Hypothesis Testing
Chi Square and ANOVA tests

Hai Vu

03 March, 2023

```
# Libraries import
library(dplyr)
library(knitr)
library(tidyverse)
library(tibble)
library(RColorBrewer)
library(xlsx)
library(ggplot2)
library(kableExtra)
library(formatR)
library(DescTools)
library(ggpubr)
library(psych)
```

I. INTRODUCTION SECTION

This project aims to apply my understanding of different Chi-square and ANOVA tests to solve a few problems. I will conduct several tests using one or more of the following methods:

- Test the goodness of fit of a distribution using Chi-square
- Test two variables for independence using Chi-square
- Test homogeneity of proportions using Chi-square
- One-way ANOVA to see if there is a significant difference between pairs of means
- Two-way ANOVA to see if there is a significant difference in the main effects or the interaction between variables

II. ANALYSIS SECTION

Task 1 (Blood Types)

```
alpha1 <- 0.10

# Expected:
exp_a <- 0.20 # Type A
exp_b <- 0.28 # Type B
exp_o <- 0.36 # Type O
exp_ab <- 0.16 # Type AB

# Observed:
```

```r
obs_a <- 12
obs_b <- 8
obs_o <- 24
obs_ab <- 6

# Data table:
expected1 <- c(exp_a,exp_b,exp_o,exp_ab)
observed1 <- c(obs_a,obs_b,obs_o,obs_ab)

df_test1 <- data.frame(expected1,observed1)

colnames(df_test1) <- c("Expected Values","Observed Values (n=50)")
rownames(df_test1) <- c("Type A","Type B","Type O","Type AB")
knitr::kable(df_test1, "simple",
             caption="Table 1.1. Blood types distribution") %>%
  kable_styling(position = "center")
```

Table 1.1. Blood types distribution

	Expected Values	Observed Values (n=50)
Type A	0.20	12
Type B	0.28	8
Type O	0.36	24
Type AB	0.16	6

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): $P_A = 0.20, P_B = 0.28, P_O = 0.36, P_{AB} = 0.16$

- Alternative Hypothesis (H_1): $P_A \neq 0.20$ or $P_B \neq 0.28$ or $P_O \neq 0.36$ or $P_{AB} \neq 0.16$, or that the blood type distribution is not similar to the stated distribution in the null hypothesis

2. Step 2-3. Find the critical values and compute the test values:

At $\alpha = 0.10$ and degree of freedom $= 3$, the critical value is 6.251 based on the Chi-Square distribution table. We begin performing the chi-square test:

```r
test1 <- chisq.test(observed1, p=expected1, correct=F)
test1
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed1
## X-squared = 5.4714, df = 3, p-value = 0.1404
```

The following graph helps visualize the differences between the observed and expected values:

```r
# Observed vs Expected values:
par(mai=c(1,1,1,1))

plot(expected1,col="blue", type="o", pch=16, ylim=c(0,1), xaxt="n",
     main="Blood Type distribution",
```

```
     sub=paste("Figure 1.1","\n"),
     ylab="Proportion",xlab="",cex.axis=0.8,cex.sub=0.9)
lines(observed1/sum(observed1),col="red",type="o", pch=16)
axis(1,at=c(1,2,3,4),
     labels=c("Type A","Type B","Type O","Type AB"),cex.axis=0.8)
legend("topright",c("Expected","Observed"),
       lty=1,col=c("blue","red"),pch=16,cex=0.8)
```

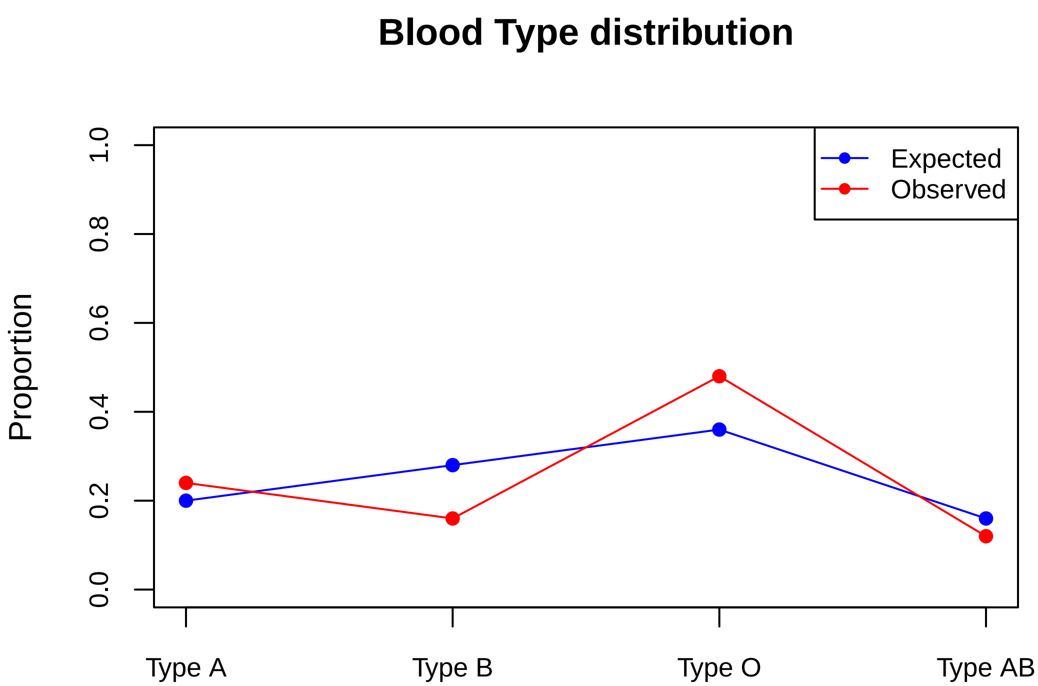

Figure 1.1

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`0.1403575`) is greater than α (`0.1`), we fail to reject the null hypothesis and conclude that the blood type distribution observed from the random sample is not different from the blood type distribution found in the general population.

Task 2 (On-time performance by airlines)

```
alpha2 <- 0.05

# Expected:
exp_ontime <- 0.708 # On time
exp_nas <- 0.082 # National Aviation System delay
exp_late <- 0.09 # Arriving late
exp_other <- 0.12 # Other reasons

# Observed:
```

```r
obs_ontime <- 125
obs_nas <- 10
obs_late <- 25
obs_other <- 40

# Data table:
expected2 <- c(exp_ontime,exp_nas,exp_late,exp_other)
observed2 <- c(obs_ontime,obs_nas,obs_late,obs_other)

df_test2 <- data.frame(expected2,observed2)

colnames(df_test2) <- c("Expected Values","Observed Values (n=200)")
rownames(df_test2) <- c("On Time","NAS delay","Late","Other reasons")
knitr::kable(df_test2, "simple",
             caption="Table 2.1. Airlines on-time performance distribution") %>%
  kable_styling(position = "center")
```

Table 2.1. Airlines on-time performance distribution

	Expected Values	Observed Values (n=200)
On Time	0.708	125
NAS delay	0.082	10
Late	0.090	25
Other reasons	0.120	40

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): $P_{on\_time} = 0.708, P_{nas} = 0.082, P_{late} = 0.09, P_{other} = 0.12$
- Alternative Hypothesis (H_1): $P_{on\_time} \neq 0.708$ or $P_{nas} \neq 0.082$ or $P_{late} \neq 0.09$ or $P_{other} \neq 0.12$, or that the on-time performance of airlines from the selected sample is not similar to the on-time performance recorded by the Bureau of Transport Statistics

