Module 2 | Assignment: Chi-Square Testing and ANOVA

Trang Tran

CPS, Northeastern University

ALY6015 | Intermediate Analytics

Professor Steve Morin

Apr 26, 2023

Section 11-1/ Chi-squared Goodness of Fit - Use the traditional method

6. Blood Types

a. H0: The hospital patients have the same blood type distribution as those in the general population, specifically type A, 20%; type B, 28%; type O, 36%; and type AB = 16% (claim)

H1: The hospital patients' distribution differs from those of the general population.

b. $\alpha = 0.10$; df = 4 - 1 = 3 => CV = 6.251 (Table G or using R)

c. Test Value = $\chi^2 = 5.471$ (using R)

d. Since $5.471 < 6.251$, the decision is not to reject the null hypothesis.

e. There is not enough evidence to reject the claim that hospital patients have the same blood type distribution as those in the general population.

8. On-Time Performance by Airlines

a. H0: The performance statistics of airlines are that 70.8% were on time, 8.2% were delayed by the National Aviation System, 9% were delayed by other aircraft arriving late, and 12% were delayed for various reasons (because of weather and other conditions).

H1: The delay proportions are different from the government's statistics above. (claim)

b. $\alpha = 0.05$; df = 4 - 1 = 3 => CV = 7.815 (Table G or using R)

c. Test Value = $\chi^2 = 17.832$ (using R)

d. Since $17.832 > 7.815$, the decision is to reject the null hypothesis.

e. There is enough evidence to support the claim that the proportions are different from the government's statistics.

Section 11-2/ Chi-squared Independence Test - Use the traditional method

8. Ethnicity and Movie Admissions

a. H0: Movie attendance by year was independent of the ethnicity of the moviegoers.

H1: Movie attendance by year was dependent upon ethnicity. (claim)

b. α = 0.05; df = 4 - 1 = 3 => CV = 7.815 (Table G or using R)

c. Test Value = χ^2 = 60.143 (using R)

d. Since 60.143 > 7.815, the decision is to reject the null hypothesis.

e. There is enough evidence to support the claim that movie attendance by year was dependent upon ethnicity.

10. Women in the Military

a. H0: There is no relationship between rank and branch of women personnel in the military.

H1: There is an existing relationship between rank and branch of women personnel in the military. (claim)

b. α = 0.05; df = 4 - 1 = 3 => CV = 7.815 (Table G or using R)

c. Test Value = χ^2 = 654.272 (using R)

d. Since 60.143 > 7.815, the decision is to reject the null hypothesis.

e. There is sufficient evidence to support the claim that a relationship exists between rank and branch of the Armed Forces of women.

Section 12-1/ One-way ANOVA Test - Use the traditional method

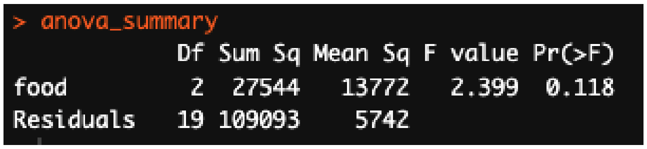
8. Sodium Contents of Foods

a. H0: $\text{mean}_{\text{cereals}} = \text{mean}_{\text{condiments}} = \text{mean}_{\text{desserts}}$

H1: At least one of the means is different than the others. (claim)

b. α = 0.05; dfN = 3 - 1 = 2; dfD = 22 – 3 = 19 => CV = 3.52 (Table H)

c. Test Value = F = 2.399 (using R)



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

d. Since 2.399 < 3.52, the decision is to not reject the null hypothesis.

e. There is not enough evidence to support the claim that at least one mean is different.

