Northeastern University
College of Professional Studies

ALY6015- Chi-Square testing and ANOVA

Prof. Ji-Young Yun

Submitted by: -

Jeseeka Shah

Introduction

In this assignment I have used Chi-square and ANOVA test to solve the various stated problems. Chi-Square test is used when the values at test are categorical in nature. ANOVA test is used is when at least one categorical or one continuous dependent variable is present. Also, when there is more than one categorical variable, and they are independent in nature I have considered using Chi-square test of independence. Mainly, this assignment has challenged me in understanding the different test and applying this knowledge in solving the problem statement that is been given.

Analysis on Problem statements

The problems are solved based on the test required as per the questionnaire. Firstly, stating a claim and making a hypothesis then finding the critical value for the claim. Moving forward carrying out the test based on the hypothesis and critical value obtained. Lastly, based on the outcome a decision has been made.

- ### Blood Type :-

  The random sample size is 50 patients and blood type count for the sample is shown below and total population percentage is also shown below :-

Blood type	Count per blood type
A	**12**
B	**8**
O	**24**
AB	**6**

Table 1: Sample based count of blood type

Blood type	Population percentage of Blood Type
A	**20**
B	**28**
O	**36**
AB	**16**

Table 1: Sample based count of blood type

Alpha =0.10 which means confidence interval is going to be of 90 percent.

Ho : Patients of large hospital have same blood type distribution as that of general hospitals.
H1 : Patients of larger hospitals do not have same blood type distribution as that of general hospital.

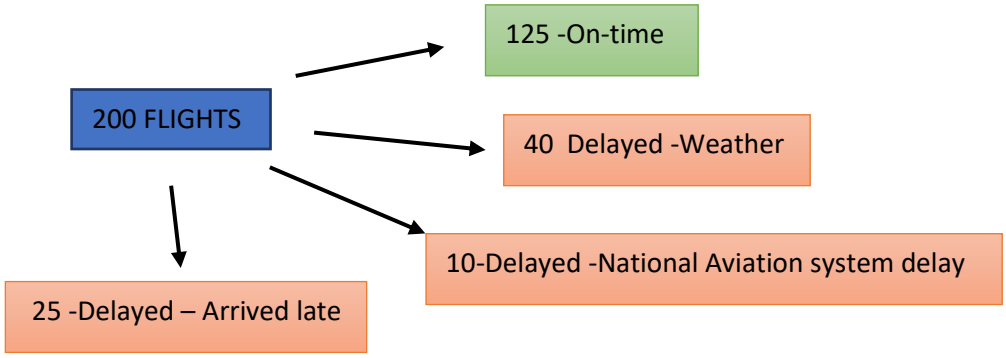
```
> test_result

        Chi-squared test for given probabilities

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

> test_result$statistic
X-squared
 5.471429
> test_result$parameter
df
 3
> test_result$p.value
[1] 0.1403575
```

Figure 1: Result of hospital data

Here, the p-value is greater than 0.10 that means it fails to reject the null hypothesis as there not enough evidence to support the claim. That means the patients of the large hospital have the same blood type distribution as that of general hospital. Also, the chi-square calculated is 5.4714 which is less than the critical value of chi-square that is check in the table with respect to the degree of freedom (3) as 6.251.

- **Performance estimation of major airlines :-**

  There are 200 randomly selected flights that are showed with following distribution of being on time or delayed **:-**



Action	Percentage of Time
On time	70.8
National Aviation System Delayed	8.2
Aircraft arriving late	9.0
Other	12.0

Table 3 : Airline performance based on being on-time

Alpha value is -0.05 and confidence interval is 95 percent.

Ho : On-time performance recorded by Government is same as Bureau of transportation statistics **H1** : On-time performance recorded by Government is not same as Bureau of transportation statistics.

```
> test_result2<-chisq.test(x = Diff_count, p = Ontime_prob)
> test_result2

        Chi-squared test for given probabilities

data:  Diff_count
X-squared = 17.674, df = 3, p-value = 0.0005134
```

Figure 2: showing the result of chi-square test

The p-value is smaller than 0.05.Therefore, I have rejected the null hypothesis. And the claim that the on-time performance statistics observed by the government is not same as the Bureau of transportation statistics.