2. Step 2-3. Find the critical values and compute the test values:

At $\alpha = 0.05$ and degree of freedom $= 3$, the critical value is 7.815 based on the Chi-Square distribution table. We begin performing the chi-square test:

```r
test2 <- chisq.test(observed2, p=expected2, correct=F)
test2
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed2
## X-squared = 17.832, df = 3, p-value = 0.0004763
```

The following graph helps visualize the differences between the observed and expected values:

```r
# Observed vs Expected values:
par(mai=c(1,1,1,1))

plot(expected2,col="orchid4", type="o", pch=16, ylim=c(0,1), xaxt="n",
```

```
     main="Airlines on-time performance distribution",
     sub=paste("Figure 2.1","\n"),
     ylab="Proportion",xlab="",cex.axis=0.8,cex.sub=0.9)
lines(observed2/sum(observed2),col="royalblue",type="o", pch=16)
axis(1,at=c(1,2,3,4),
     labels=c("On Time","NAS delay","Late","Other reasons"),cex.axis=0.8)
legend("topright",c("Expected","Observed"),
       lty=1,col=c("orchid4","royalblue"),pch=16,cex=0.8)
```

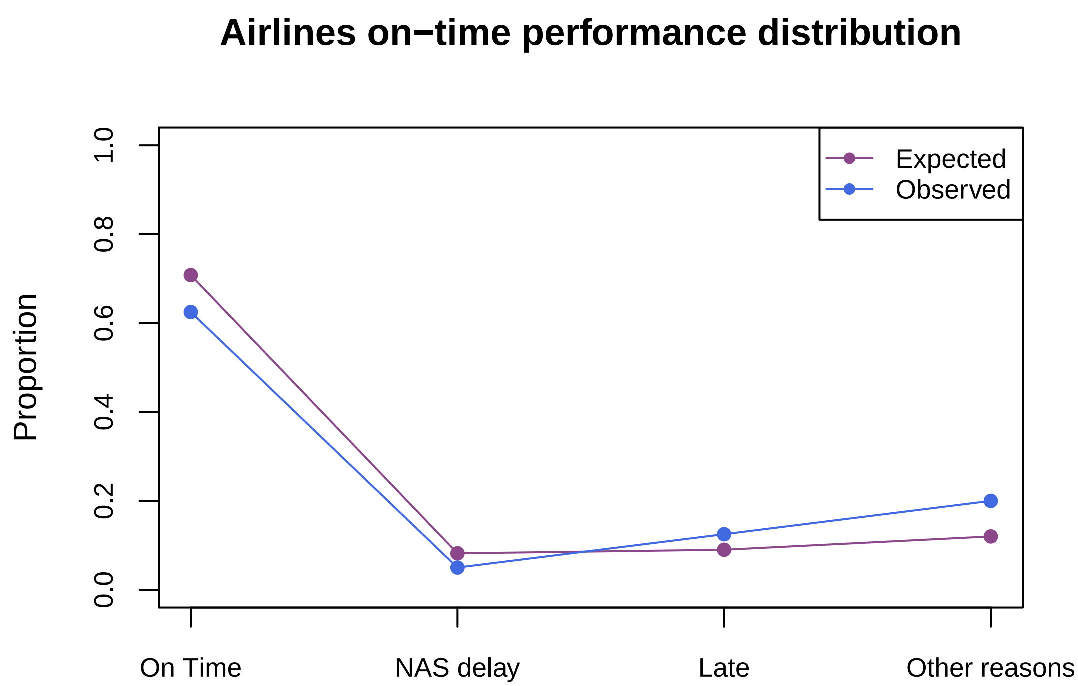

Figure 2.1

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`4.762587e-04`) is smaller than α (`0.05`), we are able to reject the null hypothesis and conclude that the on-time performance of airlines from the selected sample is not similar to the on-time performance recorded by the Bureau of Transport Statistics.

Task 3 (Ethnicity and movie admissions)

```
alpha3 <- 0.05

Y_2013 <- c(724, 335, 174, 107) # movie admissions in 2013
Y_2014 <- c(370, 292, 152, 140) # movie admissions in 2014

df_test3 <- data.frame(Y_2013,Y_2014)

colnames(df_test3) <- c(2013,2014)
```

```
rownames(df_test3) <- c("Caucasian","Hispanic","African American","Other")
knitr::kable(df_test3, "simple",
             caption="Table 3.1. Ethnicity and Movie Admissions by year") %>%
  kable_styling(position = "center")
```

Table 3.1. Ethnicity and Movie Admissions by year

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

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): Movie attendance by year is independent of ethnicity
- Alternative Hypothesis (H_1): Movie attendance by year is dependent upon ethnicity

2. Step 2-3. Find the critical values and compute the test values:

At $\alpha = 0.05$ and degree of freedom = 3, the critical value is 7.815 based on the Chi-Square distribution table. We begin performing the chi-square test:

```
test3 <- chisq.test(df_test3)
test3
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_test3
## X-squared = 60.144, df = 3, p-value = 5.478e-13
```

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`5.477507e-13`) is smaller than α (`0.05`), we are able to reject the null hypothesis and conclude that there is enough evidence to support the claim that the movie attendance by year is dependent upon ethnicity.

Task 4 (Women in the military)

```
alpha4 <- 0.05

army <- c(10791, 62491) # army
navy <- c(7816,42750) # navy
marine <- c(932,9525) # marine corps
airf <- c(11819,54344) # air force

# Data table:
df_test4 <- data.frame(army,navy,marine,airf)

rownames(df_test4) <- c("Officers","Enlisted")
colnames(df_test4) <- c("Army","Navy","Marine Corps","Air Force")
knitr::kable(format(df_test4,big.mark=","), "simple",
             caption="Table 4.1. Women personnel in the military by rank and branch") %>%
  kable_styling(position = "center")
```

Table 4.1. Women personnel in the military by rank and branch

	Army	Navy	Marine Corps	Air Force
Officers	10,791	7,816	932	11,819
Enlisted	62,491	42,750	9,525	54,344

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): Military ranking is independent of military branch for women in the Armed Forces
- Alternative Hypothesis (H_1): Military ranking is dependent upon military branch for women in the Armed Forces

2. Step 2-3. Find the critical values and compute the test values:

At $\alpha = 0.05$ and degree of freedom $= 3$, the critical value is 7.815 based on the Chi-Square distribution table. We begin performing the chi-square test:

```
test4 <- chisq.test(df_test4)
test4
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_test4
## X-squared = 654.27, df = 3, p-value < 2.2e-16
```

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`1.726418e-141`) is smaller than α (`0.05`), we are able to reject the null hypothesis and conclude that there is enough evidence to support the claim that the military ranking is dependent upon the military branch for women in the Armed Forces.

