Module 2 Assignment — Chi-Square Testing and ANOVA

Sourabh D. Khot (ID 002754952)

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics CRN 81176

Professor Behzad Ahmadi

April 24, 2022

Table of Contents

Introduction

In this assignment, my goal is to understand the concepts of the Chi-Square Goodness-of-Fit test, Chi-Square Independence test, One-Way ANOVA for testing equivalence of multiple means, Scheffé or Tukey test for identifying which of the means is different, and Two-Way ANOVA for testing the impact of multiple factors on an outcome variable.

I will implement these concepts hands-on in R on ten different datasets. My steps will include importing the data, analyzing the data, making a claim with a significance level, formulating the hypothesis, finding critical value, computing the test value in R, making a decision, and finally summarizing the results. In the conclusion section at the end, I will provide a final consolidated recommendation per dataset.

In the first eight datasets, I have used the traditional method of hypothesis testing by finding critical value from reference tables and comparing it with the computed test value from R to arrive at a decision. In the final two datasets, I will also use the p-value method by comparing the p-value obtained from R with the decided significance level and then arrive at a decision.

Analysis

Section 11-1: Goodness-of-Fit Test

For determining Goodness-of-Fit, I will use the Chi-Square test, which is right-sided.

Q6. Blood Types

As a medical researcher, I have to test whether a large hospital has the same blood group distribution as the overall population from the observed sample of 50 patients with a 0.10 level of significance. I have summarized the data and computed the expected distribution in Table 1. The frequency in each category is greater than 5, and the data is randomly sampled.

Blood Type		**Type A**	**Type B**	**Type O**	**Type AB**	**Total (computed)**
Population		20%	28%	36%	16%	100%
Sample n=50	Observed	12	8	24	6	50
	Expected (computed)	10	14	18	8	50

Table 1. Blood Type Data

a. State the hypotheses and identify the claim.

H_0: $P_A = 0.20$, $P_B = 0.28$, $P_O = 0.36$, $P_{AB} = 16\%$

H_1: The distribution differs from the one stated in null hypothesis

b. Find the critical value.

$\alpha = 0.10$, d.f. = 3

critical value = 6.251 (from reference table)

c. Compute the test value.

```
> (result1 <- chisq.test(x = observed1, p = p1))

        Chi-squared test for given probabilities

data:  observed1
X-squared = 5.4714, df = 3, p-value = 0.1404

> ifelse(result1$p.value > alpha1, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Not enough evidence to reject H0"
```

Figure 1. Blood Type Goodness-of-Fit Test

d. Make the decision.

X-squared (5.471) is less than the critical value (6.251). Hence, there is not enough evidence to reject H_0 that the hospital distribution is the same as the population distribution.

e. Summarize the results

It can be concluded that hospital patients in the large hospital have the same blood type distribution as the general population.

Q8. On-Time Performance by Airlines

I have to test whether a major airline company has the same trend of on-time performance as the government statistics at a significance level of 0.05. The data is computed in Table 2, where the sample is randomly selected, and each category is greater than 5.

Action		On-Time	National Aviation System delay	Aircraft Arriving Late	Other	Total (computed)
Population		70.8%	8.2%	9.0%	12.0%	100%
Sample n=200	Observed	125	10	25	40	200
	Expected (computed)	141.6	16.4	18	24	200

Table 2. Airline On-Time Performance Data

a. State the hypotheses and identify the claim.

H_0: $P_{OT} = 0.708$, $P_{NASD} = 0.082$, $P_{AAL} = 0.090$, $P_O = 0.120$

H_1: The distribution differs from the one stated in the null hypothesis

b. Find the critical value.

