Problem Set 1, Winter 2024

Michael Ghattas

```r
# Load necessary libraries
library(ggplot2)
library(ggpubr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(car) # For Levene's Test
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(stats) # For ANOVA and Post-hoc testing
```

CONTEXT - DOUGHNUTS DATA

This data set was derived from an experiment conducted by Lowe (1935) (obtained from Snedecor & Cochran, 1989).

Lowe wanted to learn more about how much fat doughnuts absorb when cooked in different kinds of fat. He tested four kinds of fats (fat_type): canola oil, vegetable shortening, peanut oil, and sunflower oil. He cooked six identical batches of doughnuts using each type of fat. Each batch contained 24 doughnuts. The outcome of interest was the total amount of fat (in grams) absorbed by each batch of doughnuts (total_fat).

Run the code chunk below to read the data into memory and change the type of a variable.

```r
# Load the doughnuts dataset
doughnuts <- read.csv("doughnuts-1.csv")

# Convert fat_type to a factor
doughnuts$fat_type_factor <- as.factor(doughnuts$fat_type)
```

Run the code chunk below to confirm that the variables are of the appropriate type. The str() function is useful for checking four things: The number of rows ("observations"), the number of variables, the names of the variables, and the type of the variables.

The str() function should confirm all of these for you about this data set. This data set should have 24 rows and three variables. One of these variables, fat_type, should be a character-type ("chr") variable. Another of those variables, total_fat, should be an integer-type ("int") variable. The remaining variable, fat_type_factor, should be a factor-type variable with four levels.

```r
# Check the structure of the dataset
str(doughnuts)
```

```
## 'data.frame':    24 obs. of  3 variables:
##  $ fat_type       : chr  "Canola" "Canola" "Canola" "Canola" ...
##  $ total_fat      : int  64 72 68 77 56 95 78 91 97 82 ...
##  $ fat_type_factor: Factor w/ 4 levels "Canola","Peanut",..: 1 1 1 1 1 1 3 3 3 3 ...
```

Question 1 - 10 points

Compute the mean and standard deviation for each fat type. Hint: You have sample data, not population data; this matters for computing the standard deviation.

```r
# Group by fat type and compute the mean and sample standard deviation for total fat absorbed
fat_type_stats <- doughnuts %>%
  group_by(fat_type) %>%
  summarise(mean_fat = mean(total_fat), sd_fat = sd(total_fat, na.rm = TRUE))

# Display the results
fat_type_stats
```

```
## # A tibble: 4 x 3
##   fat_type   mean_fat sd_fat
##   <chr>         <dbl>  <dbl>
## 1 Canola           72  13.3
## 2 Peanut           76   9.88
## 3 Shortening       85   7.77
## 4 Sunflower        62   8.22
```

