180,80,70,200,260,220,290,290,200,320,140,100,180,250,250,300,360,300,160))
type=c(rep("Condiments", 7),
rep("Cereals",7),
rep("Desserts",8))
data2
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 22 x 1
##       y
##   <dbl>
##  1   270
##  2   130
##  3   230
##  4   180
##  5    80
##  6    70
##  7   200
##  8   260
##  9   220
## 10   290
## # ... with 12 more rows
```

```
mod_5<-lm(y~type,data=data2)


summary(aov(mod_5))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2  27544   13772   2.399  0.118
## Residuals   19 109093    5742
```

```
#critical value
qf(0.95,2,19)
```

```
## [1] 3.521893
```

6. In this question we are asked to determine the sales for a year for the companies listed. Using these values, we conduct the ANOVA test and determine if the distribution is same as that of the general population. As p value is greater than 0.01 ie 99%. There is not enough evidence that means are different. In other words, means are the same.

```
#Question 6 Sales for leading companies
#HO : µ1 = µ2 = µ
data3<-
tibble(y=c(578,320,264,249,237,311,106,109,125,173,261,185,302,689),
             type=c(rep("Cereals",5),
                    rep("Chocolate candy",5),
                        rep("Coffee",4)))
data3
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 14 x 2
##        y type
##    <dbl> <chr>
##  1   578 Cereals
##  2   320 Cereals
##  3   264 Cereals
##  4   249 Cereals
##  5   237 Cereals
##  6   311 Chocolate candy
##  7   106 Chocolate candy
##  8   109 Chocolate candy
##  9   125 Chocolate candy
## 10   173 Chocolate candy
## 11   261 Coffee
## 12   185 Coffee
## 13   302 Coffee
## 14   689 Coffee
```

```
model_6<-lm(y~type,data=data3)
summary(aov(model_6))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2 103770   51885   2.172   0.16
## Residuals   11 262795   23890
```

```
11
```

```
## [1] 11
```

```
#critical value
qf(0.99,2,11)
```

```
## [1] 7.205713
```

7. In this question we have been asked to determine the expenditures per pupil for states in three different sections of the country. Using these values, we conduct the ANOVA test and determine if the distribution is same as that of the general population. We do not reject the null hypothesis as p value is greater than 0.05 (alpha). There is not enough evidence that means are different. In other words, means are the same.

```
#Question 7 Per Pupil Expenditure
dat<-tibble(y=c(4946,5953,6202,7243,6113,6149,7451,6000,6479,5282,8605,6528,6911),
           type=c(rep("Eastern third",5),
                 rep("Middle third",4),
                 rep("Westernthirs",4)))
dat
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 13 x 2
##        y type
##    <dbl> <chr>
##  1  4946 Eastern third
##  2  5953 Eastern third
##  3  6202 Eastern third
##  4  7243 Eastern third
##  5  6113 Eastern third
##  6  6149 Middle third
##  7  7451 Middle third
##  8  6000 Middle third
##  9  6479 Middle third
## 10  5282 Westernthirs
## 11  8605 Westernthirs
## 12  6528 Westernthirs
## 13  6911 Westernthirs
```

```
model_7<-lm(y~type,data=dat)
summary(aov(model_7))
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## type         2 1244588  622294   0.649  0.543
## Residuals   10 9591145  959114
```

8. Section 12-3

   In this question, we have been asked to determine if the growth light combinations on plant food A and plant food B are related and if yes, how do they affect the growth of both the plants. We must determine which distribution to be used, and we can do that by observing the type of variables involved. The variables here are all independent variables and not mean of the given population. Thus, we can conclude that we use the chi-square test to determine the hypotheses, critical value, and the test value. For alpha=0.05 we do not reject null hypothesis because our p value is greater than 0.05 and the chi square value is less than the critical value. We conclude that the grow light does affect the growth of the plants and there is interaction between the two variables. Thus, there is a difference in mean growth with respect to light.