$\alpha = 0.05$, d.f. = 3

critical value = 7.815 (from reference table)

c. Compute the test value.

```
> (result2 <- chisq.test(x = observed2, p = p2))

        Chi-squared test for given probabilities

data:  observed2
X-squared = 17.832, df = 3, p-value = 0.0004763

> ifelse(result2$p.value > alpha2, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Figure 2. Airline On-Time Performance Goodness-of-Fit Test

d. Make the decision.

X-squared (17.832) is greater than the critical value (7.815). Hence, there is enough evidence to reject H_0 that the airline's on-time performance is the same as government statistics.

e. Summarize the results

It can be conclusively said that the major airline company differs in its on-time performance trend from the numbers maintained by the Bureau of Transportation Statistics.

Section 11-2: Independence of Variables Test

I will use the Chi-Square Independence Test to test whether two variables are independent of each other from sample data.

Q8. Ethnicity and Movie Admissions

I have to test whether the yearly movie attendance depends on the movie-goers' ethnicity at a significance level of 0.05. The observed sample data and expected values in brackets are given in Table 3. Each expected value is greater than 5, and the sample is random.

Ethnicity Year	Caucasian	Hispanic	African American	Other	Total
2013 (Expected)	724 (639.04)	335 (366.25)	174 (190.43)	107 (144.28)	1340
2014 (Expected)	370 (454.96)	292 (260.75)	152 (135.57)	140 (102.72)	954
Total	1094	627	326	247	2294

Table 3. Movie Admissions Contingency Table

a. State the hypotheses and identify the claim.

H_0: Yearly movie attendance is independent of the movie-goer's ethnicity

H_1: Yearly movie attendance is dependent upon the movie-goer's ethnicity

b. Find the critical value.

$\alpha = 0.05$, d.f. = (2-1)*(4-1) = 3

critical value = 7.815 (from reference table)

c. Compute the test value

```
> (result3 <- chisq.test(m3))

        Pearson's Chi-squared test

data:  m3
X-squared = 60.144, df = 3, p-value = 5.478e-13

> ifelse(result3$p.value > alpha3, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Figure 3. Movie Admissions Independence Test

d. Make the decision.

X-squared (60.144) is greater than the critical value (7.815). Hence, there is enough evidence to reject H_0 that the yearly attendance and ethnicity are independent of each other.

e. Summarize the results.

The movie attendant by year is dependent upon the ethnicity of movie-goers. In other words, in 2013 and 2014, the number of movie-goers by ethnicity varied.

Q10. Women in the Military

At a 0.05 level of significance, I have to analyze whether there is a relationship between military rank and branch of military women from a sample with expected values given in Table 4. Each cell has more than 5 counts and the sample is randomly selected.

Rank **Branch**	**Officers (Expected)**	**Enlisted (Expected)**	**Total**
Army	10,791 (11,463.06)	62,491 (61,818.94)	73,282
Navy	7,816 (7909.73)	42,750 (42,656.27)	50,566
Marine Corps	932 (1635.73)	9,525 (8821.27)	10,457
Air Force	11,819 (10,349.48)	54,344 (55,813.52)	66,163
Total	31,358	169,110	200,468

Table 4. Women Military Contingency Table

a. State the hypotheses and identify the claim.

H_0: Among military women, rank and branch are independent of each other

H_1: Among military women, rank and branch are dependent on each other

b. Find the critical value.

α = 0.05, d.f. = (4-1)*(2-1) = 3

critical value = 7.815 (from reference table)

c. Compute the test value.

```
> (result4 <- chisq.test(m4))

	Pearson's Chi-squared test

data:  m4
X-squared = 654.27, df = 3, p-value < 2.2e-16

> ifelse(result4$p.value > alpha4, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Figure 4. Women Military Independence Test

d. Make the decision.

X-squared (654.270) is greater than the critical value (7.815). Hence, there is enough evidence to reject H_0 that rank and branch are independent of each other.

e. Summarize the results.

In women's armed forces, it can be concluded that rank is indeed dependent upon the branch of service, or vice versa.

Section 12-1: One-Way ANOVA

In sections 12-1 and 12-2, I will use the one-way ANOVA F-test to compare multiple means. As a criterion to perform this test, I have made the following assumptions: All variables are normally distributed, the samples are independent, the population variances are equal, and samples are simple random samples from respective populations.

Q8. Sodium Contents of Foods

From the sample given in Table 5, I have to determine whether the mean sodium is different in condiments, cereals, and desserts at α of 0.05.

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

Table 5. Sodium in Food Data

a. State the hypotheses and identify the claim.

Null hypothesis is that sodium content in condiments, cereals, and desserts is the same.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Alternative hypothesis is that sodium content in at least one of these foods is different.

H_1: At least one mean is different from others

b. Find the critical value.

α = 0.05, k = 3, N = 22, d.f.N. = k – 1 = 2, d.f.D. = N – k = 19,

critical value = 3.52 (from reference table)

c. Compute the test value.