Section 12-2/ One-way ANOVA Test/ Scheffé or Tukey test if Ho is rejected

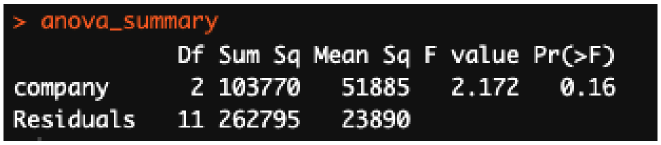
10. Sales for Leading Companies

a. H0: $\text{mean}_{\text{cereal}} = \text{mean}_{\text{chocolate\_candy}} = \text{mean}_{\text{coffee}}$

H1: At least one mean is different from the others. (claim)

b. α = 0.01; dfN = 3 - 1 = 2; dfD = 14 – 3 = 11 => CV = 7.21 (Table H)

c. Test Value = F = 2.172 (using R)



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

d. Since 2.172 < 7.21, the decision is to not reject the null hypothesis.

e. There is not enough evidence to support the claim that at least one mean is different from the others.

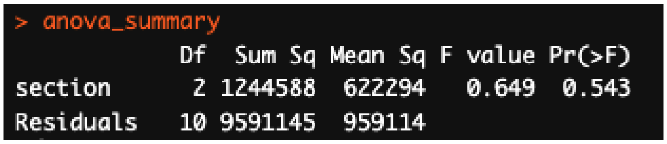
12. Per-Pupil Expenditures

a. H0: $\text{mean}_{\text{Eastern}} = \text{mean}_{\text{Middle}} = \text{mean}_{\text{Western}}$

H1: At least one mean is different from the others. (claim)

b. α = 0.05; dfN = 3 - 1 = 2; dfD = 13 – 3 = 10 => CV = 4.10 (Table H)

c. Test Value = F = 0.649 (using R)



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

d. Since 0.649 < 4.10, the decision is to not reject the null hypothesis.

e. There is not enough evidence to support the claim that at least one of the means is different from the others.

Section 12-3/ Two-way ANOVA Test

10. Increasing Plant Growth

a. *Plant food Hypotheses (F_A)*

H0: There is no difference in the mean growth of plants and the type of plant food supplement.

H1: There is a difference in the mean growth of plants and the type of plant food supplement. (claim)

Grow-light Hypotheses (F_B)

H0: There is no difference in the mean growth of plants and the strength of the grow-light.

H1: There is a difference in the mean growth of plants and the strength of the grow-light. (claim)

Plant food / Grow-light Interaction Hypotheses (F_{AxB})

H0: There is no interaction effect between the type of plant food supplement and the strength of the grow-light on plant growth.

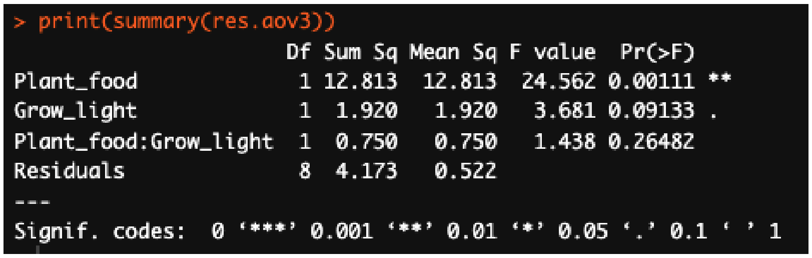
H1: There is an interaction effect between the type of plant food supplement and the strength of the grow-light on plant growth.

b. $\alpha = 0.05$

$dfN_A = dfN_B = 2 - 1 = 1$; $dfN_{AxB} = 1*1 = 1$; $dfD = 2*2*(3-1) = 8$

⇨ $CV_A = CV_B = CV_{AxB} = 5.32$ (Table H)

c. Summary table and compute test values:



```
> print(summary(res.aov3))
                     Df Sum Sq Mean Sq F value  Pr(>F)   
Plant_food            1 12.813  12.813  24.562 0.00111 **
Grow_light            1  1.920   1.920   3.681 0.09133 . 
Plant_food:Grow_light 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
```