➢ Evaluating the audience at movies based on the ethnicity :-

Admission to movie based on the ethnicity for different years at the level of significance of 0.05 . The table below shows the admission number based on ethnicity.

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

Figure 3: shows the admission number based on ethnicity

```
> test_result3<-chisq.test(movie_data)
> test_result3

        Pearson's Chi-squared test

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

Figure 4: Chisq-test result

Ho : Admission to movie based on the ethnicity of the audience. **H1** : Admission to movie is not based on the ethnicity of audience.

The p-value is less than the 0.05 . That means we reject the null hypothesis that admission is based on ethnicity.as the chi-square statistic is 60.144 which is more than the critical value which is 7.81. Hence, the null hypothesis is rejected that states that admission to movies is not based on the ethnicity.

➢ Relationship between rank and branch of Armed Forces of Women in Military:-

To check if there is sufficient information to establish a relationship between rank and branch of the armed forces.

	Officers	Enlisted
Army	10791	62491
Navy	7816	42750
Marine Corps	932	9525
Air Force	11819	54344

Figure 5: Shows the women from military enlisted and officers

Ho : There is a relation between the rank and the branch of the armed forces. **H1** : There is no relation between the rank and the branch of the armed forces.

```
> test_result4

        Pearson's Chi-squared test

data:  military_women
X-squared = 654.27, df = 3, p-value < 2.2e-16
```

Figure 6: Chi-square statistics

The p values is less than alpha values which is given as 0.05. Which shows that **we reject the null** hypothesis as there is not enough evidence to prove it Even the chi-square value calculated is 654.27 is **larger than the critical value of chi-square present** in the table based on the degree of freedom and p-value. Therefore, there is no relationship between the rank and the branch of the armed forces.

➢ Presence of Sodium :-

Random sample of three different type are given such as cereals, condiments, and dessert. Based on the hypothesis, I have used ANOVA here as we have to find the mean difference between the sodium levels in the three categories.

Ho :There is no difference in the mean sodium amount of condiments, cereals and dessert. **H1** : There is difference in the mean sodium amount of condiments, cereals, and dessert.

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

Figure7 : Summary of ANOVA test for presence of sodium

The p values here is greater than the given significance value of 0.05. Therefore, we fail to reject the null hypothesis that there is no difference in mean sodium amount of condiment, cereals, and dessert due to lack of enough evidence.

```
> F.value<-a.summary[[1]][1, "F value"]
> F.value
[1] 2.398538
> p.value<-a.summary[[1]][1, "Pr(>F)"]
> p.value
[1] 0.1178108
> ifelse(p.value > alpha, "Fail to reject null hypothesis", "Reject the null hypothesis")
[1] "Fail to reject null hypothesis"
```

Figure 8 : Values of F, P comparison based on p values result

```
> TukeyHSD(anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sodium ~ food, data = sodium_contains)

$food
                          diff        lwr       upr     p adj
condiments-cereals  -80.000000 -182.89588  22.89588 0.1456674
desserts-cereals     -8.214286 -107.84279  91.41422 0.9761344
desserts-condiments  71.785714  -27.84279 171.41422 0.1866850
```

➢ Sales for Leading Companies :-

The sales of a company in million dollars are given for different leading company with alpha value as 0.01 as shown below **:-**

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

Figure 9: Sales for different company

Ho :There no significant difference between the means of the sales for the different companies

H1 : There is a significant difference between the means of the sales for the different companies.

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

Figure 10: ANOVA test results for means

The p-value is greater than the alpha value of 0.01. Therefore, we fail to reject the null hypothesis as the p values is 0.1603 > 0.01.That means there is no significance difference between the means of the sales for the different companies.

➢ Expenditure per student

The expenditure per student from states in three sections of the country are given and alpha value to be considered is 0.05.

Ho : There is no difference in the mean of the expenditure per student belonging to different states
H1 : There is no difference in the mean of the expenditure per student belonging to different states.