```
> (result5 <- summary(aov(sodium ~ food, data = df5)))
            Df Sum Sq Mean Sq F value Pr(>F)
food         2  27544   13772   2.399  0.118
Residuals   19 109093    5742
> ifelse(result5[[1]][[1,"Pr(>F)"]] > alpha5, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Not enough evidence to reject H0"
```

Figure 5. Sodium in Food One-Way ANOVA

d. Make the decision.

F-value (2.399) is less than the critical value (3.52). Hence, there is insufficient evidence to reject H_0 that all sodium content in the three foods is the same.

e. Summarize the results.

It can be deduced from the given sample that the amount of sodium in one serving of condiments, cereals, and desserts is the same.

Section 12-2: Scheffé or Tukey Test

Similar to 12-1, I will perform a one-way ANOVA to compare means, considering all assumptions are met. In cases where the null hypothesis is rejected, I will further use either the Scheffé or Tukey test to verify which of the means has the significant difference.

Q10. Sales for Leading Companies

At α of 0.01, I have to determine whether mean sales in a million dollars differ for three products of leading companies from the sample given in Table 6.

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

Table 6. Leading Company Sales Data

a. State the hypotheses and identify the claim.

Null hypothesis states that the mean sales of all products are the same.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Alternative hypothesis states that at least one product type has a different mean sales.

H_1: At least one mean is different from others

b. Find the critical value.

α = 0.01, k = 3, N = 14, d.f.N. = k – 1 = 2, d.f.D. = N – k = 11,

critical value = 7.21 (from reference table)

c. Compute the test value.

```
> (result6 <- summary(aov(sales ~ product, data = df6)))
            Df Sum Sq Mean Sq F value Pr(>F)
product      2 103770   51885   2.172   0.16
Residuals   11 262795   23890
> ifelse(result6[[1]][[1,"Pr(>F)"]] > alpha6, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Not enough evidence to reject H0"
```

Figure 6. Leading Company Sales One-Way ANOVA

d. Make the decision.

F-value (2.172) is less than the critical value (7.21). Hence, there is not enough evidence to reject H_0 that mean sales of all products is same.

e. Summarize the results.

Basis the sample and level of significance, it can be concluded that the mean sales of the three products by leading companies are the same. As no mean is significantly different, I will not further use the Scheffé or Tukey test.

Q12. Per-Pupil Expenditures

Using the sample given in Table 7 and $\alpha = 0.05$, I will explore if there is a difference in mean expenditure per pupil in three sections of the country.

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

Table 7. Per-Pupil Expenditure Data

a. State the hypotheses and identify the claim.

Null hypothesis states the mean expenditure per pupil is the same across the sections.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Alternative hypothesis states at least one of the sections has a different mean expenditure.

H_1: At least one mean is different from others

b. Find the critical value.

$\alpha = 0.05$, $k = 3$, $N = 13$, d.f.N. = $k - 1 = 2$, d.f.D. = $N - k = 10$,

critical value = 4.10 (from reference table)

c. Compute the test value.