Task 5 (Sodium contents of foods)

```
alpha5 <- 0.05

cond <- c(270,130,230,180,80,70,200) # condiments
cereal <- c(260,220,290,290,200,320,140) # cereals
dessert <- c(100,180,250,250,300,360,300) # desserts

# Data table:
df_table <- data.frame(cond,cereal,dessert)

colnames(df_table) <- c("Condiments","Cereals","Desserts")
knitr::kable(df_table, "simple",
             caption="Table 5.1. Sodium Contents of Foods") %>%
  kable_styling(position = "center")
```

Table 5.1. Sodium Contents of Foods

Condiments	Cereals	Desserts
270	260	100
130	220	180
230	290	250
180	290	250
80	200	300
70	320	360
200	140	300

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): The mean sodium contents are similar among all three types of foods
- Alternative Hypothesis (H_1): At least one type of foods has the mean sodium content that is different from the others

2. Step 2-3. Find the critical values and compute the test values:

From table 5.1, we have N = 21 and k = 3. Therefore, d.f.N = k-1 = 2 and d.f.D = N-k = 18. Based on the F Distribution Table, the critical value at alpha = 0.05 is 3.5546. We begin performing the one-way ANOVA test:

```r
# Data preparation
df_test5 <- matrix(c(rep("Condiments",7),rep("Cereals",7),rep("Desserts",7),
                     cond,cereal,dessert),ncol=2)
df_test5 <- as.data.frame(df_test5)
names(df_test5) <- c("Food_Types","Sodium_Contents")
df_test5$Sodium_Contents <- as.numeric(as.character(df_test5$Sodium_Contents))

df_test5
```

```
##    Food_Types Sodium_Contents
## 1  Condiments             270
## 2  Condiments             130
## 3  Condiments             230
## 4  Condiments             180
## 5  Condiments              80
## 6  Condiments              70
## 7  Condiments             200
## 8     Cereals             260
## 9     Cereals             220
## 10    Cereals             290
## 11    Cereals             290
## 12    Cereals             200
## 13    Cereals             320
## 14    Cereals             140
## 15   Desserts             100
## 16   Desserts             180
## 17   Desserts             250
## 18   Desserts             250
## 19   Desserts             300
## 20   Desserts             360
## 21   Desserts             300
```

```r
# Conduct the one-way ANOVA test
test5.0 <- oneway.test(Sodium_Contents ~ Food_Types, data=df_test5, var.equal=T)
test5.1 <- aov(Sodium_Contents ~ Food_Types, data=df_test5)
summary(test5.1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Food_Types   2  30971   15486   2.727 0.0924 .
## Residuals   18 102229    5679
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (0.092) is greater than α (0.05), we fail to reject the null hypothesis and conclude that the mean sodium contents are similar among all three types of foods. Since there isn't any significant differences between the pairs of means, there is no need to conduct the Scheffe test or Tukey test.

```r
# Distribution of sodium contents of foods
ggplot(df_test5, aes(x = Food_Types, y = Sodium_Contents, fill = Food_Types)) +
    labs(title="Sodium content distribution by food types",caption="Figure 5.1") +
    scale_fill_brewer(palette="Set3") +
    theme(plot.title=element_text(hjust=0.5),plot.caption=element_text(hjust=0.5,size=11)) +
    geom_boxplot() +
    xlab(paste("Food Types","\n")) + ylab("Sodium Contents (mg)") +
    stat_summary(fun.y="mean", shape=15, color="red") +
    theme(legend.position = "none")
```

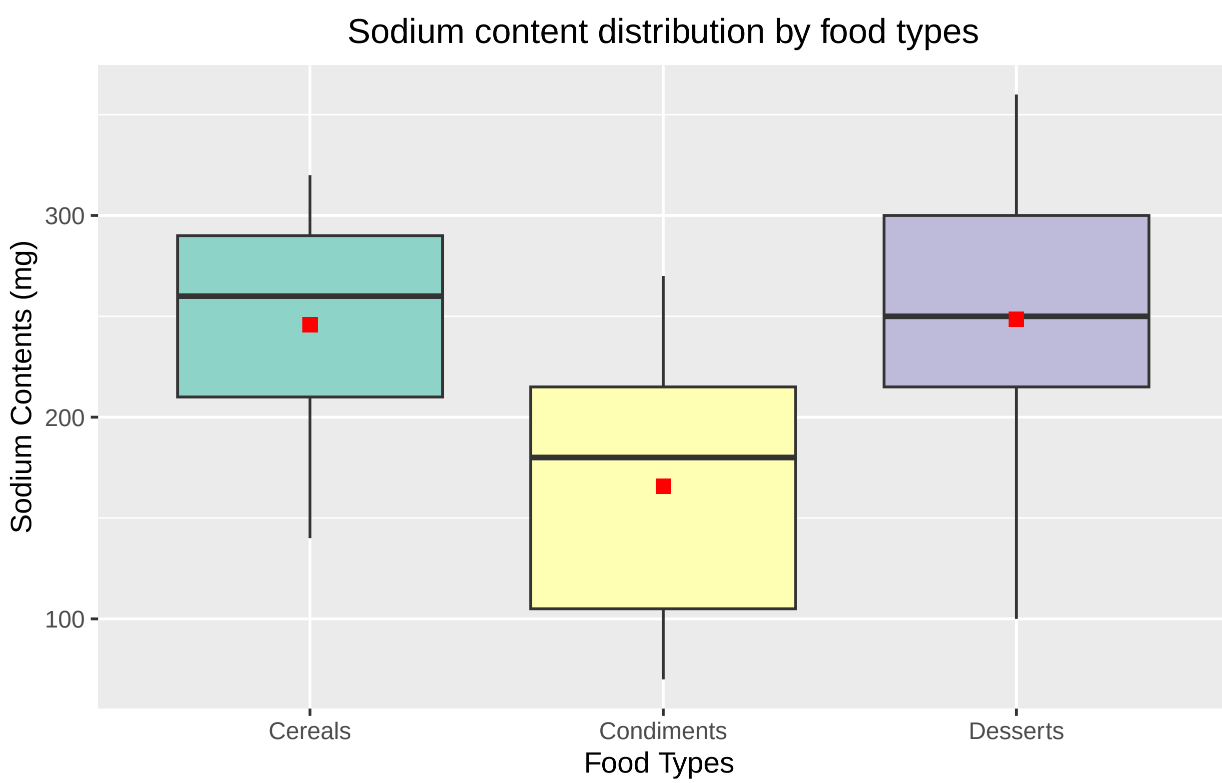

Figure 5.1

Task 6 (Sales for leading companies)

```
alpha6 <- 0.01

cereal2 <- c(578,320,264,249,237)
choco <- c(311,106,109,125,173)
coffee <- c(261,185,302,689,NA)

# Data table:
df_table2 <- data.frame(cereal2,choco,coffee)
colnames(df_table2) <- c("Cereal","Chocolate Candy","Coffee")
knitr::kable(df_table2, "simple",
             caption="Table 6.1. Sales for Leading Companies (in millions USD)") %>%
  kable_styling(position = "center")
```

Table 6.1. Sales for Leading Companies (in millions USD)

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

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): The average sales are similar among all three types of products
- Alternative Hypothesis (H_1): At least one type of products generate average sales that is different from the others

2. Step 2-3. Find the critical values and compute the test values:

From table 6.1, we have N = 14 and k = 3. Therefore, d.f.N = k-1 = 2 and d.f.D = N-k = 11. Based on the F Distribution Table, the critical value at alpha = 0.01 is 7.206. We begin performing the one-way ANOVA test:

```
# Data preparation
coffee2 <- c(261,185,302,689)
df_test6 <- matrix(c(rep("Cereal",5),rep("Chocolate Candy",5),rep("Coffee",4),
                     cereal2,choco,coffee2),ncol=2)
df_test6 <- as.data.frame(df_test6)
names(df_test6) <- c("Products","Sales")
df_test6$Sales <- as.numeric(as.character(df_test6$Sales))

df_test6
```

```
##          Products Sales
## 1          Cereal   578
## 2          Cereal   320
## 3          Cereal   264
## 4          Cereal   249
## 5          Cereal   237
```

```
## 6  Chocolate Candy   311
## 7  Chocolate Candy   106
## 8  Chocolate Candy   109
## 9  Chocolate Candy   125
## 10 Chocolate Candy   173
## 11          Coffee   261
## 12          Coffee   185
## 13          Coffee   302
## 14          Coffee   689
```

```r
# Conduct the one-way ANOVA test
test6.0 <- oneway.test(Sales ~ Products, data=df_test6, var.equal=T)
test6.1 <- aov(Sales ~ Products, data=df_test6)
summary(test6.1)
```

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

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`0.16`) is greater than α (`0.01`), we fail to reject the null hypothesis and conclude that the average sales are similar among all three types of products. Since there isn't any significant differences between the pairs of means, there is no need to conduct the Scheffe test or Tukey test.