Table 1: ANOVA Summary Table

Source	**SS**	**df**	**MS**	**F**
Plant food	12.813	1	12.813	24.562
Grow-light	1.92	1	1.92	3.681
Interaction	0.75	1	0.75	1.438
Within	4.173	8	0.522	
Total	19.656	11		

d. Since the F_A (for the plant food) is 24.562, much greater than the CV_A, 5.32, the decision is to reject the null hypothesis for the plant food. It can be concluded that there is a significant difference in the mean growth for the plant food. Meanwhile, F_B (grow-light) and F_{AxB} are 3.681 and 1.438 respectively, lower than the CV. So, the decision is to not reject the null hypotheses for the Grow-light and Plant food / Grow-light Interaction.

e. Summarize the result:

The analysis results suggest that there is a significant difference in the mean growth for the plant food. Besides, there is not enough evidence to support the claim that plant light strength influences the mean growth of plants. Additionally, there is insufficient evidence to conclude that having an interaction effect between the type of plant food supplement and the strength of the grow-light on plant growth.

Figure 1: Boxplot with multiple groups

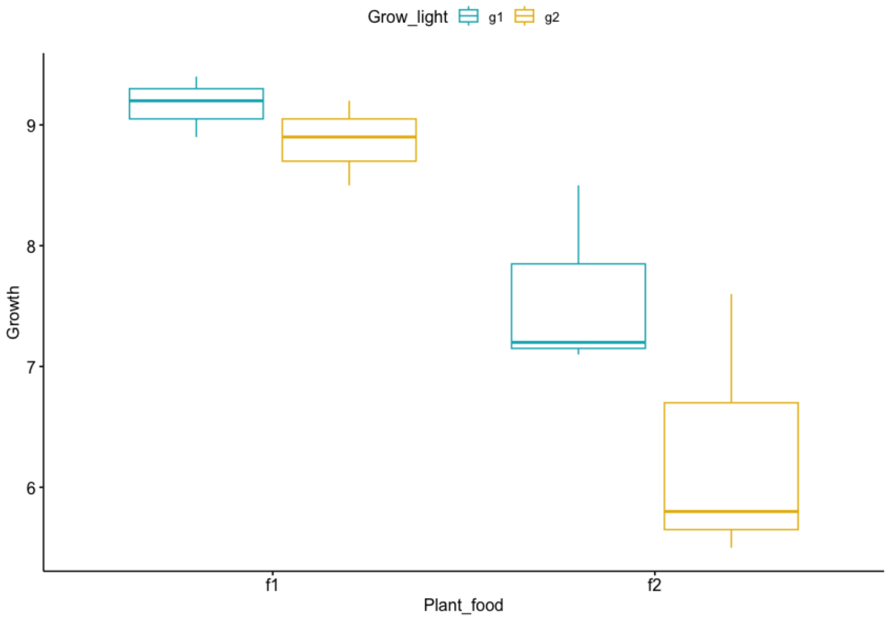
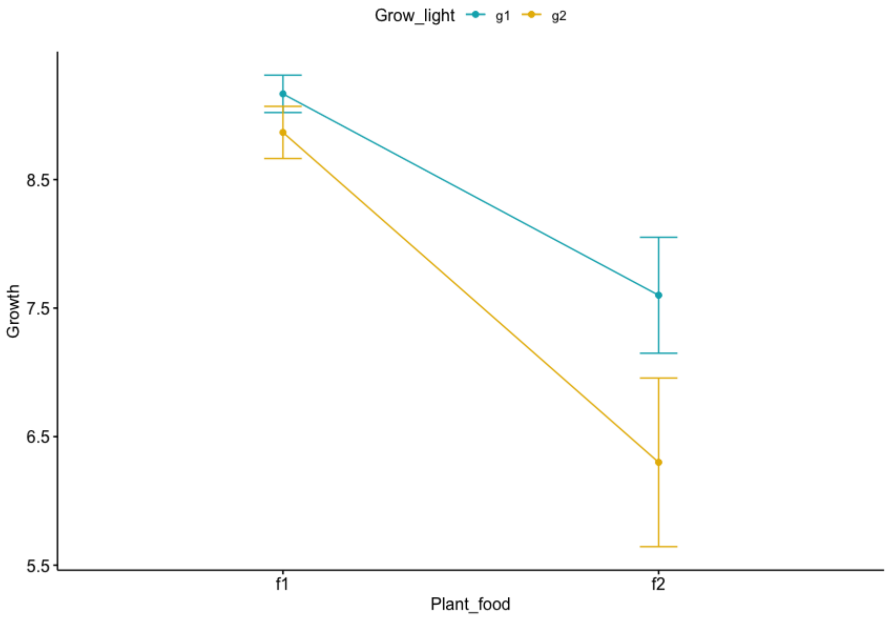