```
> (result7 <- summary(aov(expenditure ~ section, data = df7)))
            Df  Sum Sq Mean Sq F value Pr(>F)
section      2 1244588  622294   0.649  0.543
Residuals   10 9591145  959114
> ifelse(result7[[1]][[1,"Pr(>F)"]] > alpha7, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Not enough evidence to reject H0"
```

Figure 7. Per-Pupil Expenditure One-Way ANOVA

d. Make the decision.

F-value (0.649) is less than the critical value (4.10). Hence, there is not enough evidence to reject H_0 that the mean expenditures are the same across the country's three sections.

e. Summarize the results.

It can be understood from the sample data that all the three sections of the country have the same mean expenditure per pupil. Since no difference in means is detected, the Scheffé or Tukey test will not be further required.

Section 12-3: Two-Way ANOVA

I will use the two-way ANOVA to test two independent factors' individual effects and interaction on a dependent variable. I have considered that all assumptions (normal distribution, independent samples, equal variances, simple random samples) are met.

Q10. Increasing Plant Growth

A random sample of 12 plants is exposed to a combination of grow-lights (factor A) and plant food supplements (factor B), and their growth is measured and shown in Table 8a. I have to determine at $\alpha = 0.05$ whether there is an interaction between these two factors and if there is a difference in mean growth due to light and due to plant food.

Factor A **Factor B**	**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

Table 8a. Increasing Plant Growth Data

		Factor A	
		Grow light 1	Grow light 2
Factor B	Plant food A	Grow light 1 Plant food A	Grow light 2 Plant food A
	Plant food B	Grow light 1 Plant food B	Grow light 2 Plant food B

Table 8b. Increasing Plant Growth Treatment Groups

a. State the hypotheses.

This is a two-by-two ANOVA with four treatment groups. I will state the interaction hypothesis, followed by grow light groups (factor A) and plant food groups (factor B).

$H_{0,\ A*B}$: There is NO interaction between grow lights and plant foods.
$H_{1,\ A*B}$: There is an interaction between grow lights and plant foods.
$H_{0,\ A}$: The mean growth of the two grow light groups is equal
$H_{1,\ A}$: The mean growth of the two grow light groups is different
$H_{0,\ B}$: The mean growth of the two plant food groups is equal
$H_{1,\ B}$: The mean growth of the two plant food groups is different

b. Find the critical value for each F-test.

$\alpha = 0.05$, $a = 2$, $b = 2$, $n = 3$,
Factor A: $d.f.N._A = a - 1 = 1$, Factor B: $d.f.N._B = b - 1 = 1$,
Interaction (A*B): $d.f.N_{A*B} = (a - 1)(b - 1) = 1$, Within (error): $d.f.D. = ab(n - 1) = 8$
critical value = 5.32 for F_{A*B}, F_A and F_B

c. Complete the summary table and find the test value.

```
> (result8 <- summary(aov(growth ~ facA * facB, data = df8)))
            Df Sum Sq Mean Sq F value  Pr(>F)
facA         1  1.920   1.920   3.681 0.09133 .
facB         1 12.813  12.813  24.562 0.00111 **
facA:facB    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
> ifelse(result8[[1]][[3,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.A*B",
+        "Sufficient evidence to reject H0.A*B")
[1] "Not enough evidence to reject H0.A*B"
> ifelse(result8[[1]][[1,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.A",
+        "Sufficient evidence to reject H0.A")
[1] "Not enough evidence to reject H0.A"
> ifelse(result8[[1]][[2,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.B",
+        "Sufficient evidence to reject H0.B")
[1] "Sufficient evidence to reject H0.B"
```

Figure 8. Increasing Plant Growth Data Two-Way ANOVA

d. Make the decision.

F-value for interaction A*B (1.438) and Factor A (3.681) is less than the critical value (5.32), hence there is not enough evidence to reject $H_{0,\ A*B}$ (no interaction between the two factors) $H_{0,\ A}$ (equal mean growth of the two grow light groups).

However, F-value for Factor B (24.562) is more than the critical value (5.32). Hence there is sufficient evidence to reject $H_{0,\ B}$ (equal mean growth of the two plant food groups).

e. Summarize the results.

It can be concluded that plant growth varies depending upon the type of supplement food. However, there is no significant difference in growth due to the type of growth light and no significant interaction between the type of supplement food and the type of grow light.

On Your Own: A. Baseball Dataset

1. Importing Baseball Dataset

I imported the dataset into a dataframe ‘bb’ using the read.csv() function in R.

2. Exploratory Data Analysis

This dataset contains 1232 records on 30 major league baseball teams that participated in American League and National League between 1962 and 2012. There are 15 variables about each baseball team. The descriptive statistics for numerical variables is shown in Table 9a.

Wins variable has a minimum of 40, medium of 81, mean of 80.9 and maximum of 116. It is nearly symmetrical with skewness of just -0.18.