Canola mean and SD (your answer here): Mean = 72g, SD = 13.342

Shortening mean and SD (your answer here): Mean = 85g, SD = 7.772

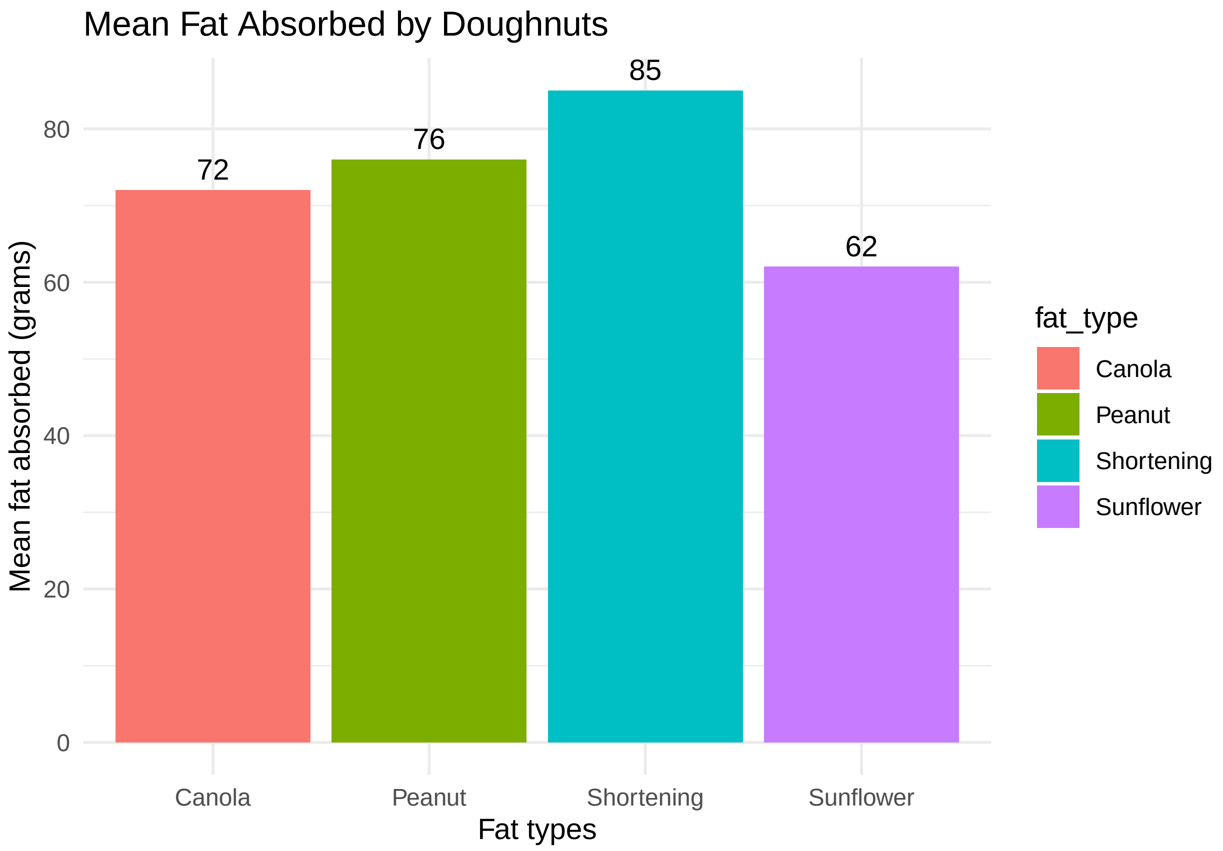
Peanut mean and SD (your answer here): Mean = 76g, SD = 9.879

Sunflower mean and SD (your answer here): Mean = 62g, SD = 8.222

Next, create a bar plot to visualize the differences in the means. There are some examples of what a bar plot is at this website: https://statisticsglobe.com/barplot-in-r. Please label your Y axis "Mean fat absorbed (grams)" and your X axis "Fat types". Please also have sub-labels for each bar that match the appropriate fat type (canola, shortening, peanut, and sunflower).

Although many bar plots also include a visualization of the variability within groups (e.g., standard error bars), visualizing the variability is not necessary for full credit on this question.

```r
# Create a bar plot of the mean fat absorbed by each fat type
ggplot(fat_type_stats, aes(x = fat_type, y = mean_fat, fill = fat_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Fat Absorbed by Doughnuts",
       x = "Fat types",
       y = "Mean fat absorbed (grams)") +
  theme_minimal() +
  geom_text(aes(label = round(mean_fat, 1)), vjust = -0.5)
```