```
#Increasing Plant Growth

data4<-matrix(c(9.2, 9.4, 8.9,

              8.5, 9.2, 8.9,

              7.1, 7.2, 8.5,

              5.5, 5.8, 7.6),

            nrow=2) #Assigning the matrix to a variable


chisq.test(data4) #Performing the chi square test
```

CONCLUSION

In this assignment we were asked to interpret the data that was given to us and come up with a way to manipulate or model the data given to decide whether to use the chi-square test or the ANOVA test. We were also asked to determine when to use which type of test based on the conditions and the type of variables required to decide the type of test.

REFERENCES:

1. Tran, J. (2019, December 20). Exploratory Data Analysis in R for beginners. Retrieved November 11, 2020, from https://towardsdatascience.com/exploratory-data-analysis-in-r-for-beginners-fe031add7072
2. Chi-squared Distribution. (n.d.). Retrieved November 11, 2020, from http://www.r-tutor.com/elementary-statistics/probability-distributions/chi-squared-distribution
3. Siegle, D. (2015, June 14). ANOVA, Regression, and Chi-Square. Retrieved November 11, 2020, from https://researchbasics.education.uconn.edu/anova_regression_and_chi-square/

APPENDIX:

CODE:

```
#ALY6015_Assignment2_Dhruv_Gujrathi
#Section 11-1
#Blood Types
chisq.test(x=c(12,8,24,6),
           p=c(0.2,0.28,0.36,0.16)) #Conducting the Chi-sq test


#Section 11-1
#On-time Performance by Airlines
chisq.test(x=c(125,10,25,40),
           p=c(0.708,0.082,0.09,0.12)) #Conducting the Chi-sq test
#Section 11-2
#Ethinicity and Movie Admissions
dat<-matrix(c(724,370,
              335,292,
              174,152,
              107,140),
            nrow=2) #assigning the matrix to a single variable


chisq.test(dat) #Conducting the Chi-sq test


#Section 11-2
#Women in the Military
data<-matrix(c(10791,62491,
          7816,42750,

          932,9525,

          11819,54344),

         nrow=2) #Assigning the matrix to a single variable


chisq.test(data) #Conducting the Chi-sq test


#Section 12-1

#Sodium Contents of Foods

install.packages("tidyverse") #Installing 'tidyverse' package for Data Manipulation & Visualization

install.packages("Hmisc") #Installing 'Hmisc' package for Imputing missing values, high level graphics, etc.

library(tidyverse)

library(Hmisc)

data2<-tibble(y=c(270,130,230,180,80,70,200,260,220,290,290,200,320,140,100,180,250,250,300,360,300,160)) #Assigning required values to tibble.

type=c(rep("Condiments", 7),

      rep("Cereals",7),

      rep("Desserts",8)) #Determining the dependant and independant variables


data2

mod_5<-lm(y~type,data=data2)

summary(aov(mod_5)) #Performing the ANOVA test

qf(0.95,2,19) #Determining the critical value
```

```
#Section 12-2

#Sales for leading companies

data3<-
tibble(y=c(578,320,264,249,237,311,106,109,125,173,261,185,302,689),
#Assigning required values to tibble.

              type=c(rep("Cereals",5),

                     rep("Chocolate candy",5),

                     rep("Coffee",4))) #Determining the dependant and independant
variables

data3

model_6<-lm(y~type,data=data3)

summary(aov(model_6)) #Performing the ANOVA test

qf(0.99,2,11) #Determining the critical value
```

```
#Section 12-2

#Per Pupil Expenditure

dat<-
tibble(y=c(4946,5953,6202,7243,6113,6149,7451,6000,6479,5282,8605,6528,6
911), #Assigning required values to tibble.

            type=c(rep("Eastern third",5),

                   rep("Middle third",4),

                   rep("Westernthirs",4))) #Determining the dependant and
independant variables

dat

model_7<-lm(y~type,data=dat) #Performing the ANOVA test

summary(aov(model_7)) #Determining the critical value
```

```
#Section 12-3

#Increasing Plant Growth

data4<-matrix(c(9.2, 9.4, 8.9,

            8.5, 9.2, 8.9,

            7.1, 7.2, 8.5,

            5.5, 5.8, 7.6),

          nrow=2) #Assigning the matrix to a variable


chisq.test(data4) #Performing the chi square test
```