Problem Set 1, Winter 2022

Krista Miller

knitr::opts\_chunk$set(echo = TRUE)  
# Load any packages, if any, that you use as part of your answers here  
  
# For example:   
library(ggplot2)  
library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.0.5

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(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble 3.1.0 v purrr 0.3.4  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readxl)  
library(nortest)  
library(ggpubr)  
library(lawstat)

## Warning: package 'lawstat' was built under R version 4.0.5

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.

doughnuts <- read.csv("doughnuts.csv",header=TRUE,sep=",") # Loads the CSV file into memory. You may need to adapt this line to work on your computer  
  
colnames(doughnuts) <- c('fat\_type', 'total\_fat')  
  
doughnuts$fat\_type\_factor <- as.factor(doughnuts$fat\_type) # Creates a new variable and tells R that the values are categorical ("factor")

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.

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.

# Write your code to compute your group means and standard deviations in this code chunk  
  
#means:   
dough\_means<- doughnuts%>%group\_by(fat\_type)%>%summarize(mean\_size = mean(total\_fat))  
  
#standard deviations:  
dough\_std<- doughnuts%>%group\_by(fat\_type)%>%summarize(std= sd(total\_fat))

Canola mean and SD (your answer here): 72; 13.34

Shortening mean and SD (your answer here): 85; 7.77

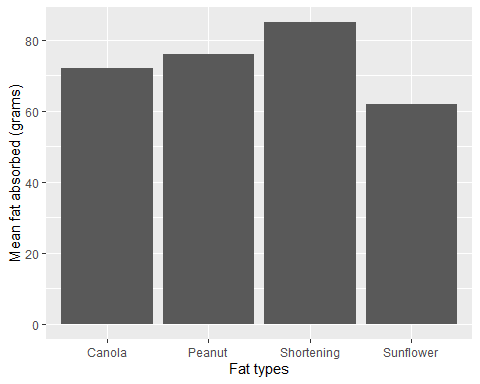
Peanut mean and SD (your answer here): 76; 9.87

Sunflower mean and SD (your answer here): 62; 8.22

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.

# Write your code for your bar plot here - be sure that it displays in your knitted document  
g <- ggplot(dough\_means, aes(fat\_type, mean\_size)) +   
 xlab("Fat types") +  
 ylab("Mean fat absorbed (grams)")  
g + geom\_col()

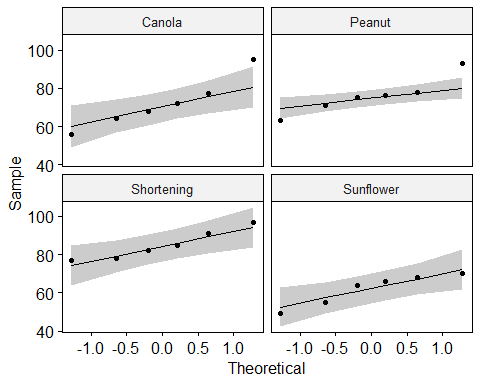


## Question 2 - 10 points

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:

# Code for visual assessment - be sure that your visual assessment is visible in your knitted document  
ggqqplot(doughnuts,x="total\_fat",facet.by = "fat\_type")



Next, assess this assumption quantitatively:

# Code for quantitative assessment - be sure that your results are visible in your knitted document  
doughnuts%>%group\_by(fat\_type)%>%summarize(pval=shapiro.test(total\_fat)$p)

## # A tibble: 4 x 2  
## fat\_type pval  
## <chr> <dbl>  
## 1 Canola 0.741  
## 2 Peanut 0.607  
## 3 Shortening 0.593  
## 4 Sunflower 0.310

Finally, answer the three questions below:

1. What type of visualization did you use to assess the assumption of normality visually?

Your answer here: qqplot

1. What type of quantitative test did you conduct to assess the assumption of normality quantitatively?

Your answer here: Shapiro-Wilk Test

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

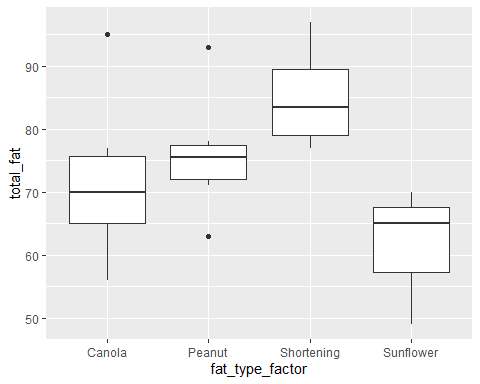
Your answer here: Visually, the data points fall along the straight diagonal line in the Q-Q plot, so the dataset likely follows a normal distribution. The Shapiro-Wilk Test tests the null hypothesis that the samples arise from a normal distribution. In this case,,he p-values (0.741, 0.607, 0.593, 0.310) are greater than the alpha level of 0.05, suggesting that the samples come from a normal distribution.

## 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:

# Code for visual assessment - be sure that your visual assessment is visible in your knitted document  
  
ggplot(doughnuts, aes(x=fat\_type\_factor, y= total\_fat)) + geom\_boxplot()



Next, assess this assumption quantitatively:

# Code for quantitative assessment - be sure that your results are visible in your knitted document  
levene.test(doughnuts$total\_fat,doughnuts$fat\_type)

##   
## 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:

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

Your answer here: Boxplot

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

Your answer here: Brown-Forsythe

1. 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: The Brown-Forsythe test p-value of 0.7942 does not give strong evidence against the null hypothesis of equal variances.

## 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:

# Write your code to conduct the one-way ANOVA here   
  
doughnuts.aov <- aov(total\_fat~fat\_type\_factor, data=doughnuts)  
   
  
# Don't forget to display the results of your analysis by using the summary() function!  
  
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:

1. What is the null hypothesis being tested in this one-way ANOVA analysis?

Your answer here: The null hypothesis is that the groups have equal means.

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

Your answer here: The small p-value (0.00688) gives strong evidence against the null hypothesis.

1. 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:

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

Your answer here: Because we have 4 groups (Canola, Peanut, Shortening, and Sunflower), we will need to make the following 6 comparisons:

Canola-Peanut Canola-Shortening Canola-Sunflower Peanut-Shortening Peanut-Sunflower Shortening-Sunflower

1. 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: 0.264 (26.4%)

Calculating the error rate for the entire set of comparisons= 1- (1-alpha)^ C (alpha is significance level for a single comparison, C equals the number of comparisons)

1- (1-0.05)^6 = 0.264

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

# Write your code to conduct the Tukey HSD test here  
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

# Be sure that your results are visible in the knitted document

1. 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: 1 (Sunflower-Shortening is significantly different at the 95 percent confidence level)

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

Your answer here: Sunflower-Shortening has a statistically significant difference at the 95 percent confidence level (p= 0.0039)