Intermediate Analytics

Module 2 Chi-Square Testing and Anova using R

ALY6015, Spring 2022

Week-2 | Group 9

Professor: Roy Wada

Submitted by: Abhinav Jain

Date: 04/24/2022

Introduction

Background

In this report performing Chi-Square testing and Anova using R programming language. There are some problems which we are going to solve using various chi-square like testing a distribution for goodness of fit, testing two variables for independence and testing proportions for homogeneity. Apart from chi square test using one- way ANOVA and two-way ANOVA technique to determine the significant difference between the variables. Major steps to perform these tests are:

- Measure the Hypothesis and find the claim
- Evaluate the critical value
- Calculate the test value
- Find out the decision
- Explain the results

Moreover, we have two datasets in this report which will be analyses after solving certain problems. One for the crop data and another one is of baseball that we will discuss later in this report while finding the outcomes as per the requirements for this report.

Task 1: Problem for Blood type

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 $\alpha = 0.10$, can it be concluded that the distribution is the same as that of the general population?

Solution:

The chi square test was done for the blood type from a random sample size of 50 patients, as shown above. The chi-square test states that the hypothesis is not significant because the alpha should be 0.10, but we receive a p-value of 0.1404, which is more than the significance level. This test's conclusion is not significant, and it is more likely to reject the null hypothesis that says two variables are independent to each other. Whereas the sample size is 4, the degree of freedom is 3 which is (n-1). The critical value is at 6.25 whereas the test value is 12. Reject null hypothesis

Solution1: for problem 6: Blood type

```
> #11.1-6 : Problem Statement : Blood Types
> #Set significance level
> aplha <- 0.10
> #Create a vector of the values
> observed <- c(12,8,24,6)
> #Create a vector for the Probability
> p <- c(0.20,0.28,0.36,0.16)
> critical_value<-qchisq(p=0.14, df=3, lower.tail=FALSE)
> result<- chisq.test(x= observed, p=p)
> result

        Chi-squared test for given probabilities

data:  observed
X-squared = 5.4714, df = 3, p-value = 0.1404

> ifelse(result$statistic<critical_value, "Null hypothesis rejected", "Null hypothesis accepted")
                X-squared
"Null hypothesis rejected"
```

Task 2: Problem for 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	9.0
Other (because of weather and other conditions)	12.0

Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of weather, 10 because of a National Aviation System delay, and the rest because of arriving late. At $\alpha = 0.05$, do these results differ from the government's statistics?

Source: Transtats: OST_R|BTS

Solution:

The chi square test was performed on airline performance using a random sample size of 200 flights, as shown above. According to the chi-square test, the hypothesis is not significant because the alpha should be 0.05, but we get a p-value of 0.21, which is more than the significance criterion. The outcome of this test is significant, and it is more likely to reject the null hypothesis. While the sample size is 4, the degree of freedom is 3 (n-1). The critical value is at 0.351 where as the test value is 12.

Solution 2: for the problem performance of Airlines

```
> #11.1-7Problem Statement 2: On Time Performance by Airlines
> #Set significance level
> aplha <- 0.05
> #Create a vector of the values
> observed <- c(125,40,10,25)
> #Create a vector for the Probability
> p <- c(.708,.082,.09,.12)
> critical_value<-qchisq(p=1.357e-08, df=3, lower.tail=FALSE)
> result<- chisq.test(x= observed, p=p)
> result

        Chi-squared test for given probabilities

data:  observed
X-squared = 39.504, df = 3, p-value = 1.357e-08

> critical_value<-qchisq(p=1.357e-08, df=3, lower.tail=FALSE)
> ifelse(result$statistic<critical_value, "Null hypothesis rejected", "Null hypothesis accepted")
                X-squared
"Null hypothesis rejected"
```

Task 3: Problem for Ethnicity and Movie Admission

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

Source: MPAA Study

Solution: The chi square test was used on ethnicity and movie admission for two years, 2013 and 2014, as shown above, using the independence test for two variables. The hypothesis is significant according to the chi-square test since the alpha is 0.05, but the p-value is 0.00000000068, which is less than the significance criteria. The result of this test is significant, and the null hypothesis is more likely to be accepted. The degree of freedom is 3, despite the fact that the sample size is 4 (n-1). The critical value is at 0.0.351 where as the test value is 60.15.