```r
# Distribution of sodium contents of foods
ggplot(df_test6, aes(x = Products, y = Sales, fill = Products)) +
    labs(title="Sales distribution by products (in mUSD)",caption="Figure 6.1") +
    scale_fill_brewer(palette="Accent") +
    theme(plot.title=element_text(hjust=0.5),plot.caption=element_text(hjust=0.5,size=11)) +
    geom_boxplot() +
    xlab(paste("Products","\n")) + ylab("Sales Amount (mUSD)") +
    stat_summary(fun.y="mean", shape=15, color="red") +
    theme(legend.position = "none")
```

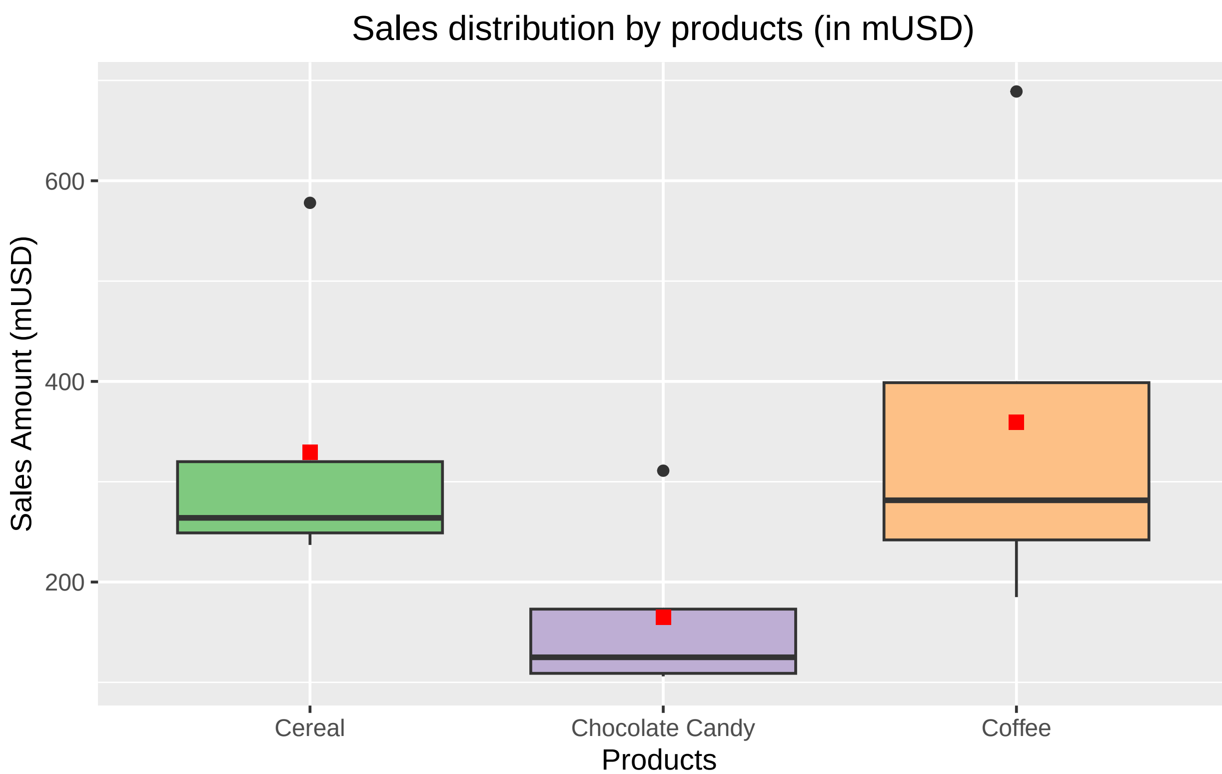

Figure 6.1

Task 7 (Per-pupil expenditures)

```
alpha7 <- 0.05

east <- c(4946,5953,6202,7243,6113)
mid <- c(6149,7451,6000,6479,NA)
west <- c(5282,8605,6528,6911,NA)

# Data table:
df_table3 <- data.frame(east,mid,west)
colnames(df_table3) <- c("Eastern third","Middle third","Western third")
knitr::kable(format(df_table3,big.mark=","), "simple",
             caption="Table 7.1. Per-Pupil Expenditures") %>%
  kable_styling(position = "center")
```

Table 7.1. Per-Pupil Expenditures

Eastern third	Middle third	Western third
4,946	6,149	5,282
5,953	7,451	8,605
6,202	6,000	6,528
7,243	6,479	6,911
6,113	NA	NA

1. Step 1. State the hypothesis and identify the claim:

- Null Hypothesis (H_0): The average expenditures per pupil are similar among all three sections of the country
- Alternative Hypothesis (H_1): At least one section of the country have different average expenditures per pupil compared to the other sections

2. Step 2-3. Find the critical values and compute the test values:

From table 7.1, we have N = 13 and k = 3. Therefore, d.f.N = k-1 = 2 and d.f.D = N-k = 10. Based on the F Distribution Table, the critical value at alpha = 0.05 is 4.1028. We begin performing the one-way ANOVA test:

```r
# Data preparation
mid2 <- c(6149,7451,6000,6479)
west2 <- c(5282,8605,6528,6911)

df_test7 <- matrix(c(rep("Eastern",5),rep("Middle",4),rep("Western",4),
                     east,mid2,west2),ncol=2)
df_test7 <- as.data.frame(df_test7)
names(df_test7) <- c("Location","Expenditure_per_Pupil")
df_test7$Expenditure_per_Pupil <- as.numeric(as.character(df_test7$Expenditure_per_Pupil))

df_test7
```

```
##    Location Expenditure_per_Pupil
## 1   Eastern                  4946
## 2   Eastern                  5953
## 3   Eastern                  6202
## 4   Eastern                  7243
## 5   Eastern                  6113
## 6    Middle                  6149
## 7    Middle                  7451
## 8    Middle                  6000
## 9    Middle                  6479
## 10  Western                  5282
## 11  Western                  8605
## 12  Western                  6528
## 13  Western                  6911
```

```r
# Conduct the one-way ANOVA test
test7.0 <- oneway.test(Expenditure_per_Pupil ~ Location, data=df_test7, var.equal=T)
test7.1 <- aov(Expenditure_per_Pupil ~ Location, data=df_test7)
summary(test7.1)
```

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

3. Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`0.543`) is greater than α (`0.05`), we fail to reject the null hypothesis and conclude that the average expenditures per pupil are similar among all three sections of the country. Since there isn't any significant differences between the pairs of means, there is no need to conduct the Scheffe test or Tukey test.