Figure 2: Interaction plot

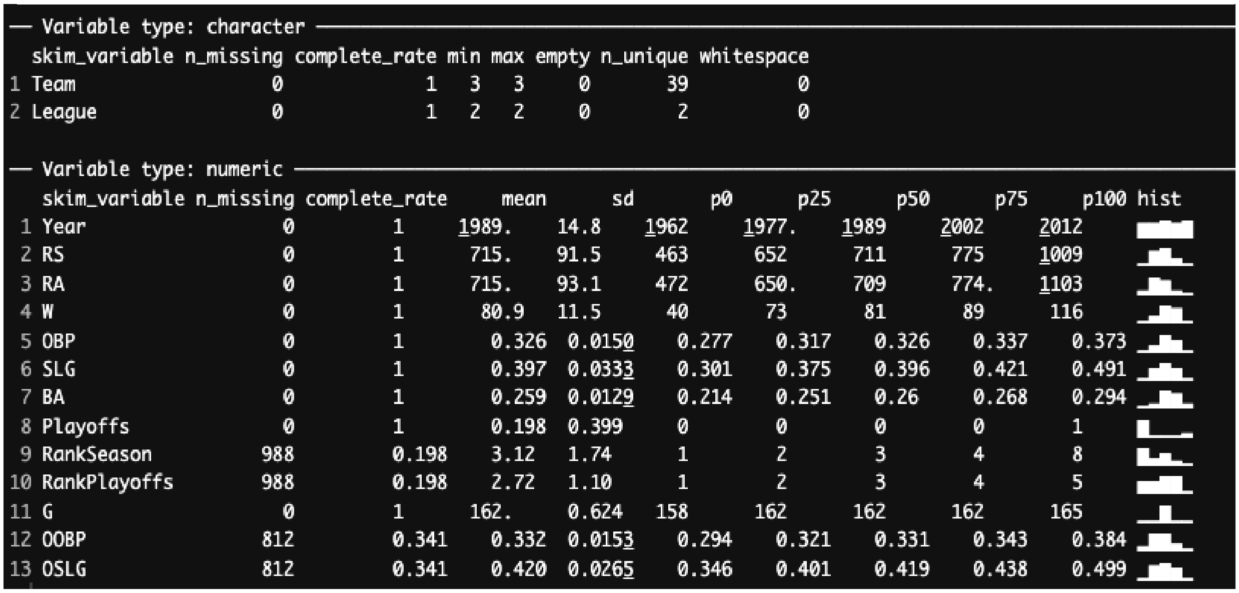


ON MY OWN PART

###Baseball dataset

#EDA

Figure 3: Baseball Dataset Summary



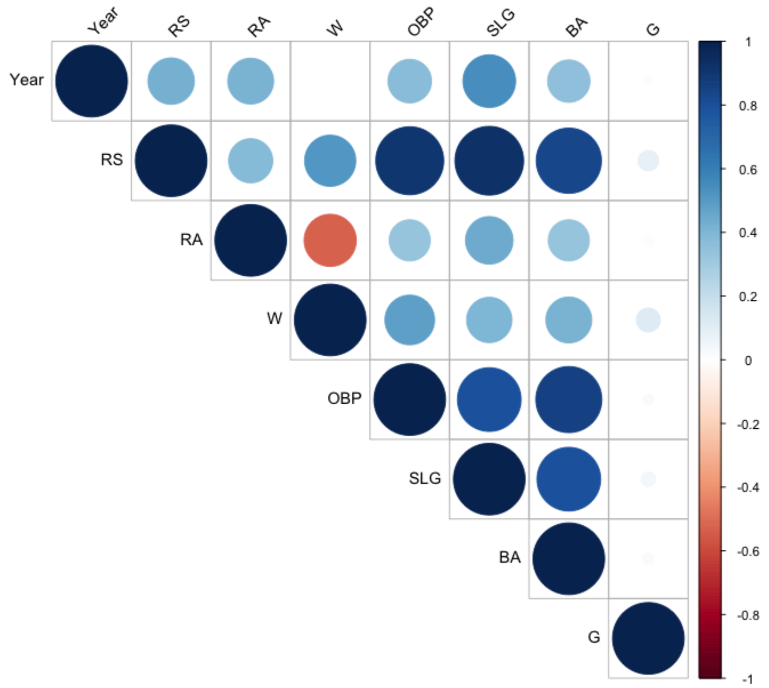
```
── Variable type: character ──────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 Team                  0             1   3   3     0       39          0
2 League                0             1   2   2     0        2          0

── Variable type: numeric ────────────────────────────────────────────────────
   skim_variable n_missing complete_rate     mean      sd       p0      p25      p50      p75     p100 hist
 1 Year                  0         1      1989.    14.8     1962     1977.    1989     2002     2012    ▇▆▇▇▇
 2 RS                    0         1       715.    91.5      463      652      711      775     1009   ▁▇▆▁▁
 3 RA                    0         1       715.    93.1      472      650.     709      774.    1103   ▂▇▃▁▁
 4 W                     0         1        80.9   11.5       40       73       81       89      116   ▁▃▇▆▁
 5 OBP                   0         1         0.326  0.0150     0.277    0.317    0.326    0.337    0.373 ▁▃▇▃▁
 6 SLG                   0         1         0.397  0.0333     0.301    0.375    0.396    0.421    0.491 ▁▅▇▃▁
 7 BA                    0         1         0.259  0.0129     0.214    0.251    0.26     0.268    0.294 ▁▂▇▅▁
 8 Playoffs              0         1         0.198  0.399      0        0        0        0        1     ▇▁▁▁▂