```
> summary(anova)
            Df  Sum Sq Mean Sq F value Pr(>F)
section      2 1244588  622294   0.649  0.543
Residuals   10 9591145  959114
> a.summary<-summary(anova)
> df.numerator<-a.summary[[1]][1, "Df"]
> df.numerator
[1] 2
> df.denominator<-a.summary[[1]][2, "Df"]
> df.denominator
[1] 10
> F.value<-a.summary[[1]][1, "F value"]
> F.value
[1] 0.6488214
> p.value<-a.summary[[1]][1, "Pr(>F)"]
> p.value
[1] 0.5433264
> ifelse(p.value > alpha, "Fail to reject null hypothesis", "Reject the null hypothesis")
[1] "Fail to reject null hypothesis"
> TukeyHSD(anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pupil ~ section, data = pupil)

$section
                           diff       lwr      upr     p adj
middlethird-eastherntthird  428.35 -1372.582 2229.282 0.7954670
westernthird-eastherntthird 740.10 -1060.832 2541.032 0.5203918
westernthird-middlethird    311.75 -1586.599 2210.099 0.8954324
```

Figure 11: ANOVA test results for population

Here, its seen that the p-values are 0.54332 which is greater than the significant value of 0.05 that is alpha. This means we fail to reject the null hypothesis as there is not enough evidence to support the claim. Hence, the mean of expenditure per student belonging to different stated has no difference in the mean.

➢ Growth in Plants

There are two different factors that are considered here one is the growth light and other is the plant food supplement with different minerals. The interaction between these two factors is to be checked. So, I am going to consider the ANOVA test to check the relationship. The alpha given to us is 0.05.

	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

Figure 12: Grow-light relation with Plant Food A& B

Ho : There is no change in the growth with respect to the supplement or light **H1** : There is change in the growth with respect to the supplement or light.

```
> # extract the p test value from summary
> ipgpvalue1 <- a.summary3[[1]][[1,"Pr(>F)"]]
> ipgpvalue1
[1] 0.2954224
> ipgpvalue2 <- a.summary4[[1]][[1,"Pr(>F)"]]
> ipgpvalue2
[1] 0.1776955
> # Comparing p value with alpha to make decision
> ifelse(ipgpvalue1 > 0.05,"Fail to reject null hypothesis","reject null hypothesis")
[1] "Fail to reject null hypothesis"
> ifelse(ipgpvalue2 > 0.05,"Fail to reject null hypothesis","reject null hypothesis")
[1] "Fail to reject null hypothesis"
```

Figure 13: P-value for light and plant food A& B

As shown in the figure 13, the p-value is greater than that of the alpha value in both growth-light and food supplement . This mean that we fail to reject the null hypothesis as there is not enough evidence to support the claim. Therefore, there is no change in the growth with respect to the supplement or light.

➢ Own your own problem : Baseball Team

Here, the dataset is given to us. So , I read the csv file in R . Performed the descriptive statistics and then plotted the graphs as shown in the figure 14 and 15. The figure 14 and figure 15 is shows the density plot and boxplot for runs allowed and scored. The density of both shows a normal curve but boxplot of runs scored has more outliers than that of runs allowed.

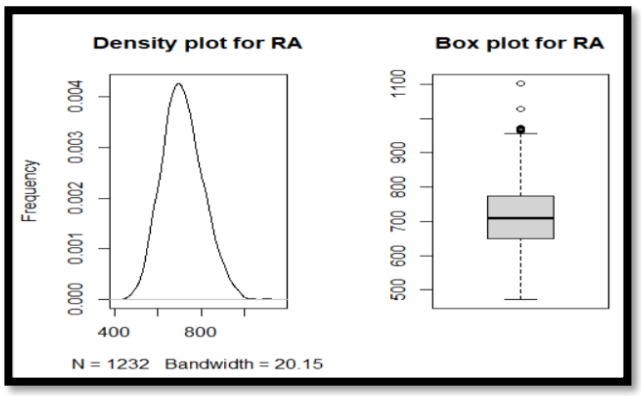


Figure 14: showing the runs allowed

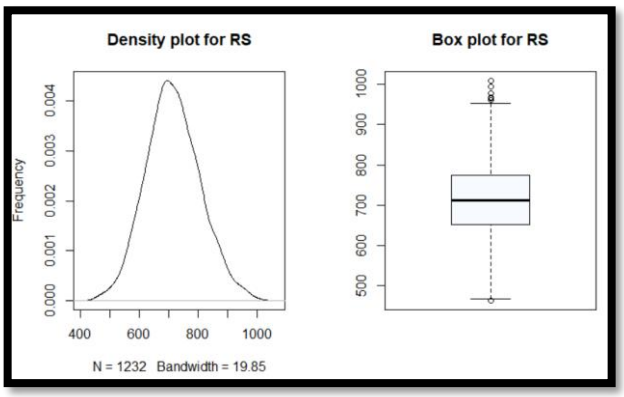


Figure 15 : Showing the runs scored

Ho : There is no difference in the wins by decade **H1** :There is a difference in the wins by the decade.

