Problem Set 1, Fall 2021

Ben Karabinus

```
# load required packages
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

library(ggeasy)
library(lawstat)
```

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.

```
# Loads the CSV file into memory. You may need to adapt this line to work on
your computer
doughnuts <- read.csv("doughnuts.csv",header=TRUE,sep=",")
# Creates a new variable and tells R that the values are categorical ("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.

```
# View dataset attributes
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
# Function to calculate sample standard deviation(built-in sd is alternative)
std_samp <- function(data){
  sqrt(sum((data-mean(data))^2/(length(data)-1)))
}

# compute sample mean and standard deviation for each fat type as tibble usin
g dplyr
fat_type_mean_sd <- doughnuts %>%
                  group_by(fat_type_factor) %>%
                  summarise_at(vars(total_fat),
                  list(mean=mean, standard_deviation=std_samp))
fat_type_mean_sd
```

```
## # A tibble: 4 × 3
##   fat_type_factor  mean standard_deviation
##   <fct>           <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):

$\bar{x} = 72, S = 13.341664$

Shortening mean and SD (your answer here):

$\bar{x} = 85, S = 7.771744$

Peanut mean and SD (your answer here):

$\bar{x} = 76, S = 9.879271$

Sunflower mean and SD (your answer here):

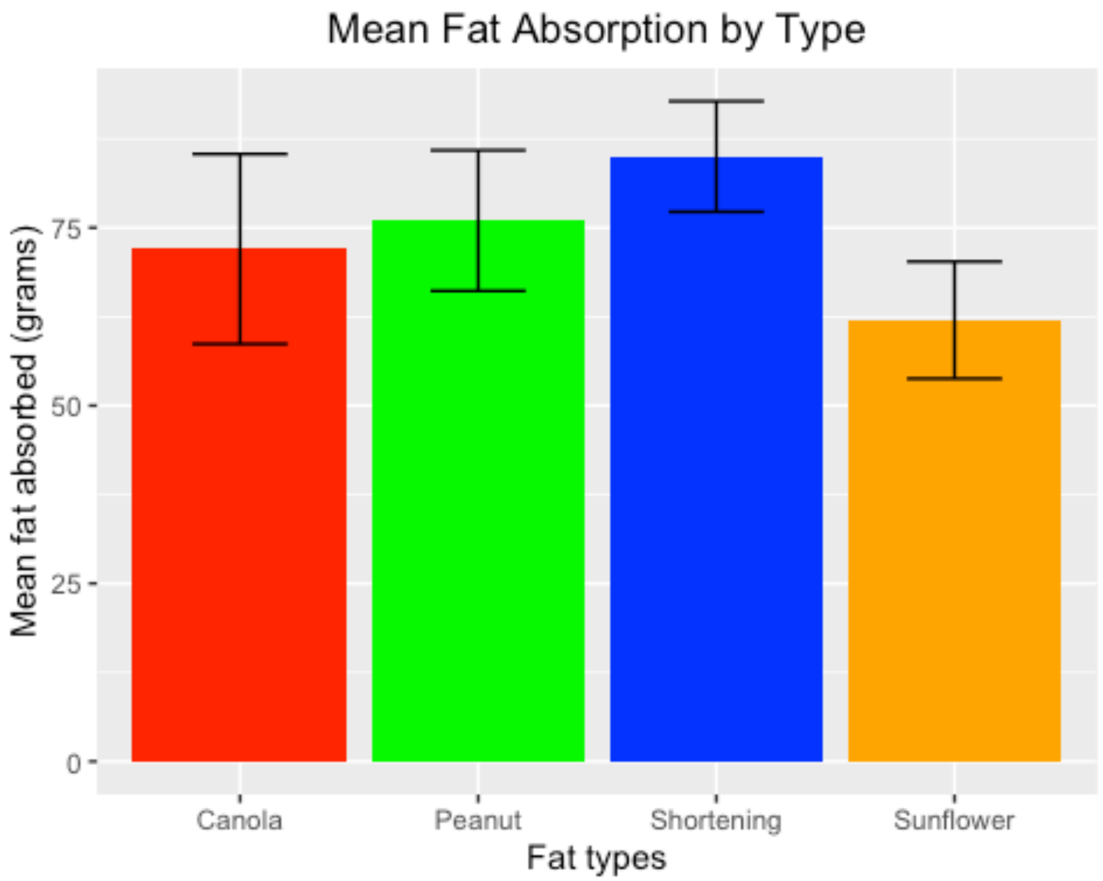
$\bar{x} = 62, S = 8.221922$

Next, create a bar plot to visualize the differences in the means. If you are unsure of what a bar plot looks like, there are some examples 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.

```
# create bar plot using ggplot2 - geom_bar()
d_plt <- ggplot(fat_type_mean_sd)+
                geom_bar(aes(x=fat_type_factor, y=mean),
                stat ="identity",fill=c("red","green", "blue", "orange"))+
                xlab("Fat types")+
                ylab("Mean fat absorbed (grams)")+
                ggtitle("Mean Fat Absorption by Type")+
                geom_errorbar(aes(x=fat_type_factor, ymin=mean-standard_devia
tion,
                                  ymax=mean+standard_deviation),width=0.4)+
                ggeasy::easy_center_title()

d_plt
```

