Problem Set 2, Winter 2024

Michael Ghattas

```r
# Load necessary libraries
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

Donna is the owner of a boutique doughnut shop. Because many of her customers are conscious of their fat intake but want the flavor of fried doughnuts, she decided to develop a doughnut recipe that minimizes the amount of fat that the doughnuts absorb from the fat in which the doughnuts are fried.

She conducted a factorial experiment that had a similar procedures as Lowe (1935). Like Lowe, she used four types of fats (fat_type). She also used three types of flour (flour_type): all-purpose flour, whole wheat flour, and gluten-free flour. For each combination of fat type and flour type, she cooked six identical batches of doughnuts. Each batch contained 24 doughnuts, and the total fat (in grams) absorbed by the doughnuts in each batch was recorded (sim_tot_fat).

Question 1 - 5 points

You may need to process your data before you begin your analysis. Specifically, you will need to make sure that the variable type is set to 'factor' for both of your grouping variables and 'num' for your outcome variable.

```r
# Load the data
doughnuts.factorial <- read.csv("doughnutsfactorial.csv", header=TRUE, sep=",")
head(doughnuts.factorial)
```

```
##   fat_type flour_type sim_tot_fat
## 1   Canola         ap          78
## 2   Canola         ap          71
## 3   Canola         ap          80
## 4   Canola         ap          88
## 5   Canola         ap          62
## 6   Canola         ap          72
```

Like in Problem Set 1, please create two new variables in the doughnuts.factorial data set. The first new variable will be called fat_type_factor and will contain the same values as in the fat_type variable but will have a variable type of factor. The second new variable will be called flour_type_factor and will contain the same values as in the flour_type variable but will also have a variable type of factor.

```r
# Create new factor variables
doughnuts.factorial$fat_type_factor <- as.factor(doughnuts.factorial$fat_type)
doughnuts.factorial$flour_type_factor <- as.factor(doughnuts.factorial$flour_type)
```

Check your work by running the following code chunk. Be sure that fat_type_factor and flour_type_factor are factor-type variables before you complete the rest of the problem set.