 9 RankSeason          988         0.198     3.12   1.74       1        2        3        4        8     ▇▃▅▁▁
10 RankPlayoffs        988         0.198     2.72   1.10       1        2        3        4        5     ▅▅▆▇▂
11 G                     0         1       162.     0.624    158      162      162      162      165     ▁▁▇▁▁
12 OOBP                812         0.341     0.332  0.0153     0.294    0.321    0.331    0.343    0.384 ▂▇▇▃▁
13 OSLG                812         0.341     0.420  0.0265     0.346    0.401    0.419    0.438    0.499 ▁▅▇▅▁
```

The data set consists of 1232 observations and 15 variables, including team, league, year, and various performance statistics. After importing and performing exploratory data analysis using the [skimr] package, we can see that there are two character variables: 'Team' and 'League'. The 'Team' variable has no missing values and 39 unique teams. The 'League' variable has no missing values and only two unique values. There are also 13 numeric variables with different ranges and standard deviations. The 'Year' variable has no missing values and ranges from 1962 to 2012. The 'RankSeason' and 'RankPlayoffs' variables have 988 missing values which is around 80% of the data set. The ‘OOBP’ and ‘OSLG’ variables also have 812 missing values, accounting for around 65% of the total rows. The remaining variables have no missing values. So, I decided to remove variables “'RankSeason', 'RankPlayoffs', ‘OOBP’ and ‘OSLG’ for their missing data.