```r
# Distribution of sodium contents of foods
ggplot(df_test7, aes(x = Location, y = Expenditure_per_Pupil, fill = Location)) +
    labs(title="Expenditure per pupil by country section (in USD)",caption="Figure 7.1") +
    scale_fill_brewer(palette="Pastel2") +
    theme(plot.title=element_text(hjust=0.5),plot.caption=element_text(hjust=0.5, size=11)) +
    geom_boxplot() +
    xlab(paste("Country Section","\n")) + ylab("Expenditure (USD)") +
    stat_summary(fun.y="mean", shape=15, color="red") +
    theme(legend.position = "none")
```

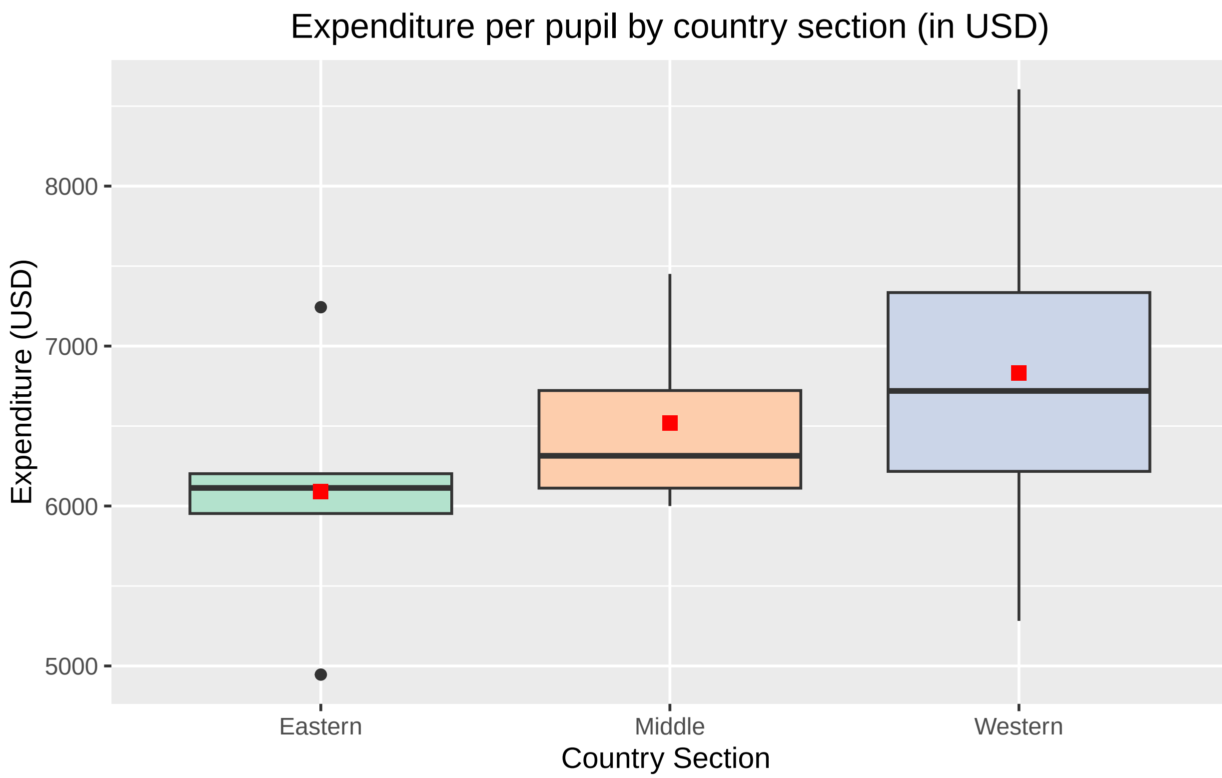

Figure 7.1

Task 8 (Increasing Plant Growth)

```r
alpha8 <- 0.05

fooda_l1 <- c(9.2,9.4,8.9)
fooda_l2 <- c(8.5,9.2,8.9)
foodb_l1 <- c(7.1,7.2,8.5)
foodb_l2 <- c(5.5,5.8,7.6)
growth <- c(fooda_l1,fooda_l2,foodb_l1,foodb_l2)

# Data preparation
df_test8 <- matrix(c(rep("Plant food A",6),rep("Plant food B",6),
            rep("Grow light 1",3),rep("Grow light 2",3),
            rep("Grow light 1",3),rep("Grow light 2",3),
```

```
            growth),ncol=3)

df_test8 <- as.data.frame(df_test8)
names(df_test8) <- c("Food_Supplement","Grow_Light","Growth_Length")
df_test8$Growth_Length <- as.numeric(df_test8$Growth_Length)
df_test8$Food_Supplement <- as.factor(df_test8$Food_Supplement)
df_test8$Grow_Light <- as.factor(df_test8$Grow_Light)

# Data table
knitr::kable(df_test8, "simple",
             caption="Table 8.1. Plant Growth by Food Supplement & Grow Light (in inches)") %>%
  kable_styling(position = "center")
```

Table 8.1. Plant Growth by Food Supplement & Grow Light (in inches)

Food_Supplement	Grow_Light	Growth_Length
Plant food A	Grow light 1	9.2
Plant food A	Grow light 1	9.4
Plant food A	Grow light 1	8.9
Plant food A	Grow light 2	8.5
Plant food A	Grow light 2	9.2
Plant food A	Grow light 2	8.9
Plant food B	Grow light 1	7.1
Plant food B	Grow light 1	7.2
Plant food B	Grow light 1	8.5
Plant food B	Grow light 2	5.5
Plant food B	Grow light 2	5.8
Plant food B	Grow light 2	7.6