```
> summary(bb)
     Team              League               Year            RS               RA        
 Length:1232        Length:1232        Min.   :1962   Min.   : 463.0   Min.   : 472.0  
 Class :character   Class :character   1st Qu.:1977   1st Qu.: 652.0   1st Qu.: 649.8  
 Mode  :character   Mode  :character   Median :1989   Median : 711.0   Median : 709.0  
                                       Mean   :1989   Mean   : 715.1   Mean   : 715.1  
                                       3rd Qu.:2002   3rd Qu.: 775.0   3rd Qu.: 774.2  
                                       Max.   :2012   Max.   :1009.0   Max.   :1103.0  

       W              OBP              SLG               BA            Playoffs     
 Min.   : 40.0   Min.   :0.2770   Min.   :0.3010   Min.   :0.2140   Min.   :0.0000  
 1st Qu.: 73.0   1st Qu.:0.3170   1st Qu.:0.3750   1st Qu.:0.2510   1st Qu.:0.0000  
 Median : 81.0   Median :0.3260   Median :0.3960   Median :0.2600   Median :0.0000  
 Mean   : 80.9   Mean   :0.3263   Mean   :0.3973   Mean   :0.2593   Mean   :0.1981  
 3rd Qu.: 89.0   3rd Qu.:0.3370   3rd Qu.:0.4210   3rd Qu.:0.2680   3rd Qu.:0.0000  
 Max.   :116.0   Max.   :0.3730   Max.   :0.4910   Max.   :0.2940   Max.   :1.0000  

   RankSeason     RankPlayoffs         G              OOBP             OSLG       
 Min.   :1.000   Min.   :1.000   Min.   :158.0   Min.   :0.2940   Min.   :0.3460  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:162.0   1st Qu.:0.3210   1st Qu.:0.4010  
 Median :3.000   Median :3.000   Median :162.0   Median :0.3310   Median :0.4190  
 Mean   :3.123   Mean   :2.717   Mean   :161.9   Mean   :0.3323   Mean   :0.4197  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:162.0   3rd Qu.:0.3430   3rd Qu.:0.4380  
 Max.   :8.000   Max.   :5.000   Max.   :165.0   Max.   :0.3840   Max.   :0.4990  
 NA's   :988     NA's   :988                     NA's   :812      NA's   :812     
```

Table 9a. Baseball Data Descriptive Statistics

3. Determine if Number of Wins differs by decade

I will check if total wins by decades are equal using the Chi-Square Goodness-of-Fit test. I have considered all assumptions (normal distribution, independent samples, equal variances, simple random samples) are met. In the original dataset, I have converted years into decades and summed the wins by decade to get data in Table 9b, where I have also added the expected wins.

Decade	1960s	1970s	1980s	1990s	2000s	2010s	Total
Exp Freq	1/6	1/6	1/6	1/6	1/6	1/6	1
Observed	13,267	17,934	18,926	17,972	24,286	7,289	99,674
Expected	16,612.3	16,612.3	16,612.3	16,612.3	16,612.3	16,612.3	99,674

Table 9b. Baseball Wins by Decades Data

a. State the hypotheses and identify the claim.

H_0: $P_{1960s} = P_{1970s} = P_{1980s} = P_{1990s} = P_{2000s} = P_{2010s} = 1/6$ (total wins are equal each decade)
H_1: The distribution differs from the one stated in the null hypothesis

b. Find the critical value (α = 0.05).

α = 0.05, d.f. = 5
critical value = 11.07 (from reference table)

c. Compute the test value.

```
> (result9 <- chisq.test(x = winsbydecade$wins, p = p9))