```
# Check structure of the data
str(doughnuts.factorial)
```

```
## 'data.frame':    72 obs. of  5 variables:
##  $ fat_type         : chr  "Canola" "Canola" "Canola" "Canola" ...
##  $ flour_type       : chr  "ap" "ap" "ap" "ap" ...
##  $ sim_tot_fat      : int  78 71 80 88 62 72 78 75 89 74 ...
##  $ fat_type_factor  : Factor w/ 4 levels "Canola","Peanut",..: 1 1 1 1 1 1 3 3 3 3 ...
##  $ flour_type_factor: Factor w/ 3 levels "ap","gf","ww": 1 1 1 1 1 1 1 1 1 1 ...
```

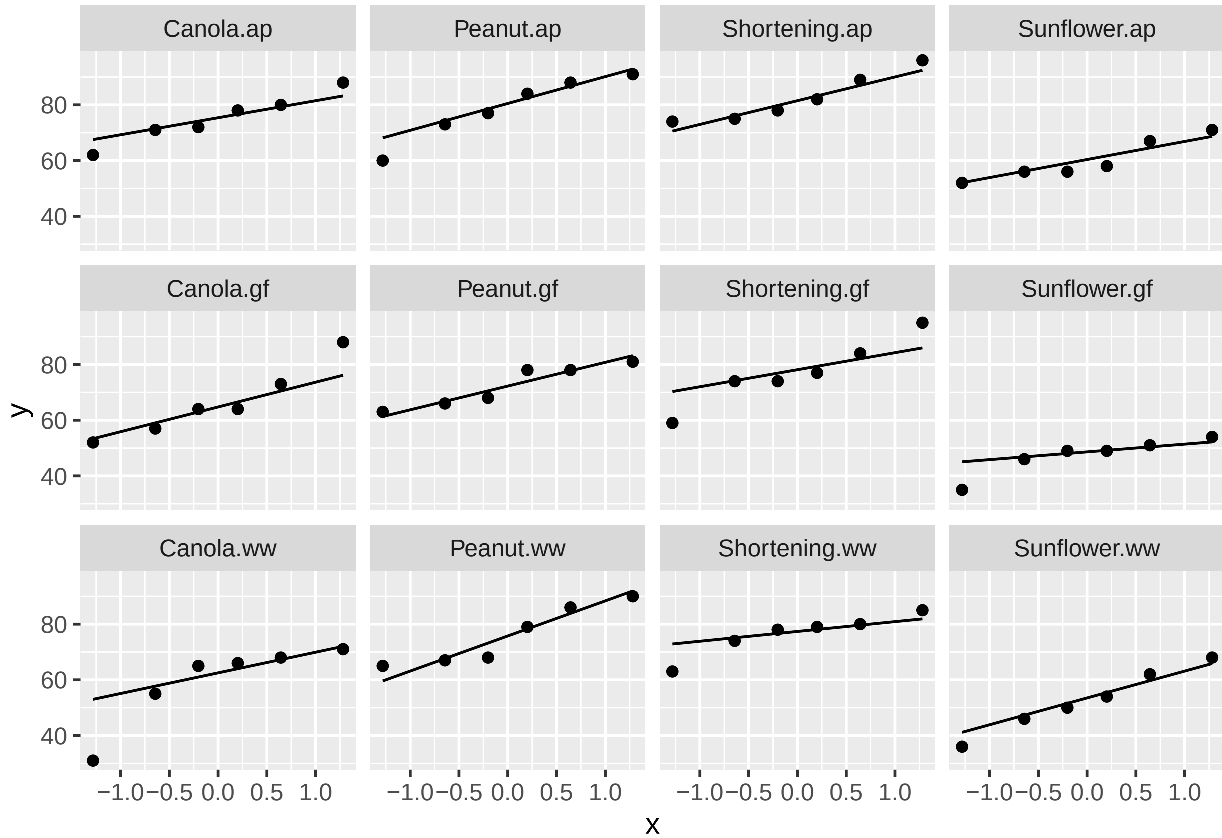
Question 2 - 5 points

Provide a visual assessment and a quantitative assessment for the assumption of *normality* for each cell. Hint: Remember that a cell contains the observations that make up a particular combination of two factors. Therefore, there will be as many graphs/quantitative tests as are unique combinations of flour and fat types.

Code for your visual assessment of normality

```
# Create a combination of the fat and flour factors for the cell-wise analysis
doughnuts.factorial$cell <- interaction(doughnuts.factorial$fat_type_factor,
                                        doughnuts.factorial$flour_type_factor)

# Loop through each combination and create Q-Q plots
ggplot(doughnuts.factorial, aes(sample = sim_tot_fat)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~cell)
```



Code for your quantitative assessment of normality

```r
# Quantitative assessment of normality using Shapiro-Wilk test
shapiro_test_results <- by(doughnuts.factorial$sim_tot_fat,
                           doughnuts.factorial$cell, shapiro.test)
shapiro_test_results
```

```
## doughnuts.factorial$cell: Canola.ap
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.98525, p-value = 0.9745
##
## ------------------------------------------------------------
## doughnuts.factorial$cell: Peanut.ap
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.942, p-value = 0.6754
##
## ------------------------------------------------------------
## doughnuts.factorial$cell: Shortening.ap
##
##  Shapiro-Wilk normality test
```