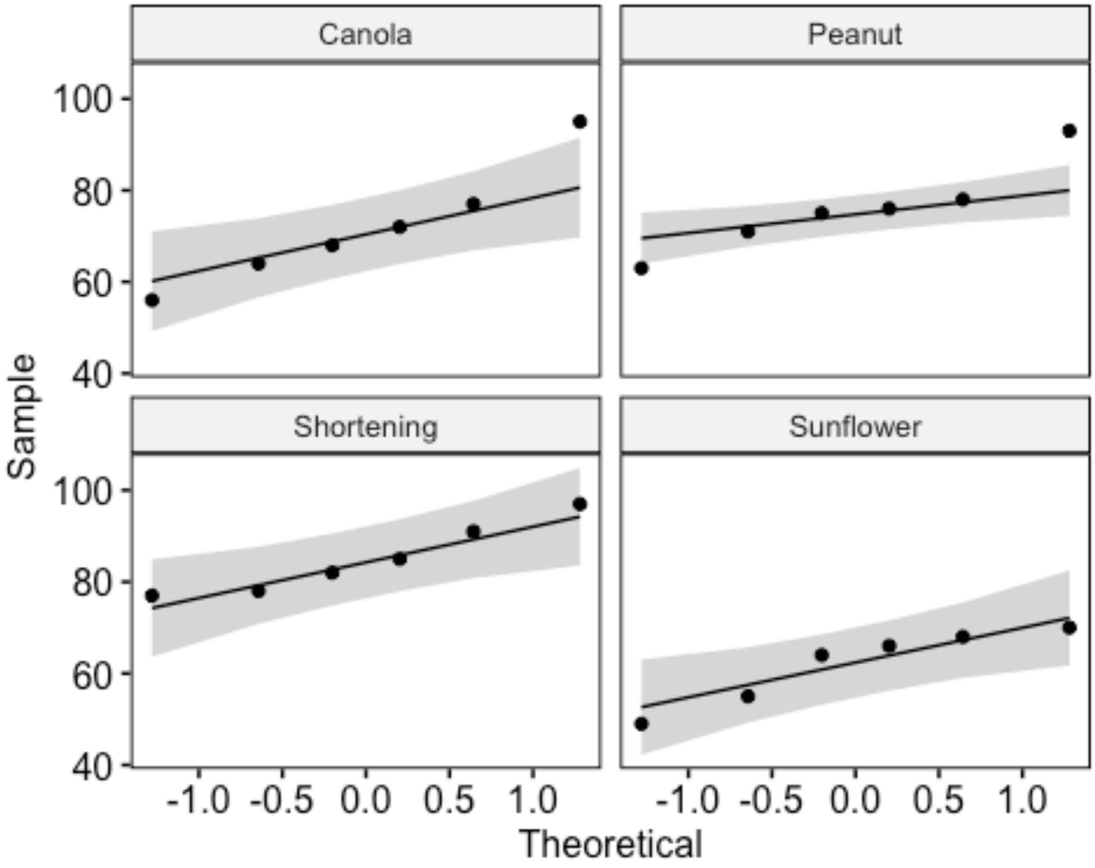


Question 2 - 10 points

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:

```
# use ggqqplot to assess the assumption of normality for data by factor level
ggqqplot(doughnuts, x="total_fat", facet.by = "fat_type_factor")
```



Next, assess this assumption quantitatively:

```
# use the Shapiro-Wilk test with dplyr to assess normality and display as tib
ble
sw_doughnuts <- doughnuts %>%
                group_by(fat_type_factor)%>%
                summarize(pval=shapiro.test(total_fat)$p)
sw_doughnuts
```

```
## # A tibble: 4 × 2
##   fat_type_factor  pval
##   <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 a qqplot, specifically ggqqplot available in the ggpubr library. Basic rule of