        Chi-squared test for given probabilities

data:  winsbydecade$wins
X-squared = 9989.5, df = 5, p-value < 2.2e-16

> ifelse(result9$p.value > alpha9, "Not enough evidence to reject H0",
+        "Sufficient evidence to reject H0")
[1] "Sufficient evidence to reject H0"
```

Figure 9. Baseball Wins by Decades Goodness-of-Fit Test

d. Make the decision.

By the traditional method of comparing test value with critical value, it can be observed that X-squared (9989.5) is greater than the critical value (11.07). Hence, there is enough evidence to reject H_0 that wins are equal every decade. Thus, at least one or more decades had a different total win than the other decades.

e. Comparing the results from two methods: 'the critical value with the test value' vs. 'the p-value from R with the significance level'.

By the p-value method of comparing p-value with significance level, the p-value ($2.2*10^{-16}$) is much lesser than the alpha of 0.05. Hence, we have enough evidence to reject H_0 that wins are equal every decade. This is the same result that I got from the traditional method.

On Your Own: B. Crop Dataset

4. Importing Crop Dataset

I imported the dataset into a dataframe ‘crop’ using the read.csv() function in R.

5. How Yield is impacted by Fertilizer and Density

To understand if fertilizer (factor A) and density (factor B) have an impact on impact yield, I will perform a Two-Way ANOVA test using $\alpha = 0.05$. I have considered that all assumptions (normal distribution, independent samples, equal variances) are met and verified all groups are of equal size of 16. This test is of 2*3 design having six treatment groups, as given in Table 10.

		Factor A Fertilizer type	
Factor B Density type	Fertilizer 1 Density 1	Fertilizer 2 Density 1	Fertilizer 3 Density 1
	Fertilizer 1 Density 2	Fertilizer 2 Density 2	Fertilizer 3 Density 2

Table 10. Increasing Crop Yield Treatment Groups

a. State the hypotheses.

First, I will state the interaction hypothesis, followed by the fertilizer group (factor A) and the density groups (factor B) hypotheses.

$H_{0,\,A*B}$: There is NO interaction between fertilizer and density types.
$H_{1,\,A*B}$: There is an interaction between fertilizer and density types.
$H_{0,\,A}$: The mean yield of the groups by fertilizer type is equal
$H_{1,\,A}$: The mean yield of the groups by fertilizer type is different
$H_{0,\,B}$: The mean yield of the groups by density type is equal
$H_{1,\,B}$: The mean yield of the groups by density type is different

b. Find the critical value for each F-test.

$\alpha = 0.05$, $a = 3$, $b = 2$, $n = 16$,
Factor A: d.f.N.$_A$ = 2, Factor B: d.f.N.$_B$ = 1,
Interaction (A*B): d.f.N$_{A*B}$ = 2, Within (error): d.f.D. = 90
critical value = 3.098 for F_{A*B}, F_A critical value = 3.947 for F_B

c. Complete the summary table and find the test value.