```
##
## data:  dd[x, ]
## W = 0.90965, p-value = 0.4341
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Sunflower.ap
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.88761, p-value = 0.3059
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Canola.gf
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9346, p-value = 0.6161
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Peanut.gf
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.87763, p-value = 0.2583
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Shortening.gf
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.96166, p-value = 0.8325
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Sunflower.gf
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.85355, p-value = 0.1681
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Canola.ww
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.78216, p-value = 0.04037
##
## ---------------------------------------------------------------
## doughnuts.factorial$cell: Peanut.ww
```

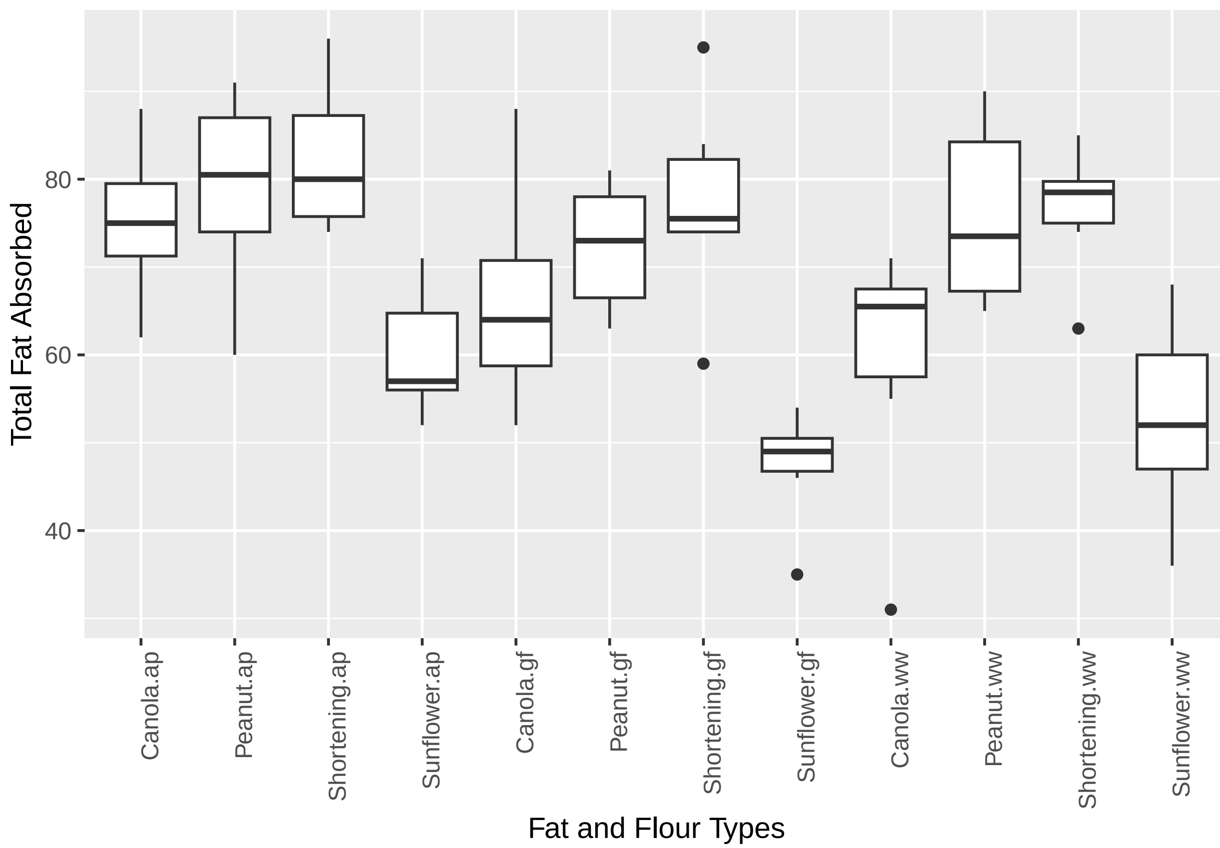
```
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.87732, p-value = 0.257
##
## ------------------------------------------------------------
## doughnuts.factorial$cell: Shortening.ww
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.9001, p-value = 0.3745
##
## ------------------------------------------------------------
## doughnuts.factorial$cell: Sunflower.ww
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.98901, p-value = 0.9866
```

Question 3 - 5 points

Provide a visual assessment and a quantitative assessment for the assumption of *equality of variances* for each cell.

Code for your visual assessment of equality of variances

```
# Visual assessment of equality of variances using boxplots
ggplot(doughnuts.factorial, aes(x = interaction(fat_type_factor, flour_type_factor),
                                y = sim_tot_fat)) + geom_boxplot() + xlab("Fat and Flour Types") +
  ylab("Total Fat Absorbed") + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Code for your quantitative assessment of equality of variances

```
# Quantitative assessment of equality of variances using Levene's test
leveneTest(sim_tot_fat ~ fat_type_factor * flour_type_factor, data = doughnuts.factorial)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 11  0.4064 0.9478
##       60
```

Question 4 - 10 points

Before conducting your two-way ANOVA with an interaction, start by conducting one-way ANOVAs for each of your factors. You wouldn't do this in practice - you would just conduct the two-way ANOVA - but you'll do it here to allow you to make some comparisons between one-way ANOVA and two-way ANOVA with an interaction in Question 7. You do not need to interpret these ANOVAs, but be sure to display the output in your knitted document.