Solution3: for the problem Ethnicity and Movie Admission

```
> #Problem od Ethnicity and Movie Admissions
> #2014 Vector
> #Set significance level
> aplha <- 0.05
> #Create a vector of the values
> observed <- matrix(c(724,370,335,292,174,152,107,140), nrow= 2)
> observed
     [,1] [,2] [,3] [,4]
[1,]  724  335  174  107
[2,]  370  292  152  140
> row.names(observed)<- c("2013", "2014")
> colnames(observed)<- c("Caucasian","Hispanic","African Americal","Other")
> result<- chisq.test(x= observed)
> result

        Pearson's Chi-squared test

data:  observed
X-squared = 60.144, df = 3, p-value = 5.478e-13

> critical_value<-qchisq(p=5.478e-13, df=3, lower.tail=FALSE)
> ifelse(result$statistic>critical_value, "Movie depends on Ethnicity", "Movie does'nt depend on ethnicity")
                X-squared
"Movie depends on Ethnicity"
```

Task 4: Problem on 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

Source: New York Times Almanac

Solution: The chi square test was used on Women in the Military for officers on different position Army, Navy, Marine Corps, and Air Force, as shown above, between the rank and the branch o the armed force. The hypothesis is significant according to the chi-square test since the alpha is 0.05, but the p-value is 2.2e-16, which is less than the significance criteria. The result of this test is significant, and the null hypothesis is more likely to be accepted. The degree of freedom is 3, despite the fact that the sample size is 4 (n-1). The critical value is at 0.05 where as the test value is 60.15.

Solution10: for the problem Women in the Military

```
> #Women in the Military
> observed <- matrix(c(10791,7816,932,11819))
> r1<-c(10791,62419)
> r2<-c(7816,42750)
> r3<-c(932,9525)
> r4<-c(11819,54344)
> rows=4
> matrix<- matrix(c(r1,r2,r3,r4),nrow=rows,byrow=TRUE)
> row.names(matrix)<-c("Army","Navy","Marine Corps","Air Force")
> colnames(matrix)<-c("Officers","Enlisted")
> matrix
             Officers Enlisted
Army            10791    62419
Navy             7816    42750
Marine Corps      932     9525
Air Force       11819    54344
> result<-chisq.test(matrix)
> result

        Pearson's Chi-squared test

data:  matrix
X-squared = 652.57, df = 3, p-value < 2.2e-16

> critical_value<-qchisq(p=2.2e-16, df=3, lower.tail=FALSE)
> ifelse(result$statistic>critical_value, "Relation-Rank and Branch", "No Relation-Rank and Branch")
                X-squared
"Relation-Rank and Branch"
```

Task 5: Problem for sodium which include Condiments, cereals, and desserts

The amount of 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

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter

Solution:

```
             Df Sum Sq Mean Sq F value Pr(>F)
food          2  27544   13772   2.399  0.118
Residuals    19 109093    5742
> df1<-summary[[1]][1,"Df"]
> df1
[1] 2
> df2<-summary[[1]][2,"Df"]
> df2
[1] 19
> critical<- qf(p=0.05, df1, df2, lower.tail=TRUE)
> F_test<-summary[[1]][[1,"F value"]]
> ifelse(F_test>critical, "Difference in Mean", "Same Mean")
[1] "Difference in Mean"
```

In this task critical value is 0.0514 whereas the test value is 2.398, the is no significant difference in the mean. The null hypothesis rejected for different food.

Task 6: Problem 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	

Source: Information Resources, Inc.