```
> (result10 <- summary(aov(yield ~ fertilizer * density, data = crop)))
                  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
> ifelse(result10[[1]][[3,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.A*B",
+        "Sufficient evidence to reject H0.A*B")
[1] "Not enough evidence to reject H0.A*B"
> ifelse(result10[[1]][[1,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.A",
+        "Sufficient evidence to reject H0.A")
[1] "Sufficient evidence to reject H0.A"
> ifelse(result10[[1]][[2,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.B",
+        "Sufficient evidence to reject H0.B")
[1] "Sufficient evidence to reject H0.B"
```

Figure 10. Increasing Crop Yield Two-Way ANOVA

d. Make the decision.

Regarding interaction A*B, the F-value (0.447) is less than critical value (3.098) and p-value (0.505) is greater than α (0.05). Hence, we do not have sufficient evidence to reject $H_{0, A*B}$ (no interaction between the two factors).

For factor A and factor B, their F-values (17.078, 15.230) are greater than their respective critical values (3.098, 3.947), and their p-values ($7.9*10^{-5}$, 0.000181) are less than α (0.05). Hence, we have sufficient evidence to reject $H_{0, A}$ (equal mean yield by fertilizer type) and reject $H_{0, B}$ (equal mean yield by density type).

e. Summarize the results.

It can be concluded that both fertilizer type and density type DO have an impact on yield. However, there is no interaction between fertilizer and density, meaning they work independently on yield.

Conclusion

I have completed testing the ten datasets for various claims made and will summarize the interpretation and make my recommendations in this section. Both traditional and p-value methods yielded the same results, and any of them can be used.

The patients of the large hospital have the same blood type distribution as the general population; hence the hospital fairly represents the population, and there is no anomaly in the large hospital that needs to be investigated.

The major airline company has a different on-time performance than the government's average. It needs to be further investigated if the airline is better or worse. Accordingly other airlines can learn from it, or this airline needs to be improved.

Movie-goers for any year depend upon the ethnicity mix. Hence, Hollywood needs to understand the yearly trend of ethnicity to make films that appeal to the mix. Similarly, a movie theater needs to predict the trend of ethnicity to project future annual revenue.

In women's military, rank is dependent upon the branch. So if a woman wants to rise higher in ranks, she must analyze the branch having the best promotions and choose accordingly.

The mean amount of sodium is the same in condiments, cereals, and desserts. Hence, a patient with high blood pressure who may be sensitive to sodium need not be concerned about choosing one among these food types.

Mean sales of cereal, chocolate candy, and coffee are the same. Hence, a new company targeting the leading companies' market share, needs to launch a portfolio mix that equally focuses on all three product types.

Per-pupil expenditure in all sections of the country is the same. Hence a private school should not consider the per-pupil expenditure, when deciding where to open a new school.

Plant growth is dependent upon the type of food supplement used. Hence a farmer should further analyze which plant supplement is better for growth, and choose accordingly. The farmer should not be concerned about the type of growth light or its interaction with the supplement.

Total wins in all baseball matches are different for each decade. Hence, baseball tournament organizers need to review their budget to reward the winning teams every decade.

Crop yield is dependent on both fertilizer type and density type, with no interaction between themselves. Hence, a farming company should investigate the type of fertilizer and the type of density increases yield and choose accordingly.

References

APA Style Table: APA.org. (n.d.). Retrieved from https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables

Bluman, A. G. (2018). *Elementary Statistics.* New York: McGraw Hill Education.

Canvas Module 1 Assignment. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110053/assignments/1345425

Chi-squared Test of Independence: R-Tutor.com. (n.d.). Retrieved from http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

Create DataFrame Row by Row in R: GeeksforGeeks.org. (n.d.). Retrieved from https://www.geeksforgeeks.org/create-dataframe-row-by-row-in-r/

Hypothesis Testing: UC Denver. (n.d.). Retrieved from http://math.ucdenver.edu/~ssantori/MATH2830SP13/Math2830-Chapter-08.pdf

R – Data Frames: GeeksforGeeks.org. (n.d.). Retrieved from https://www.geeksforgeeks.org/r-data-frames/

R – Matrices: GeeksforGeeks.org. (n.d.).

Two-Way ANOVA Test in R: STHDA.com. (n.d.). Retrieved from http://www.sthda.com/english/wiki/two-way-anova-test-in-r

Writing a dissertation: University of Cambridge. (n.d.). Retrieved from https://www.cl.cam.ac.uk/~pr10/teaching/dissertation.html

Appendix

R Code

```
# Importing libraries
library(dplyr)
library(tidyverse)
library(psych)

# 11-1: Goodness-of-Fit                    ####
## Q6. Blood Types                         ####
alpha1 = 0.10                   # significance level
observed1 <- c(12, 8, 24, 6)      # observations
p1 <- c(0.20, 0.28, 0.36, 0.16)   # expected probabilities
(result1 <- chisq.test(x = observed1, p = p1))
ifelse(result1$p.value > alpha1, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

## Q8. On-Time Performance by Airlines      ####
alpha2 = 0.05
observed2 <- c(125, 10, 25, 40)
p2 <- c(0.708, 0.082, 0.090, 0.120)
(result2 <- chisq.test(x = observed2, p = p2))
ifelse(result2$p.value > alpha2, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

# 11-2: Independence of Variables          ####
## Q8. Ethnicity and Movie Admissions       ####
alpha3 = 0.05
m3 <- matrix(c(724,335,174,107,370,292,152,140), nrow = 2, byrow = TRUE)
rownames(m3) = c(2013,2014)
colnames(m3) = c("Caucasian","Hispanic","African American","Other")
(result3 <- chisq.test(m3))
ifelse(result3$p.value > alpha3, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

## Q10. Women in the Military               ####
alpha4 = 0.05
m4 <- matrix(c(10791,62491,7816,42750,932,9525,11819,54344), nrow = 4, byrow = TRUE)
rownames(m4) = c("Army","Navy","Marine Corps","Air Force")
colnames(m4) = c("Officers","Enlisted")
(result4 <- chisq.test(m4))
ifelse(result4$p.value > alpha4, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