```
# Distribution of plant growth by groups:
ggplot(df_test8, aes(x=Food_Supplement, y=Growth_Length, fill=Grow_Light)) +
  geom_boxplot() +
  labs(title="Plant growth distribution by food supplement and grow light",caption="Figure 8.1") +
  scale_fill_brewer(palette="Pastel1") +
  theme(plot.title=element_text(hjust=0.5),plot.caption=element_text(hjust=0.5, size=11)) +
  xlab(paste("Food Supplement","\n")) + ylab("Growth Length (inch)") +
  stat_summary(fun.y="mean", geom="point",shape=16, color="red",
               position = position_dodge2(width = 0.75,preserve = "single"))
```

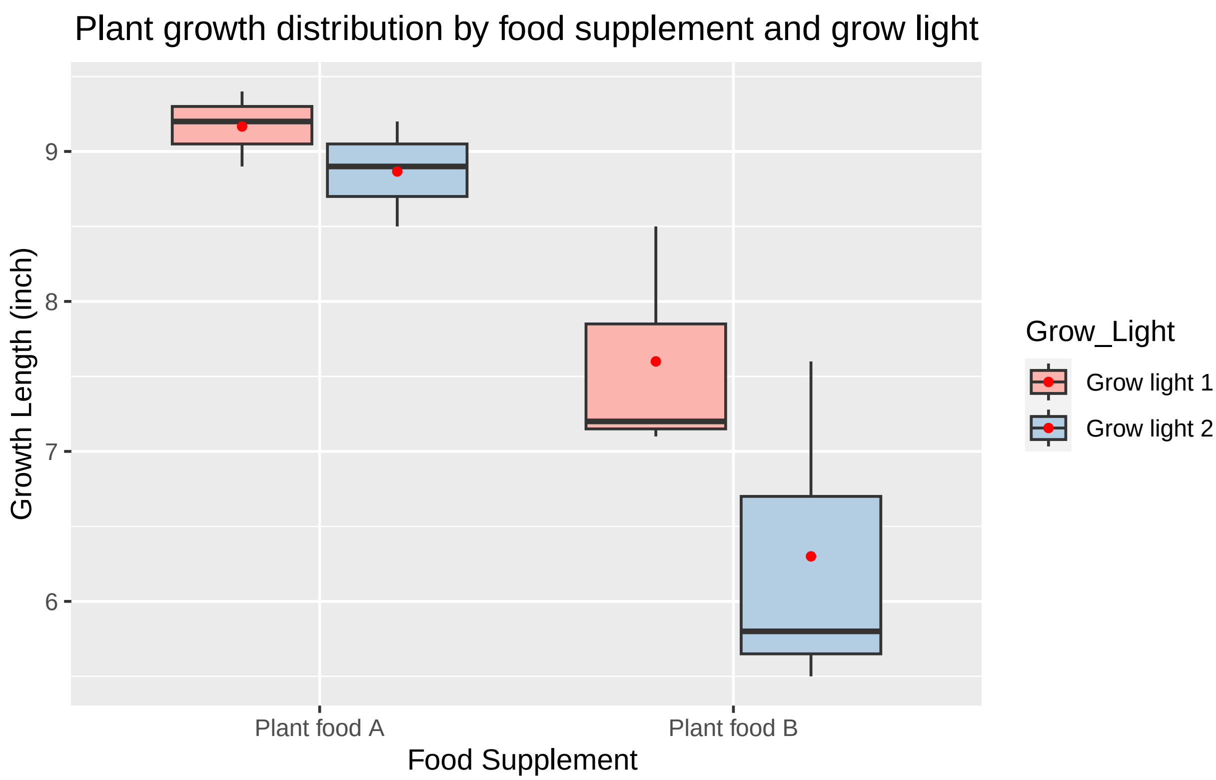

Figure 8.1

1. Step 1. State the hypothesis and identify the claim:

- The hypothesis for the interaction are ($F_{A \times B}$ test):
  - Null Hypothesis (H_0): There is no interaction effect between type of food supplement used and type of grow light used on the plant growth length
  - Alternative Hypothesis (H_1): There is an interaction effect between type of food supplement used and type of grow light used on the plant growth length
- The hypothesis for the food supplement types are (F_A test):
  - Null Hypothesis (H_0): There is no difference between the means of plant growth length for two types of food supplement
  - Alternative Hypothesis (H_1): There is a difference between the means of plant growth length for two types of food supplement
- The hypothesis for the grow light types are (F_B test):
  - Null Hypothesis (H_0): There is no difference between the means of plant growth length for two types of grow light
  - Alternative Hypothesis (H_1): There is a difference between the means of plant growth length for two types of grow light

2. Step 2-3. Find the critical values and compute the test values:

There are 2 types of food supplement and 2 types of grow light, and there are 3 data points in each group. Assume factor A and factor B represent the food supplement types and grow light types, respectively. Thus, we have a=2, b=2 and n=3. From that, we can calculate d.f.D = ab(n-1) = 2*2(3-1) = 8. At alpha = 0.05, we can determine the critical values for each test:

- For the F_A test: We have d.f.N = a-1 = 1, d.f.D = 8. Based on the F Distribution table, $F_A = 5.3177$
- For the F_B test: We have d.f.N = b-1 = 1, d.f.D = 8. Based on the F Distribution table, $F_B = 5.3177$
- For the $F_{A \times B}$ test: We have d.f.N = (a-1)*(b-1) = 1, d.f.D = 8. Based on the F Distribution table, $F_{A \times B} = 5.3177$

```
# Conduct the two-way ANOVA test
# F{A*B} test:
f_ab <- aov(Growth_Length ~ Food_Supplement * Grow_Light,data=df_test8)
summary(f_ab)
```

```
##                             Df Sum Sq Mean Sq F value  Pr(>F)
## Food_Supplement              1 12.813  12.813  24.562 0.00111 **
## Grow_Light                   1  1.920   1.920   3.681 0.09133 .
## Food_Supplement: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
```

After conducting the $F_{A \times B}$ test, we can see that the p-value of Food_Supplement:Grow_Light (0.26482) is greater than our significant level (0.05). Thus, we fail to reject the null hypothesis and conclude that there is no interaction effect between type of food supplement used and type of grow light used on the plant growth length.

```
# Interaction plot
par(mai=c(1,1,1,1))
interaction.plot(df_test8$Food_Supplement,df_test8$Grow_Light,df_test8$Growth_Length,type="b",
                 col=c(2:3),leg.bty="o",leg.bg="beige",lwd=2,pch=c(18,24), trace.label="Grow Light",
                 xlab=paste("Food Supplement","\n"),ylab="Growth Length (inch)",
                 main="Interaction plot between Food Supplement & Grow Light",
                 sub=paste("Figure 8.2","\n"),cex.axis=0.8,cex.lab=0.9,cex.sub=0.9)
```

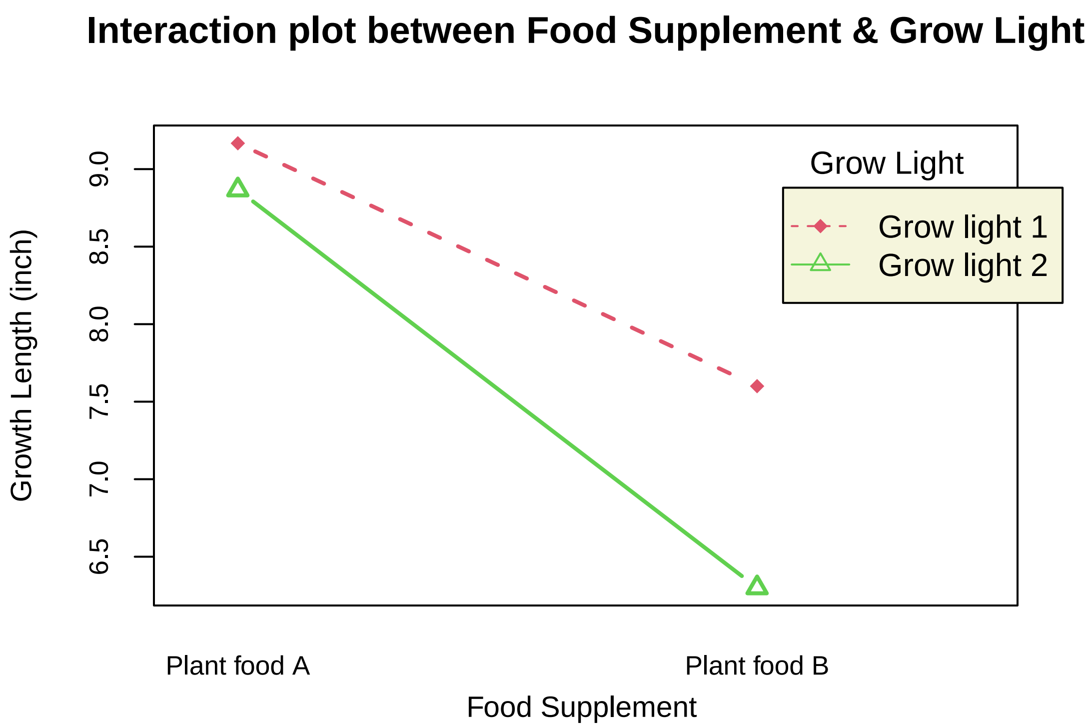

Figure 8.2

From the interaction plot above, we can see that the two lines are approximately parallel. This shows that there is no significant interaction between the 2 variables.