Solution:In this tasl the critical value is 0.01 whereas the test value is 2.171, which show there is a significant difference in mean. Then we implemented the tukey test.

```
> summary
            Df Sum Sq Mean Sq F value Pr(>F)
prod         2 103770   51885   2.172   0.16
Residuals   11 262795   23890
> df1<-summary[[1]][1,"Df"]
> df1
[1] 2
> df2<-summary[[1]][2,"Df"]
> df2
[1] 11
> critical<- qf(p=0.01, df1, df2, lower.tail=TRUE)
> F_test<-summary[[1]][[1,"F value"]]
> ifelse(F_test>critical, "Difference in Mean", "Same Mean")
[1] "Difference in Mean"
```

#Performed the Tukey Test

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sales ~ prod, data = Sales)

$prod
                       diff        lwr       upr     p adj
ChocoCandy-Cereal -164.80 -428.82409  99.22409 0.2535458
Coffee-Cereal       29.65 -250.38983 309.68983 0.9561014
Coffee-ChocoCandy  194.45  -85.58983 474.48983 0.1916553
```

Task 7: Problem for 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		

Source: New York Times Almanac

Solution: In this task, the critical value is 0.051, whereas the test value is 0.648. which show the is a there is a significant difference in mean. The lowest p-value is 0.52 in coffee which has eastern and western.

```
> summary(anova)
            Df  Sum Sq Mean Sq F value Pr(>F)
country      2 1244588  622294   0.649  0.543
Residuals   10 9591145  959114
> summary<-summary(anova)
> summary
            Df  Sum Sq Mean Sq F value Pr(>F)
country      2 1244588  622294   0.649  0.543
Residuals   10 9591145  959114
> df1<-summary[[1]][1,"Df"]
> df1
[1] 2
> df2<-summary[[1]][2,"Df"]
> df2
[1] 10
> critical<- qf(p=0.01, df1, df2, lower.tail=TRUE)
> df2
[1] 10
> critical<- qf(p=0.05, df1, df2, lower.tail=TRUE)
> F_test<-summary[[1]][[1,"F value"]]
> ifelse(F_test>critical, "Difference in Mean", "Same Mean")
[1] "Difference in Mean"
```

Task 8: Problem of 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 $\alpha = 0.05$.

Solution:

```
> summary(anova)
                 Df Sum Sq Mean Sq F value  Pr(>F)
PF1_light         1  1.920   1.920   3.681 0.09133 .
PF2_food          1 12.813  12.813  24.562 0.00111 **
PF1_light:PF2_food  1  0.750   0.750   1.438 0.26482
Residuals         8  4.173   0.522
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary<-summary(anova)
> df1<-summary[[1]][1,"Df"]
> df1
[1] 1
> df2<-summary[[1]][2,"Df"]
> df2
[1] 1
> critical<- qf(p=0.05, df1, df2, lower.tail=TRUE)
> F_test<-summary[[1]][[1,"F value"]]
> ifelse(F_test>critical, "Difference in light and food", "No diffrence in light and food")
[1] "Difference in light and food"
```

In this task, the critical value is -0.006 whereas the test value is 3.680 which show there is a difference growth of the plant which is affected by light and food.

#Baseball Dataset

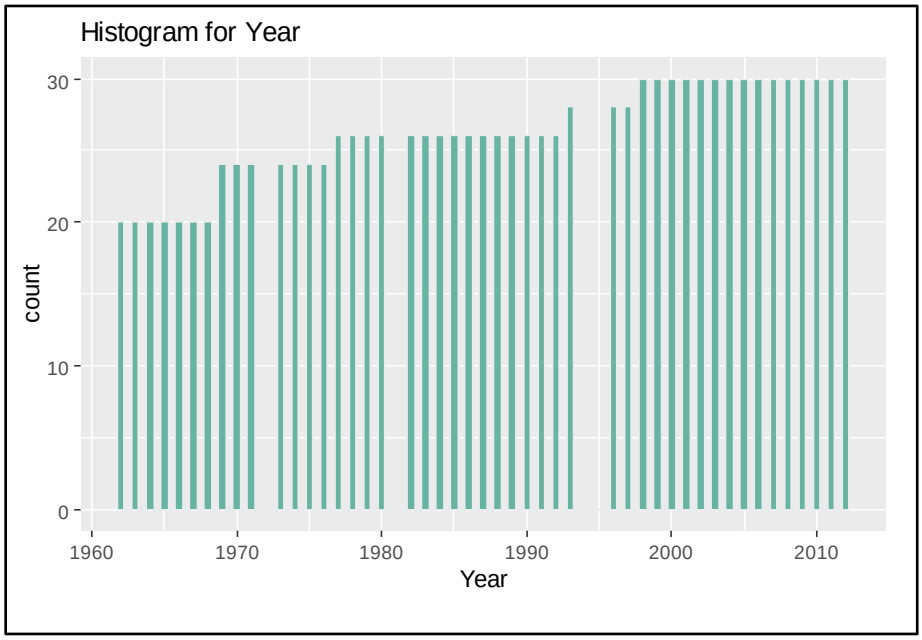
Task 1: In this task imported the file

	Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
1	ARI	NL	2012	734	688	81	0.328	0.418	0.259	0	NA	NA	162	0.317	0.415
2	ATL	NL	2012	700	600	94	0.320	0.389	0.247	1	4	5	162	0.306	0.378
3	BAL	AL	2012	712	705	93	0.311	0.417	0.247	1	5	4	162	0.315	0.403
4	BOS	AL	2012	734	806	69	0.315	0.415	0.260	0	NA	NA	162	0.331	0.428
5	CHC	NL	2012	613	759	61	0.302	0.378	0.240	0	NA	NA	162	0.335	0.424

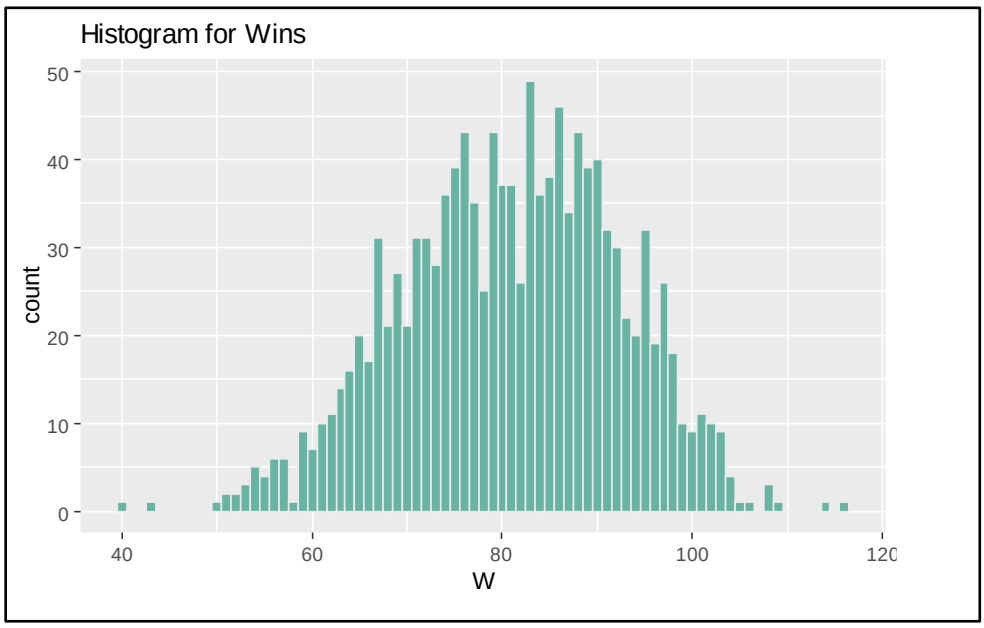
Task2: EDA Explanatory data analysis for the baseball dataset