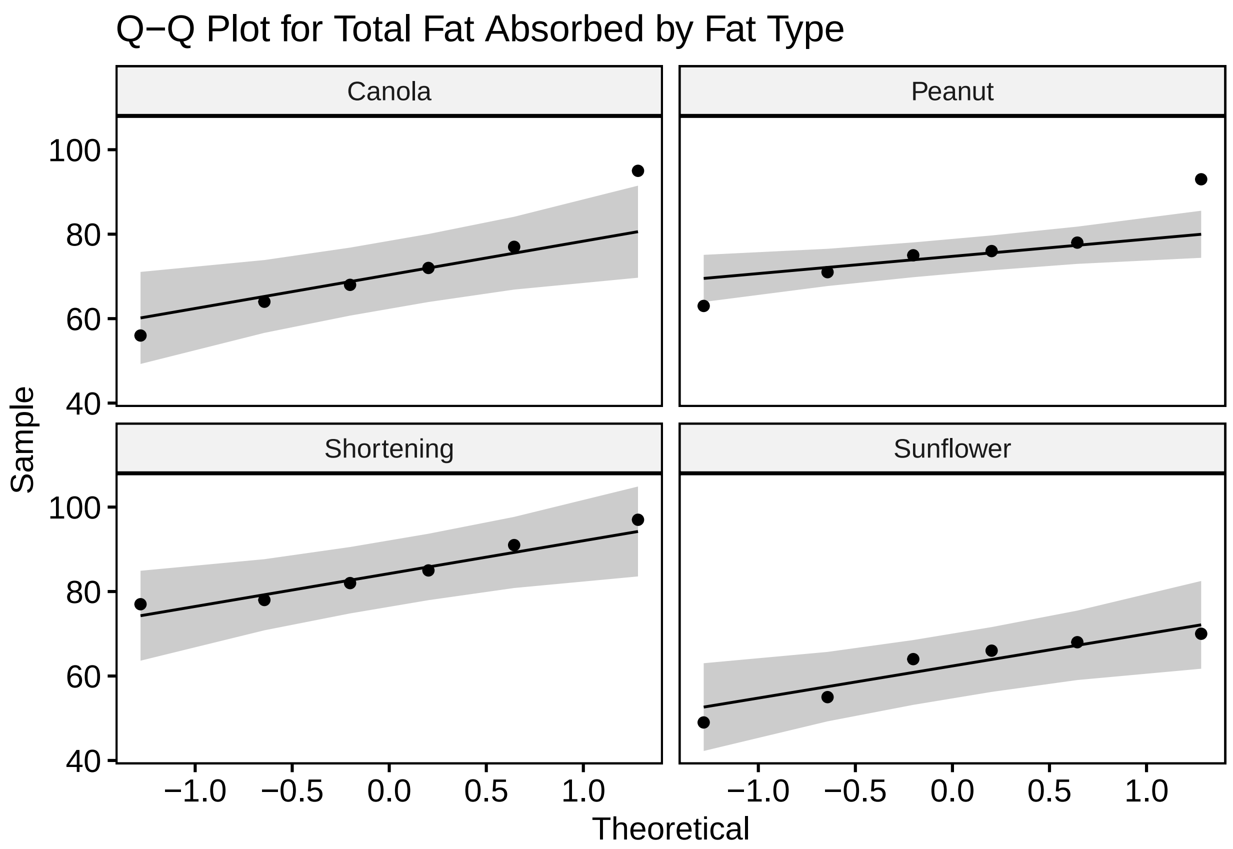


Question 2 - 10 points

You will conduct a one-way ANOVA, but let's assess our assumptions first. Assess the assumption of *normality* visually and quantitatively and comment on how well the data met this assumption.

First, assess this assumption visually:

```r
# Visual assessment of normality using a Q-Q plot for each fat type
ggqqplot(doughnuts, x = "total_fat", facet.by = "fat_type_factor",
         title = "Q-Q Plot for Total Fat Absorbed by Fat Type")
```



Next, assess this assumption quantitatively:

```r
# Quantitative assessment of normality using the Shapiro-Wilk test for each fat type
shapiro_results <- doughnuts %>%
  group_by(fat_type_factor) %>%
  summarise(p_value = shapiro.test(total_fat)$p.value)

# Display the results of the Shapiro-Wilk test
shapiro.test(doughnuts$total_fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  doughnuts$total_fat
## W = 0.97525, p-value = 0.7949
```

```r
shapiro_results
```

```
## # A tibble: 4 x 2
##   fat_type_factor p_value
##   <fct>             <dbl>
## 1 Canola            0.741
## 2 Peanut            0.607
## 3 Shortening        0.593
## 4 Sunflower         0.310
```

Finally, answer the three questions below:

A) What type of visualization did you use to assess the assumption of normality visually?

Your answer here: I used Q-Q plots to visually assess the assumption of normality for each fat type. The Q-Q plots compare the observed quantiles of the data to the theoretical quantiles from a normal distribution.

B) What type of quantitative test did you conduct to assess the assumption of normality quantitatively?

Your answer here: I conducted the Shapiro-Wilk test for each fat type to quantitatively assess the assumption of normality.

C) Based on the results of your quantitative assessment, do you conclude that your data meet the assumption of normality?