Besides, in a baseball context, variables that are measured on a categorical scale (e.g. Playoffs, RankSeason, RankPlayoffs) are considered qualitative data. So, I think we should convert them into factors.

The correlation plot below (Figure 4) shows the relationship between numeric variables in the dataset.

Figure 4: Correlation Plot



There are several trends and interesting findings in this correlation matrix. We see only one negative correlation between RA (runs allowed) and W (wins). This makes sense since teams that allow fewer runs are likely to win more games. There are strong positive correlations between OBP (on-base percentage), SLG (slugging percentage), and BA (batting average) with correlation coefficients ranging from 0.7 to over 0.9. This makes sense since these three variables are all measures of a team's batting performance. There is a strong positive correlation between RS and SLG, with a correlation coefficient of 0.91. This makes sense since the slugging percentage is a measure of a team's ability to hit for extra bases, which typically leads to more runs scored.

#Perform a Chi-Square Goodness-of-Fit test

a. H0: There is no significant difference in the number of wins by decade.

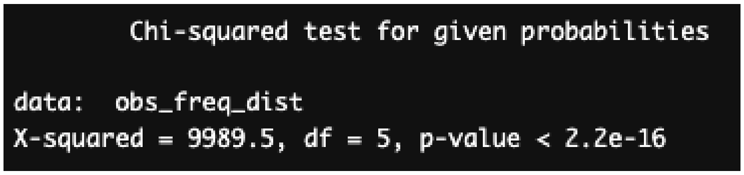
H1: There is a difference in the number of wins by decade. (claim)

b. α = 0.05; df = 6 – 1 = 5 => CV = 11.071 (Table G)

c. Test Value = χ^2 = 9989.54 (using R)

d. Since 9989.54 is much higher than 11.071, the decision is to reject the null hypothesis.

e. If we use the p-value method in R, we have the same result when comparing the p-value (<2.2e – 16) with a significant level (0.05). The p-value is much lower than the α so we have enough evidence to reject the null hypothesis and can conclude that there is a significant difference in the number of wins by decade.



```
	Chi-squared test for given probabilities

data:  obs_freq_dist
X-squared = 9989.5, df = 5, p-value < 2.2e-16
```

###Crop Dataset

#Perform a Two-way ANOVA test

a. *Density Hypotheses (F_{A})*

H0: There is no difference in the mean yield by the type of density.

H1: There is a difference in the mean yield by the type of density. (claim)

Fertilizer Hypotheses (F_{B})

H0: There is no difference in the mean yield by the type of fertilizer.

H1: There is a difference in the mean yield by the type of fertilizer. (claim)

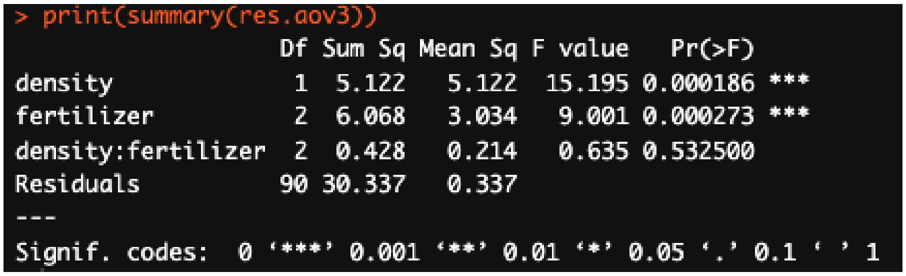
Density / Fertilizer Interaction Hypotheses (F_{AxB})

H0: There is no interaction effect between the type of density and the type of fertilizer on crop yield.