# 12-1: One-Way ANOVA                      ####
```

```
## Q8. Sodium Contents of Foods            ####
alpha5 = 0.05
df5 <- data.frame(
  sodium =
c(270,130,230,180,80,70,200,260,220,290,290,200,320,140,100,180,250,250,300,360,300,160),
  food = c( rep('condiments',7), rep('cereals',7), rep('desserts',8) )
)
(result5 <- summary(aov(sodium ~ food, data = df5)))
ifelse(result5[[1]][[1,"Pr(>F)"]] > alpha5, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

# 12-2: Scheffé or Tukey Test             ####
## Q10. Sales for Leading Companies        ####
alpha6 = 0.01
df6 <- data.frame(
  sales = c(578,320,264,249,237,311,106,109,125,173,261,185,302,689),
  product = c( rep('Cereal',5), rep('Chocolate Candy',5), rep('Coffee',4) )
)
(result6 <- summary(aov(sales ~ product, data = df6)))
ifelse(result6[[1]][[1,"Pr(>F)"]] > alpha6, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

## Q12. Per-Pupil Expenditures            ####
alpha7 = 0.05
df7 <- data.frame(
  expenditure = c(4946,5953,6202,7243,6113,6149,7451,6000,6479,5282,8605,6528,6911),
  section = c( rep('Eastern third',5), rep('Middle third',4), rep('Western third',4) )
)
(result7 <- summary(aov(expenditure ~ section, data = df7)))
ifelse(result7[[1]][[1,"Pr(>F)"]] > alpha7, "Not enough evidence to reject H0",
       "Sufficient evidence to reject H0")

# 12-3: Two-Way ANOVA                      ####
## Q10. Increasing Plant Growth            ####
alpha8 = 0.05
df8 <- data.frame(
  growth = c(9.2, 9.4, 8.9, 7.1, 7.2, 8.5, 8.5, 9.2, 8.9, 5.5, 5.8, 7.6),
  facA = c( rep('Grow-light 1',6), rep('Grow-light 2',6) ),
  facB = rep( c( rep('Plant food A',3), rep('Plant food B',3)), 2)
)
(result8 <- summary(aov(growth ~ facA * facB, data = df8)))
ifelse(result8[[1]][[3,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.A*B",
       "Sufficient evidence to reject H0.A*B")
ifelse(result8[[1]][[1,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.A",
       "Sufficient evidence to reject H0.A")
ifelse(result8[[1]][[2,"Pr(>F)"]] > alpha8, "Not enough evidence to reject H0.B",
```

```
      "Sufficient evidence to reject H0.B")

# On Your Own                        ####
## 1.   Importing Baseball Dataset          ####
bb = read.csv('Assign2/baseball.csv')

## 2.   Exploratory Data Analysis           ####
glimpse(bb)
head(bb)
summary(bb)
describe(bb)

## 3. Testing if Number of Wins differ by decade  ####
# derive decade from year
bb$Decade <- bb$Year - (bb$Year %% 10)

# wins sum by decade into table
winsbydecade <- bb %>%
  group_by(Decade) %>%
  summarize(wins = sum(W)) %>%
  as.tibble()
winsbydecade

# goodness-of-fit test assuming equal expected frequencies
alpha9 = 0.05                   # significance level
p9 <- rep(1/6,6)   # expected probabilities
(result9 <- chisq.test(x = winsbydecade$wins, p = p9))
ifelse(result9$p.value > alpha9, "Not enough evidence to reject H0",
      "Sufficient evidence to reject H0")

## 4. Importing Crop Dataset              ####
crop = read.csv('Assign2/crop_data.csv')

## 5. How Yield is impacted by Fertilizer & Density   ####
alpha10 = 0.05
(result10 <- summary(aov(yield ~ fertilizer * density, data = crop)))
ifelse(result10[[1]][[3,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.A*B",
      "Sufficient evidence to reject H0.A*B")
ifelse(result10[[1]][[1,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.A",
      "Sufficient evidence to reject H0.A")
ifelse(result10[[1]][[2,"Pr(>F)"]] > alpha10, "Not enough evidence to reject H0.B",
      "Sufficient evidence to reject H0.B")
```