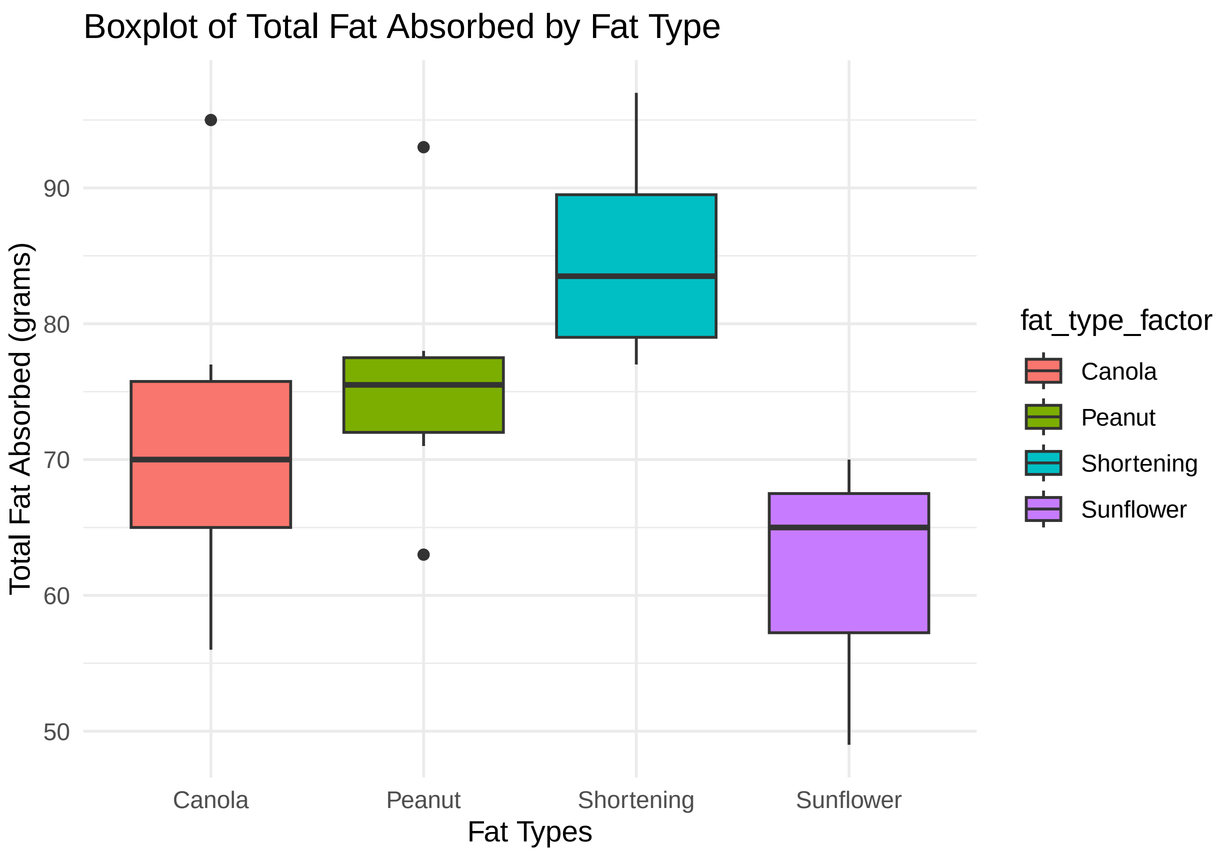
Your answer here: Based on the p-values from the Shapiro-Wilk test (Canola: 0.741, Peanut: 0.607, Shortening: 0.593, Sunflower: 0.310), all p-values are above 0.05, indicating that the data do not significantly deviate from a normal distribution. Therefore, we can conclude that the assumption of normality is reasonably met for all fat types. The visual Q-Q plots confirm the same. Thus, the data meet the assumption of normality across the groups.

Question 3 - 10 points

Assess the assumption of *equality of variances* visually and quantitatively and comment on how well the data met this assumption.

First, assess this assumption visually:

```r
# Visual assessment of equality of variances using a boxplot
ggplot(doughnuts, aes(x = fat_type_factor, y = total_fat, fill = fat_type_factor)) +
  geom_boxplot() +
  labs(title = "Boxplot of Total Fat Absorbed by Fat Type",
       x = "Fat Types", y = "Total Fat Absorbed (grams)") +
  theme_minimal()
```



Next, assess this assumption quantitatively:

```
# Quantitative assessment of equality of variances using Levene's Test
leveneTest(total_fat ~ fat_type_factor, data = doughnuts)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.3434 0.7942
##       20
```

Finally, answer the three questions below:

A) What type of visualization did you use to assess the assumption of equal variances across groups visually?

Your answer here: I used a boxplot to visually assess the assumption of equal variances across the different fat types. The box-plot shows the spread of fat absorption for each fat type, and differences in the spread of the boxes can indicate differences in variance.

B) What type of quantitative test did you conduct to assess the assumption of equal variances across groups quantitatively?

Your answer here: I conducted Levene's Test to assess the equality of variances quantitatively.

C) Based on the results of your quantitative assessment, do you conclude that your data meet the assumption of equal variances across groups?