```
             vars    n    mean    sd     min     max  range   se
Team*           1 1232   18.93 10.61    1.00   39.00  38.00 0.30
League*         2 1232    1.50  0.50    1.00    2.00   1.00 0.01
Year            3 1232 1988.96 14.82 1962.00 2012.00  50.00 0.42
RS              4 1232  715.08 91.53  463.00 1009.00 546.00 2.61
RA              5 1232  715.08 93.08  472.00 1103.00 631.00 2.65
W               6 1232   80.90 11.46   40.00  116.00  76.00 0.33
OBP             7 1232    0.33  0.02    0.28    0.37   0.10 0.00
SLG             8 1232    0.40  0.03    0.30    0.49   0.19 0.00
BA              9 1232    0.26  0.01    0.21    0.29   0.08 0.00
Playoffs       10 1232    0.20  0.40    0.00    1.00   1.00 0.01
RankSeason     11  244    3.12  1.74    1.00    8.00   7.00 0.11
RankPlayoffs   12  244    2.72  1.10    1.00    5.00   4.00 0.07
G              13 1232  161.92  0.62  158.00  165.00   7.00 0.02
OOBP           14  420    0.33  0.02    0.29    0.38   0.09 0.00
OSLG           15  420    0.42  0.03    0.35    0.50   0.15 0.00
```

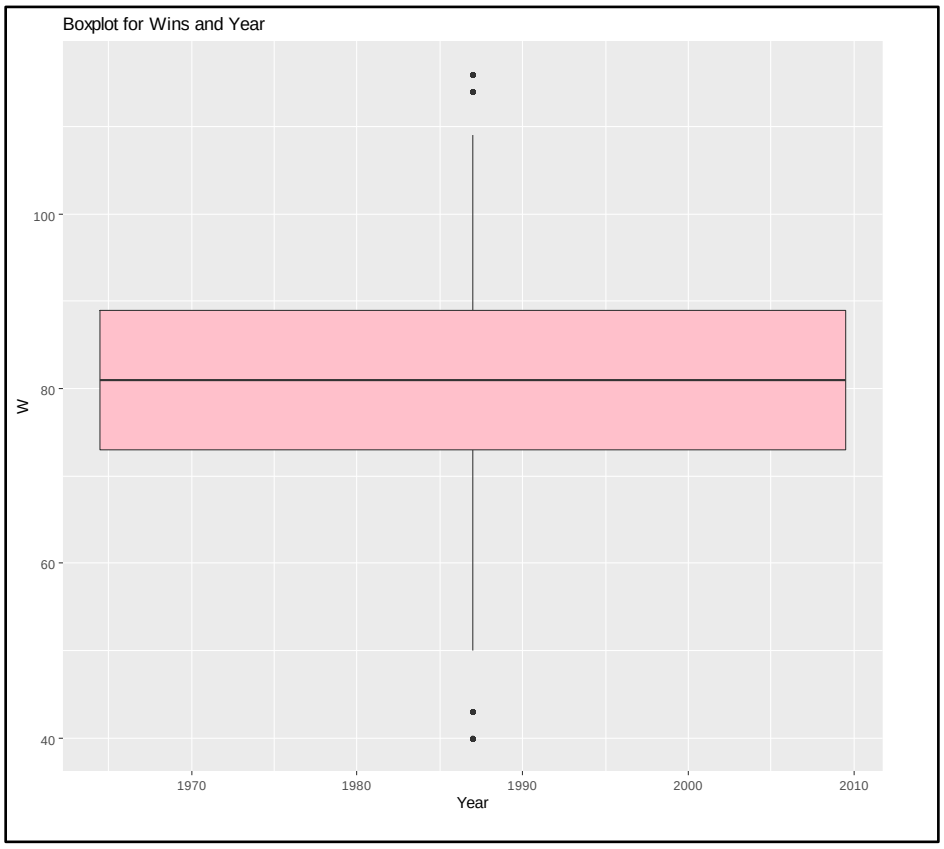
Task3: In this task created a histogram by using ggplot to analyze the pattern in the dataset for the year variable



Task4: In this task created a histogram by using ggplot to analyze the pattern in the dataset for the Wins from 50 to 110 majority of wins occurs.



Task 4: In this task created a boxplot to check the outliers of the variables in the dataset with respect to wins in year outlier below 50 and above 110 which we can clearly see in histogram.



Task5: In this task implementing chi square test in the baseball dataset

In this data set the critical value and the test value is 1.145 and 1558.50 respectively, which proves the null hypothesis has been rejected.