thumb for ggqqplot is if points fall within or close to the gray shaded region data are approximately normally distributed.

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

Your answer here:

I used the Shapiro-Wilk test. The null hypothesis of the Shapiro-Wilk test is that data from an underlying population are normally distributed. If the resulting p-value $<= \alpha$ the null hypothesis will be rejected.

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

Your answer here:

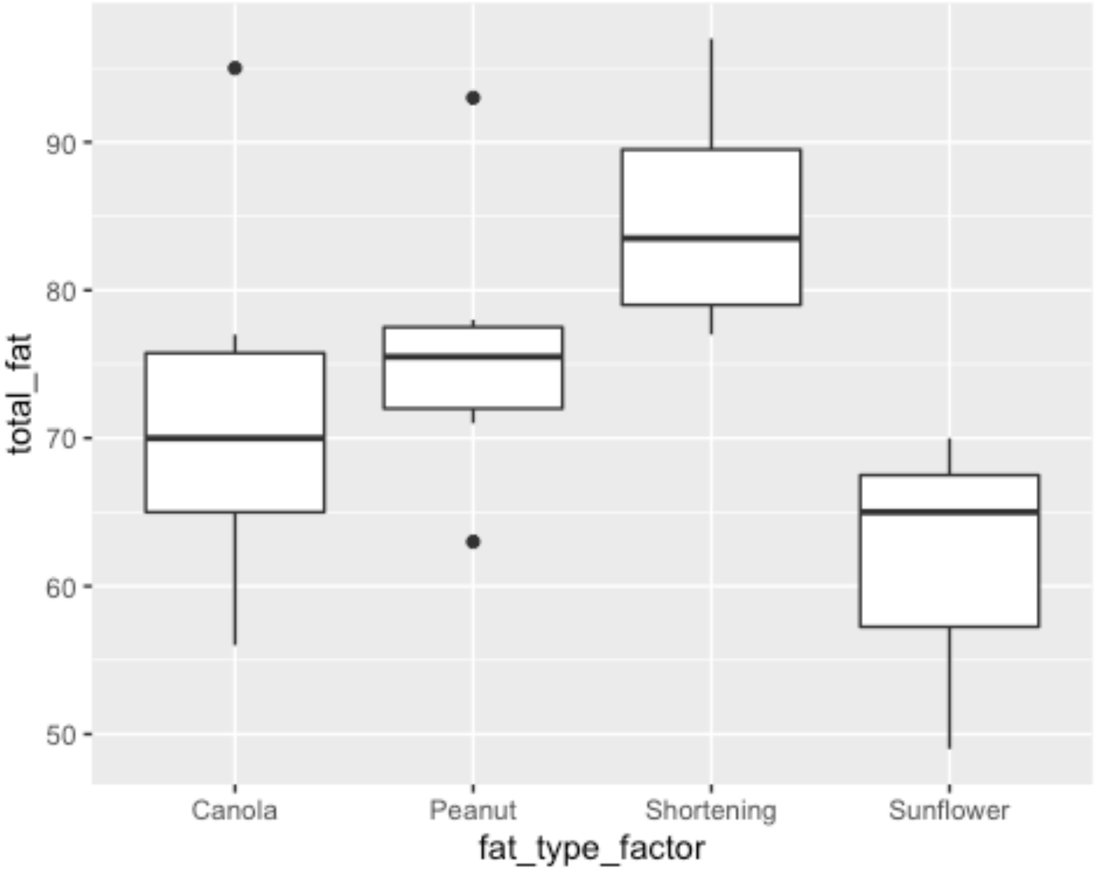
The results of the above Shapiro-Wilk test as applied to data at each factor level fail to reject the null hypothesis. The data meet the assumption of normality.

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:

```
# create ggplot geom_boxplot to assess variance between groups visually
ggplot(doughnuts, aes(x=fat_type_factor, y=total_fat))+
       geom_boxplot()
```



Next, assess this assumption quantitatively:

```
# conduct the Brown-Forsythe test
levene.test(doughnuts$total_fat, doughnuts$fat_type_factor)

##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  doughnuts$total_fat
## Test Statistic = 0.3434, p-value = 0.7942
```

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:

To assess the assumption of equal variances visually I used a box plot. The length of each box in the box plot corresponds to the amount of variance in the data for each factor level. The box plot seems to show differences in variance between factor levels. Quantitative analysis is warranted. It's also worth noting the sample size is fairly small and may affect the outcome of both visual and quantitative assessments.

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

Your answer here:

To assess the assumption of equal variances across groups quantitatively I used a Brow-Forsythe test. The Brown-Forsythe test assumes a null hypothesis of equal variances across groups. If the test produces a p-value $<= \alpha$ the null hypothesis is rejected.

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 results of the Brown-Forsythe test the null hypothesis should not be rejected. Thus the data meet the assumption of equal variance across groups.It's also worth noting the sample size is fairly small and may affect the outcome of both visual and quantitative assessments.

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
doughnuts.aov <- aov(total_fat~fat_type_factor, data=doughnuts)
#display summary of results
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 being tested is that there is no difference in mean total fat absorption of doughnuts being cooked in the fat types Canola Oil, Peanut Oil, Sunflower Oil and Shortening.

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

Your answer here:

With a $p - value = 0.00688$ the null hypothesis that mean total fat absorption is equal among the four fat types used during the cooking process is rejected.

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

Statement A: “I rejected the null hypothesis and concluded that at least one fat type had a statistically significantly different mean amount of fat absorbed than the other fat types.”

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

Statement C: “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 (A, B, or C):

A.

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.

In my experience, 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:

There a 6 unique pair-wise comparisons of fat type means to test in the data set, calculated as follows:

- $k = 4$
- pair-wise comparisons = $\big(k(k-1)\big)/2 = 6$

B) As discussed in class, multiple pairwise comparisons cause the family-wise 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 family-wise 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 family-wise Type 1 error rate for those tests be?

Your answer here:

The accumulation of the family-wise type 1 error rate can be calculated as follows:

- $\alpha = 0.05$
- $n = number\ pair - wise\ comparisons = 6$
- $accumulation\ of\ family - wise\ rate = 1 - (1 - \alpha)^n$
- $accumulation\ for\ the\ doughnuts = 1 - (1 - 0.05)^6 = 0.2649081094$

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

```
# use the TukeyHSD function to perform analysis
TukeyHSD(doughnuts.aov, conf.level = .95)

##   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:

At a 95% confidence level ($\alpha = 0.05$) the difference between mean fat in sunflower oil and shortening is the only pair whose difference between means that shows statistical significance.

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:

At a 95% confidence level ($\alpha = 0.05$) the difference between mean fat absorbed by doughnuts cooked in sunflower oil and doughnuts cooked in shortening ($sunflower - shortening$) is the only pair which shows statistical significance.