Your answer here: Based on the visual assessment from the box-plot, the spread of values does not appear drastically different between the groups. Additionally, since the p-value (0.7942) is much greater than 0.05, we fail to reject the null hypothesis of Levene’s Test. This means there is no significant evidence to suggest that the variances are unequal across the different fat types. Thus, the data meet the assumption of equal variances across the groups.

Question 4 - 10 points

You will now conduct a one-way ANOVA analysis using total_fat as the outcome and fat_type_factor as the grouping variable.

First, conduct the analysis and display the result:

```
# Conduct the one-way ANOVA using total_fat as the outcome and fat_type_factor as the grouping variable
doughnuts.aov <- aov(total_fat ~ fat_type_factor, data = doughnuts)

# Display the results of the ANOVA
summary(doughnuts.aov)
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## fat_type_factor  3   1636   545.5   5.406 0.00688 **
## Residuals       20   2018   100.9
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Second, answer the three questions below:

A) What is the null hypothesis being tested in this one-way ANOVA analysis?

Your answer here: The null hypothesis is that the mean amount of fat absorbed is the same across all fat types. In other words, there is no significant difference in the mean fat absorption among the four fat types (Canola, Peanut, Shortening, and Sunflower).

B) Based on the results of your analysis, do you reject or fail to reject the null hypothesis?

Your answer here: Since the p-value = 0.00688 is less than the significance level of 0.05, we reject the null hypothesis. This means that there is a statistically significant difference in the mean fat absorption among the different fat types.

C) Which of the three statements (1, 2, or 3) is an appropriate conclusion based on the results of your analysis? Only one of the statements is fully correct.

Statement 1: “I rejected the null hypothesis and concluded that at least one fat type’s mean amount of fat absorbed was significantly different than the other fat types.”

Statement 2: “I rejected the null hypothesis and concluded that all of the fat types had significant differences in mean amounts of fat absorbed.”

Statement 3: “I failed to reject the null hypothesis and concluded that there was not a statistically significant difference in the mean amounts of fat absorbed among the fat types.”

Your answer here (1, 2, or 3): 1

Question 5 - 10 points

When the null hypothesis in ANOVA is rejected, you conclude that at least one group mean is different than the others. You may then wonder which of the means is different. There are numerous tests that have been developed to answer this question. These are sometimes referred to as “post hoc” tests because they are usually done after an ANOVA has returned a significant result.

One of the most common of these is the Tukey Honest Significant Difference test, often shortened to Tukey’s HSD. You will conduct this analysis to determine which of the fat type means had statistically significant differences from each other. You will need to do some reading on your own to figure out how to conduct and interpret this test.

First, answer the following two questions:

A) How many unique pairwise comparisons of fat type means are possible to test in this data set?

Your answer here: The number of unique pairwise comparisons has 6 unique pairwise comparisons.

B) As discussed in class, multiple pairwise comparisons cause the familywise Type 1 error rate to increase as the number of pairwise comparisons increases; this is why you will use Tukey’s HSD, which adjusts for this increase to keep the familywise error rate at 0.05 (5%). If you were not aware of this problem and conducted as many independent-samples t tests as there are unique pairwise comparisons in this data set, what would the familywise Type 1 error rate for those tests be?

Your answer here: Familywise error rate = 1 - $(1 - 0.05)^6$ = 1 - 0.735 = 0.265. Thus, the familywise Type 1 error rate would be 26.5%.

Next, conduct the Tukey HSD test and answer the two questions below:

```r
# Conduct the Tukey HSD test
TukeyHSD(doughnuts.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = total_fat ~ fat_type_factor, data = doughnuts)
##
## $fat_type_factor
##                       diff        lwr       upr     p adj
## Peanut-Canola            4 -12.232221 20.232221 0.8998057
## Shortening-Canola       13  -3.232221 29.232221 0.1461929
## Sunflower-Canola       -10 -26.232221  6.232221 0.3378150
## Shortening-Peanut        9  -7.232221 25.232221 0.4270717
## Sunflower-Peanut       -14 -30.232221  2.232221 0.1065573
## Sunflower-Shortening   -23 -39.232221 -6.767779 0.0039064
```

C) Based on the results of your Tukey HSD test, how many pairs of means have a statistically significant difference from each other?

Your answer here: There is 1 pair of means that has a statistically significant difference from each other, as indicated by a p-value less than 0.05.

D) List the pair/s of means that have statistically significant differences here. Be sure to include the names of the groups.

Your answer here: The pair of means that has a statistically significant difference is Sunflower-Shortening, with a p-value of 0.0039.