Your one-way ANOVA for testing if the means in total fat (sim_total_fat) are the same across fat types:

```
# One-way ANOVA for fat type
fat.aov <- aov(sim_tot_fat ~ fat_type_factor, data = doughnuts.factorial)
summary(fat.aov)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
```

```
## fat_type_factor  3   6967  2322.5   20.17 1.86e-09 ***
## Residuals       68   7831   115.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your one-way ANOVA for testing if the means in total fat (sim_total_fat) are the same across flour types:

```r
# One-way ANOVA for flour type
flour.aov <- aov(sim_tot_fat ~ flour_type_factor, data = doughnuts.factorial)
summary(flour.aov)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## flour_type_factor  2   1063   531.3   2.669 0.0765 .
## Residuals         69  13736   199.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5 - 10 points

Conduct a two-way ANOVA with an interaction between fat type and flour type. Use sim_total_fat as the outcome and fat_type_factor and flour_type_factor as the grouping variables. Please be sure to display your ANOVA results using the summary() function.

```r
# Two-way ANOVA with interaction between fat and flour type
fat_flour_int.aov <- aov(sim_tot_fat ~ fat_type_factor * flour_type_factor, data = doughnuts.factorial)
summary(fat_flour_int.aov)
```

```
##                                   Df Sum Sq Mean Sq F value   Pr(>F)
## fat_type_factor                    3   6967  2322.5  21.976 1.01e-09 ***
## flour_type_factor                  2   1063   531.3   5.028  0.00958 **
## fat_type_factor:flour_type_factor  6    427    71.2   0.674  0.67095
## Residuals                         60   6341   105.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note:

Since the interaction between fat type and flour type is not significant ($p = 0.67095$), you can interpret the main effects in a straightforward manner without worrying about how one factor may influence the other. However, since both main effects for fat type ($p = 1.01e\text{-}09$) and flour type ($p = 0.00958$) are significant, it is appropriate to conduct post hoc tests to understand which specific levels of each factor are significantly different from one another.

```r
# Tukey's HSD for fat type
TukeyHSD(fat_flour_int.aov, "fat_type_factor")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sim_tot_fat ~ fat_type_factor * flour_type_factor, data = doughnuts.factorial)
```

```
##
## $fat_type_factor
##                              diff         lwr       upr     p adj
## Peanut-Canola            8.722222  -0.3329092  17.77735 0.0630718
## Shortening-Canola       11.722222   2.6670908  20.77735 0.0060539
## Sunflower-Canola       -13.611111 -22.6662425  -4.55598 0.0010846
## Shortening-Peanut        3.000000  -6.0551314  12.05513 0.8174979
## Sunflower-Peanut       -22.333333 -31.3884648 -13.27820 0.0000001
## Sunflower-Shortening -25.333333 -34.3884648 -16.27820 0.0000000
```