```
# Conduct the two-way ANOVA test
# F{A} and F{B} test:
f_ab_ind <- aov(Growth_Length ~ Food_Supplement + Grow_Light,data=df_test8)
summary(f_ab_ind)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## Food_Supplement  1 12.813  12.813   23.42 0.000921 ***
## Grow_Light       1  1.920   1.920    3.51 0.093781 .
## Residuals        9  4.923   0.547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the interaction effect is insignificant, we can continue conducting the independent tests for F_A and F_B. We see that the p-value for Food_Supplement (0.0009) is less than our alpha (0.05), but the p-value for Grow_Light (0.094) is greater than our alpha (0.05). Therefore, we can reject the null hypothesis for the F_A test but fail to reject the null hypothesis for the F_B test. The conclusions are that there is a difference between the means of plant growth length for two types of food supplement, but there is no difference between the means of plant growth length for two types of grow light.

Task 9 (On your own)

Task 9.1 (baseball.csv)

1. Import file into R

```r
df_bb <- read.csv("baseball.csv")
str(df_bb)
```

```
## 'data.frame':    1232 obs. of  15 variables:
##  $ Team        : chr  "ARI" "ATL" "BAL" "BOS" ...
##  $ League      : chr  "NL" "NL" "AL" "AL" ...
##  $ Year        : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ RS          : int  734 700 712 734 613 748 669 667 758 726 ...
##  $ RA          : int  688 600 705 806 759 676 588 845 890 670 ...
##  $ W           : int  81 94 93 69 61 85 97 68 64 88 ...
##  $ OBP         : num  0.328 0.32 0.311 0.315 0.302 0.318 0.315 0.324 0.33 0.335 ...
##  $ SLG         : num  0.418 0.389 0.417 0.415 0.378 0.422 0.411 0.381 0.436 0.422 ...
##  $ BA          : num  0.259 0.247 0.247 0.26 0.24 0.255 0.251 0.251 0.274 0.268 ...
##  $ Playoffs    : int  0 1 1 0 0 0 1 0 0 1 ...
##  $ RankSeason  : int  NA 4 5 NA NA NA 2 NA NA 6 ...
##  $ RankPlayoffs: int  NA 5 4 NA NA NA 4 NA NA 2 ...
##  $ G           : int  162 162 162 162 162 162 162 162 162 162 ...
##  $ OOBP        : num  0.317 0.306 0.315 0.331 0.335 0.319 0.305 0.336 0.357 0.314 ...
##  $ OSLG        : num  0.415 0.378 0.403 0.428 0.424 0.405 0.39 0.43 0.47 0.402 ...
```

2. Perform EDA on the data set

```r
df_bb_num <- select_if(df_bb,is.numeric)
df_bb_char <- select_if(df_bb,is.character)

# Show descriptive statistics of the numerical values
num_stats <- describe(df_bb_num)[,c('n','mean','median','sd','min','max')]
num_stats <- as.data.frame(num_stats)
num_stats$na_count <- nrow(df_bb) - num_stats$n
num_stats$mean <- round(num_stats$mean,2)
num_stats$median <- round(num_stats$median,2)
num_stats$sd <- round(num_stats$sd,2)
num_stats$min <- round(num_stats$min,2)
num_stats$max <- round(num_stats$max,2)

knitr::kable(num_stats, "simple",
             caption="Table 9.1. Descriptive statistics of the data set numerical values") %>%
  kable_styling(position = "center")
```

Table 9.1. Descriptive statistics of the data set numerical values

	n	mean	median	sd	min	max	na_count
Year	1232	1988.96	1989.00	14.82	1962.00	2012.00	0
RS	1232	715.08	711.00	91.53	463.00	1009.00	0
RA	1232	715.08	709.00	93.08	472.00	1103.00	0
W	1232	80.90	81.00	11.46	40.00	116.00	0
OBP	1232	0.33	0.33	0.02	0.28	0.37	0
SLG	1232	0.40	0.40	0.03	0.30	0.49	0
BA	1232	0.26	0.26	0.01	0.21	0.29	0
Playoffs	1232	0.20	0.00	0.40	0.00	1.00	0
RankSeason	244	3.12	3.00	1.74	1.00	8.00	988
RankPlayoffs	244	2.72	3.00	1.10	1.00	5.00	988

	n	mean	median	sd	min	max	na_count
G	1232	161.92	162.00	0.62	158.00	165.00	0
OOBP	420	0.33	0.33	0.02	0.29	0.38	812
OSLG	420	0.42	0.42	0.03	0.35	0.50	812

```
# Show descriptive statistics of the categorical values
char_uniq <- as.matrix(lengths(lapply(df_bb_char,unique)))
char_n <- describe(df_bb_char)[,c('n')]
char_stats <- cbind(char_uniq,char_n)
colnames(char_stats) <- c("Unique","Total Count")

knitr::kable(char_stats, "simple",
             caption="Table 9.2. Descriptive statistics of the data set categorical values") %>%
  kable_styling(position = "center")
```

Table 9.2. Descriptive statistics of the data set categorical values

	Unique	Total Count
Team	39	1232
League	2	1232

```
# Wins by top 5 teams by league
df_bb_wins <- df_bb[,c("Team","League","W")]
df_bb_wins1 <-
  df_bb_wins %>%
  group_by(Team,League) %>%
  summarise(total_w=sum(W)) %>%
  arrange(desc(total_w))

df_bb_wins1 <- head(df_bb_wins1,10) %>% arrange(League)

ggplot(df_bb_wins1, aes(x = League, y = total_w, fill=Team)) +
  geom_bar(stat="identity",position="dodge") +
  geom_text(aes(label=total_w), position=position_dodge(width=0.9), vjust=-0.5,size=3) +
  labs(title="Top 5 teams with the highest games won by league",caption="Figure 9.1") +
  theme(plot.title=element_text(hjust=0.5),plot.caption=element_text(hjust=0.5, size=11)) +
  xlab(paste("League","\n")) + ylab("Total Wins")
```