H1: There is an interaction effect between the type of density and the type of fertilizer on crop yield. (claim)

b. α = 0.05

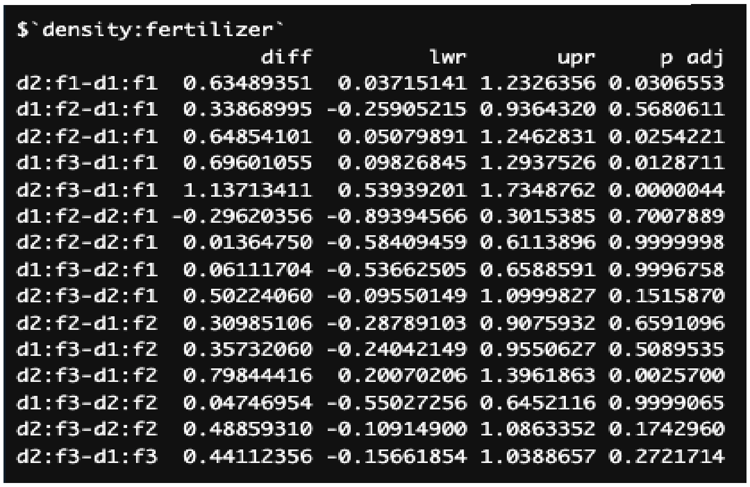
dfN$_A$ = 2 - 1 = 1; dfN$_B$ = 3 - 1 = 2; dfN$_{AxB}$ = 1*2 = 2; dfD = 2*3*(16-1) = 90



```
> print(summary(res.aov3))
                    Df Sum Sq Mean Sq F value   Pr(>F)    
density              1  5.122   5.122  15.195 0.000186 ***
fertilizer           2  6.068   3.034   9.001 0.000273 ***
density:fertilizer   2  0.428   0.214   0.635 0.532500    
Residuals           90 30.337   0.337                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Since the P-value for the density and the P-value for the fertilizer are 0.000186 and 0.000273 respectively, much lower than the significant level (.05), the decision is to reject the null hypotheses for the density and fertilizer. It can be concluded that there is a significant impact on the yield by the density and fertilizer. Meanwhile, the p-value for the interaction between density and fertilizer is 0.532, higher than the significant level (.05), so the decision is to not reject the null hypotheses for the density type/ fertilizer type interaction. We can conclude that there is an interaction effect between the type of density and the type of fertilizer on crop yield.

Figure 5: Tukey Test for density/fertilizer interaction



```
$`density:fertilizer`
                 diff         lwr       upr     p adj
d2:f1-d1:f1  0.63489351  0.03715141 1.2326356 0.0306553
d1:f2-d1:f1  0.33868995 -0.25905215 0.9364320 0.5680611
d2:f2-d1:f1  0.64854101  0.05079891 1.2462831 0.0254221
d1:f3-d1:f1  0.69601055  0.09826845 1.2937526 0.0128711
d2:f3-d1:f1  1.13713411  0.53939201 1.7348762 0.0000044
d1:f2-d2:f1 -0.29620356 -0.89394566 0.3015385 0.7007889
d2:f2-d2:f1  0.01364750 -0.58409459 0.6113896 0.9999998
d1:f3-d2:f1  0.06111704 -0.53662505 0.6588591 0.9996758
d2:f3-d2:f1  0.50224060 -0.09550149 1.0999827 0.1515870
d2:f2-d1:f2  0.30985106 -0.28789103 0.9075932 0.6591096
d1:f3-d1:f2  0.35732060 -0.24042149 0.9550627 0.5089535
d2:f3-d1:f2  0.79844416  0.20070206 1.3961863 0.0025700
d1:f3-d2:f2  0.04746954 -0.55027256 0.6452116 0.9999065
d2:f3-d2:f2  0.48859310 -0.10914900 1.0863352 0.1742960
d2:f3-d1:f3  0.44112356 -0.15661854 1.0388657 0.2721714
```

References

1. STHDA. *Two-Way ANOVA Test in R*, Retrieved April 25, 2023.

   http://www.sthda.com/english/wiki/two-way-anova-test-in-r

2. Bluman, A. G. (2018). Chapters 11 & 12. In *Elementary statistics Book*. McGraw-Hill.