```
> # Create a wins table by summing the wins by decade
> df_win<-aggregate(df_bb$W,by = list(Decade = df_bb$Decade),FUN = sum) %>%
+   as_tibble()
> critical<-qchisq(p=0.05,5,lower.tail = T)
> df_chitest<-chisq.test(df_win)
> df_chitest

        Pearson's Chi-squared test

data:  df_win
X-squared = 1558.5, df = 5, p-value < 2.2e-16

> ifelse(df_chitest$statistic>critical_value_baseball,"Reject Null Hypothesis","Reject alternative hypothesis")
             X-squared
"Reject Null Hypothesis"
```

#Crop Dataset

Task 1: Import the datset and implemented the anova test in r.

In this dataset analysis of variance divide the observed value into different component like in this fertilizer, density which give the hypothesis is rejected based on the interaction between the yield and with the fertilizer, so the Null Hypothesis rejected. Whereas the critical value and the test value is 0.006 and 17.7 respectively

```
> summary(anova)
                  Df Sum Sq Mean Sq F value   Pr(>F)
Fertilizer         1  5.743   5.743  17.078  7.9e-05 ***
Density            1  5.122   5.122  15.230 0.000181 ***
Fertilizer:Density  1  0.150   0.150   0.447 0.505630
Residuals          92 30.939   0.336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary_cd<-summary(anova)
> alpha_crop<-0.05
> df1<-summary_cd[[1]][1,"Df"]
> df2<-summary_cd[[1]][2,"Df"]
> critical<- qf(p=0.05, df1, df2, lower.tail=TRUE)
> F_test<-summary_cd[[1]][[1,"F value"]]
> ifelse(F_test>critical,"Null Hypothesis rejected","Null Hypothesis approved")
[1] "Null Hypothesis rejected"
```

Conclusion and Interpretation

We have used different chi-square to solve problems such as evaluating a distribution for goodness of fit, testing two variables for independence, and testing proportions for homogeneity. Aside from the chi square test, one-way ANOVA and two-way ANOVA techniques are used to assess the significant difference between variables.

Implemented in problem one the chi square test was performed on a random sample of 50 patients for blood type, as indicated above. According to the chi-square test, the hypothesis is not significant because the alpha should be 0.10, but we get a p-value of 0.1404, which is more than the significance criterion. The outcome of this test is not significant, and it is more likely to reject the null hypothesis, which states that two variables are independent of each other. While the sample size is 4, the degree of freedom is 3 (n-1). As demonstrated above, the chi square test was done on airline performance using a random sample size of 200 flights. The hypothesis is not significant, according to the chi-square test, because the alpha should be 0.05, but we receive a p-value of 0.21, which is more than the significance criteria. The result of this test is significant, and the null hypothesis is more likely to be rejected. The degree of freedom is 3, even though the sample size is 4 (n-1). For two years, 2013 and 2014, the chi square test was applied on ethnicity and movie admission, as indicated above, using the independence test for two variables. According to the chi-square test, the hypothesis is significant since the alpha is 0.05, but the p-value is 0.00000000068, which is less than the significance requirement. This test yielded a significant result, indicating that the null hypothesis is more likely to be accepted. Even though the sample size is 4, the degree of freedom is 3 (n-1).

In conclusion, Chi-Square test results are more proficient and provide the best understating about the dataset. Details distribution of the data and the detailed information about the data can be derived from by implementing chi-square test. Moreover, it is used to where the parametric value can not be matched. In understanding the group and handling the multiple tasks to get the better variance of the independent variables among the distribution on the data. On the other hand, ANOVA doesn’t provide a better performance than the chi-square another variance test however it’s a good tool to provide the understanding about the null hypothesis

References:

[0] Bluman book McGrew Hills

[1] ggplot2 histogram plot : Quick start guide - R software and data visualization

http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

[2] Histograms and frequency polygons - geom_freqpoly
https://ggplot2.tidyverse.org/reference/geom_histogram.html

[3] Chi-Square Test in R: Explore the Examples and Essential concepts!
https://data-flair.training/blogs/chi-square-test-in-r/

[4] ANOVA in R: A step-by-step guide

Bevans
https://www.scribbr.com/statistics/anova-in-r/

[5] How to Perform Tukey HSD Test in R: R-bloggers

Finnstats
https://www.r-bloggers.com/2021/08/how-to-perform-tukey-hsd-test-in-r/