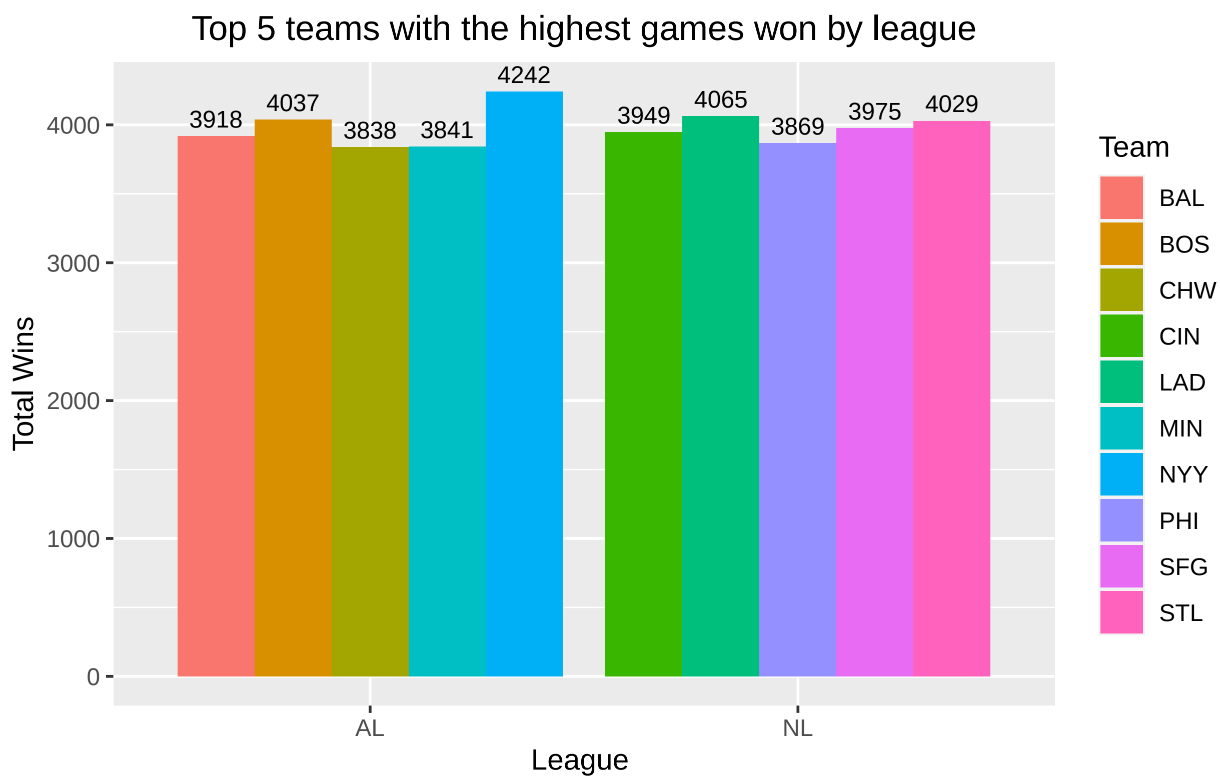

Figure 9.1

The baseball data set contains 1232 rows and 15 columns. Within the 15 data fields, there are 13 numerical variables and 2 categorical variables. There are a total of 39 teams in the data set, within which, 19 teams are in the NL league and 20 teams are in the AL league. Table 9.1 and 9.2 shows the descriptive statistics of the variables given in the data set. The bar chart above helps visualize the top 5 performing teams in each league. The 5 teams with the highest wins in AL league are BAL, BOS, CHW, MIN, NYY, and the 5 teams with the highest wins in NL league are CIN, LAD, PHI, SFG, STL.

3. Chi-square goodness-of-fit test

```
# Data preparation
df_bb$Decade <- df_bb$Year - (df_bb$Year%%10)
wins_decade <- df_bb %>%
  group_by(Decade) %>%
  summarise(wins=sum(W)) %>%
  as.tibble()
wins_decade
```

```
## # 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
```

```r
alpha9 <- 0.05
expected9 <- c(1/6,1/6,1/6,1/6,1/6,1/6)
observed9 <- wins_decade$wins
```

- Step 1. State the hypothesis and identify the claim:
  - Null Hypothesis (H_0): There is no difference in the number of wins by decade
  - Alternative Hypothesis (H_1): There is a difference in the number of wins by decade

- Step 2-3. Find the critical values and compute the test values:

At $\alpha = 0.05$ and degree of freedom $= 5$, the critical value is 11.07 based on the Chi-Square distribution table. We begin performing the chi-square test:

```r
# Conduct chi-square test:
test9 <- chisq.test(observed9, p=expected9, correct=F)
test9
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed9
## X-squared = 9989.5, df = 5, p-value < 2.2e-16
```

- Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

In this test, since the test p-value (`0e+00`) is smaller than α (`0.05`), we can reject the null hypothesis and conclude that there is significant difference in the number of wins by decade.

Task 9.2 (crop_data.csv)

1. Import file into R

```r
df_crop <- read.csv("crop_data.csv")
str(df_crop)
```

```
## 'data.frame':    96 obs. of  4 variables:
##  $ density   : int  1 2 1 2 1 2 1 2 1 2 ...
##  $ block     : int  1 2 3 4 1 2 3 4 1 2 ...
##  $ fertilizer: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yield     : num  177 178 176 178 177 ...
```

2. Conduct the two-way ANOVA test

```r
df_crop$density <- as.factor(df_crop$density)
df_crop$fertilizer <- as.factor(df_crop$fertilizer)

table(df_crop$density, df_crop$fertilizer)
```

```
##
##      1  2  3
##   1 16 16 16
##   2 16 16 16
```

Step 1. State the hypothesis and identify the claim:

- The hypothesis for the interaction are ($F_{A\times B}$ test):
  - Null Hypothesis (H_0): There is no interaction effect between type of density and fertilizer on the crop yield
  - Alternative Hypothesis (H_1): There is an interaction effect between type of density and fertilizer on the crop yield
- The hypothesis for the density types are (F_A test):
  - Null Hypothesis (H_0): There is no difference between the means of crop yield for two types of density
  - Alternative Hypothesis (H_1): There is a difference between the means of crop yield for two types of density
- The hypothesis for the fertilizer types are (F_B test):
  - Null Hypothesis (H_0): There is no difference between the means of crop yield for three types of fertilizer
  - Alternative Hypothesis (H_1): There is a difference between the means of crop yield for three types of fertilizer

Step 2-3. Find the critical values and compute the test values:

There are 2 types of density and 3 types of fertilizer, and there are 16 data points in each group. Assume factor A and factor B represent the density types and fertilizer types, respectively. Thus, we have a=2, b=3 and n=16. From that, we can calculate d.f.D = ab(n-1) = 2*3(16-1) = 90. At alpha = 0.05, we can determine the critical values for each test:

- For the F_A test: We have d.f.N = a-1 = 1, d.f.D = 90. Based on the F Distribution table, F_A ~ 0
- For the F_B test: We have d.f.N = b-1 = 2, d.f.D = 90. Based on the F Distribution table, F_B ~ 0.05
- For the $F_{A\times B}$ test: We have d.f.N = (a-1)*(b-1) = 2, d.f.D = 90. Based on the F Distribution table, $F_{A\times B}$ ~ 0.05

Step 4-5. Make the decision to accept/reject the null hypothesis and summarize the result

```
# Conduct the two-way ANOVA test
# F{A*B} test:
f_ab2 <- aov(yield ~ density * fertilizer,data=df_crop)
summary(f_ab2)
```

```
##                    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
```

After conducting the $F_{A\times B}$ test, we can see that the p-value of density:fertilizer (0.5325) is greater than our significant level (0.05). Thus, we fail to reject the null hypothesis and conclude that there is no interaction effect between type of density and fertilizer on the crop yield.

```
# Conduct the two-way ANOVA test
# F{A} and F{B} test:
f_ab_ind2 <- aov(yield ~ density + fertilizer,data=df_crop)
summary(f_ab_ind2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## density       1  5.122   5.122  15.316 0.000174 ***
## fertilizer    2  6.068   3.034   9.073 0.000253 ***
## Residuals    92 30.765   0.334                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the interaction effect is insignificant, we can continue conducting the independent tests for F_A and F_B. We see that the p-value for density (0.000174) is less than our alpha (0.05), and the p-value for fertilizer (0.000253) is also smaller than our alpha (0.05). Therefore, we can reject the null hypothesis for both the F_A and F_B tests and conclude that there is a difference between the means of crop yield for two types of density, and there is a difference between the means of crop yield for three types of fertilizer.

III. CONCLUSION SECTION

This projects help to remind me how to properly conduct the Chi-square test as well as the ANOVA test for different purposes. In future projects, I hope to be able to apply this knowledge and perform the tests myself on other data sets.