```r
# Tukey's HSD for flour type
TukeyHSD(fat_flour_int.aov, "flour_type_factor")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sim_tot_fat ~ fat_type_factor * flour_type_factor, data = doughnuts.factorial)
##
## $flour_type_factor
##             diff        lwr        upr     p adj
## gf-ap -8.2916667 -15.423477 -1.1598561 0.0188546
## ww-ap -8.0000000 -15.131811 -0.8681894 0.0243440
## ww-gf  0.2916667  -6.840144  7.4234773 0.9946891
```

Question 6 - 10 points

Be sure to have completed the two-way ANOVA with an interaction analysis before answering the following four questions.

Main effects hypotheses - two questions to answer

A) Please select the statement that is the best interpretation of the p-value associated with the main effect of fat type.

Statement 1: I reject the null hypothesis and conclude that at least one fat type has a statistically significantly different mean fat absorption than the other groups.

Statement 2: I fail to reject the null hypothesis and conclude that there is no statistically significant difference in the mean amount of fat absorbed among fat types.

Your answer here: From the two-way ANOVA output, the p-value for the main effect of fat_type_factor is very small ($p = 1.01e-09$), which is less than 0.05. I reject the null hypothesis and conclude that at least one fat type has a statistically significantly different mean fat absorption than the other groups. (Statement 1)

B) Please select the statement that is the best interpretation of the p-value associated with the main effect of flour type.

Statement 1: I reject the null hypothesis and conclude that at least one flour type has a statistically significantly different mean fat absorption than the other groups.

Statement 2: I fail to reject the null hypothesis and conclude that there is no statistically significant difference in the mean amount of fat absorbed among flour types.

Your answer here: Interpretation: The p-value for the main effect of flour_type_factor is also significant (p = 0.00958), which is less than 0.05. I reject the null hypothesis and conclude that at least one flour type has a statistically significantly different mean fat absorption than the other groups. (Statement 1)

Interaction hypothesis - 2 questions to answer

C) Please select the statement that is the best interpretation of the p-value associated with the interaction between fat type and flour type.

Statement 1: The interaction between fat type and flour type is statistically significant.

Statement 2: The interaction between fat type and flour type is not statistically significant.

Your answer here: The p-value for the interaction term (fat_type_factor:flour_type_factor) is not significant (p = 0.67095), which is greater than 0.05. The interaction between fat type and flour type is not statistically significant. (Statement 2)

D) Based on your response to the previous question about the interaction, can you interpret the main effects in a straightforward fashion? Put differently, is it justifiable to make a conclusion about the effect of fat type while ignoring the effect of flour type (and vice versa)?

Your answer here (yes or no): Yes, it is justifiable to interpret the main effects of fat type and flour type in a straightforward fashion.

This is because the interaction between fat type and flour type is not statistically significant (p = 0.67095), meaning that the effect of one factor (fat type) does not depend on the level of the other factor (flour type). Therefore, you can evaluate the main effects of fat type and flour type independently, and conclusions can be made about each factor without accounting for their interaction. it is important to mention the post hoc results when interpreting the significant main effects of fat type and flour type. While the two-way ANOVA indicates that there is a significant main effect for both factors, the post hoc tests provide detailed information on which specific levels (e.g., which types of fats or flours) differ from one another. This allows for a more comprehensive interpretation of the results.

Question 7 - 5 points

You conducted 2 one-way ANOVAs in Question 4 and 1 two-way ANOVA with an interaction in Question 5. In this question, you will answer four questions comparing the results of these analyses.

A) Look at the lines for *fat_type_factor* in both the one-way ANOVA with fat_type_factor (fat.aov in Question 4) used as the grouping variable and the two-way ANOVA with an interaction (fat_flour_int.aov in Question 5). Is there any difference in the degrees of freedom or the sums of squares between these lines?

Your answer here (yes/no): No. The degrees of freedom (3) and sums of squares (6967) for fat_type_factor are the same in both the one-way ANOVA and the two-way ANOVA. This is expected because the sums of squares for the main effect of fat type do not change whether or not other factors (such as flour type) are included in the model.

B) Looking at the same lines as the previous question, is there a difference between the F test statistic or the p-values?

Your answer here (yes/no): No. The F test statistic (20.17) and p-value (1.86e-09) for fat_type_factor are the same in both the one-way and two-way ANOVAs. This confirms that adding the additional factor (flour type) did not affect the results for the main effect of fat type.

C) Look at the lines for *flour_type_factor* in both the one-way ANOVA with flour_type_factor (flour.aov in Question 4) used as the grouping variable and the two-way ANOVA with an interaction (fat_flour_int.aov in Question 5). Is there any difference in the degrees of freedom and the sums of squares between these lines?

Your answer here (yes/no): No. The degrees of freedom (2) and sums of squares (1063) for flour_type_factor are the same in both the one-way ANOVA and the two-way ANOVA. As with fat type, the sums of squares for the main effect of flour type remain the same regardless of the inclusion of another factor in the model.

D) Looking at the same lines as the previous question, is there a difference between the F test statistic or the p-values?

Your answer here (yes/no): Yes. There is a difference in the F test statistic and p-values for flour_type_factor between the one-way ANOVA and the two-way ANOVA. - In the one-way ANOVA, the F value is 2.669 with a p-value of 0.0765, which is not significant at the 0.05 level. - In the two-way ANOVA, the F value is 5.028 with a p-value of 0.00958, which is significant at the 0.05 level. This shows that including the fat type factor in the two-way ANOVA improved the model's ability to detect a significant effect for flour type.