```
> wins
# A tibble: 6 x 2
  Decade  wins
   <dbl> <int>
1   1960 13267
2   1970 17934
3   1980 18926
4   1990 17972
5   2000 24286
6   2010  7289
> Values_obsv <- c(13267, 17934, 18926, 17972, 24286, 7289)
> Values_expt<- c(1/6,1/6,1/6,1/6,1/6,1/6)
> qf(p=0.05, df1 = 5, df2 = 1, lower.tail = FALSE)
[1] 230.1619
```

```
> chisq.test(wins)

        Pearson's Chi-squared test

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

Figure 16 : Showing the observations of the baseball

The p-value is less than 0.05 and the Critical value from Chi-Squared table is 11.07 which is less than test statistic 1558.5. This means that we reject the null hypothesis. Hence, there is a difference in the wins by the decade.

Conclusion :-

Initially , the scenarios such as **women in military, performance of airlines, and audience at the movies based on the ethnicity** all are having p-values which are less than the alpha values and chi-square values which are greater than critical value stated in the table. Therefore, we reject the null hypothesis and the accept the claims or the alternative hypothesis in these examples. In case of checking the **blood type**, **presence of sodium in three different food types, expenditure per student** for different state , and the **mean difference in the sales for different companies** have the result which has **p values** greater than the alpha which fails to reject the null hypothesis as there is not enough evidence to support the claim. The **growth in plants** scenario the result after the **ANOVA test** gave that the p- value was greater than the alpha therefore we **fail to reject** the null hypothesis which states that there **is no change in the growth with respect to supplement or light**. In baseball match after performing the **chi-square** test gave the values which were more than chi-square table and **p values which was less than 0.05**. Therefore, reject the null hypothesis which means that there is difference in the wins by the decade.

References:

- How Analysis of Variance (ANOVA) Works. (2021, October 6). Investopedia. Retrieved 25th July 2021, from https://www.investopedia.com/terms/a/anova.asp.

- One-way ANOVA - An Introduction To When You Should Run This Test And the Test Hypothesis | Laerd Statistics. (n.d.). One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics. Retrieved 25h July 2021 from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php.

- Z., & Posts By Zach, V. A. (2021, August 25). Chi-Square Test Vs. ANOVA: What's the Difference? - Statology. Statology. Retrieved 25th July 2021 https://www.statology.org/chi-square-vs-anova/#:~:text=As%20a%20basic%20rule%20of,and%20one%20continuous%20dependent%20variable..