Lesson 6: Do Diets Really Work? (continued)

CDT I.M.Smrt

Background

Recall that in Lesson 5 we investigated the impact diet (Vegetarian, Keto, and None) has on weight loss (kg lost).

```
library(tidyverse)
library(ggResidpanel)

## Load data
dietdat <- read_csv('https://raw.githubusercontent.com/jkstarling/MA376/main/WeightLoss.csv')

# glimpse(dietdat)
#Change the data types accordingly.
dietdat <- dietdat %>% mutate(diet = as.factor(diet)) %>%
  mutate(exercise = as.factor(exercise))
```

(1) Before we move on, let's create a sources of variation diagram.

Observed variation in:	Sources of explained variation:	Sources of unexplained variation:

Inclusion criteria:
Design:

(2) We can also calculate the F statistic and find the p-value of the statistic using R:

```
# dietdat <- dietdat %>% mutate(diet.cont = diet)
# contrasts(dietdat$diet.cont) = contr.sum
# diet.lm <- lm(xxx)
# SST <- sum(xxx); #SST
# SSE <- sum(xxx); #SSE
# SSM <- sum(xxx); #SSESST-SSE; #SSM
# n <- nrow(dietdat)
# dfmod <- xxx
# dfred <- xxx
# Fstat <- (xxx/dfmod) / (xxx/dfred); Fstat
#
# pf(Fstat, xxx, xxx, lower.tail = FALSE)
```

(3) Remind ourselves how to conduct an ANOVA with R:

```
# one-way ANOVA
# diet.aov <- aov(xxx)
# summary(diet.aov)

#alternative
# diet.lm2 <- dietdat %>% lm(xxx, data=.)
# anova(diet.lm2)
```

Based on the results of the F-test we concluded that the observed differences in mean weight loss among the diets is more than we would expect by chance alone. However, knowing there is a statistically significant association between diet and average kilograms lost does not tell us the entire story.

(4) What else might we want to know?

(5) Let's draw a picture to help us visualize the various means.

```
# dietdat %>% ggplot(aes(x=xxx, y=xxx)) +
  # geom_xxx() + stat_summary(fun.y=mean, geom="point", shape=20, size=5, color="red", fill="red") + la
```

(6) To figure this out, we conduct *post-hoc analyses*. Note: We only do post-hoc analyses when the F-test is statistically significant to protect against an inflated experiment-wise Type I error rate. What is Type I error? What happens to the experiment-wise Type I error rate as the number of comparisons increase?

A. All Pairwise Comparisons For this method we compare average kilograms lost for all possible diet pairs. There are different methods for this type of pairwise comparison, but our textbook uses the formula:

$$\text{difference in means } (\bar{y}_i - \bar{y}_j) \pm \text{(Multiplier) (SE of Residuals)} \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

(7) For 95% confidence intervals the multiplier is approximately 2. Why?

There are other methods to compare the mean for one population to the mean of another, e.g., Bonferroni corrections and Tukey's Honest Significant Difference (TukeyHSD in R).

Note: when a confidence interval includes 0 (null hypothesized value), we do not have enough evidence to conclude there is a difference in the mean weight loss for the two diets.

(8) Calculate the 95% CIs for the pairwise comparisons between None and Keto. What is a potential problem with doing this for all 3 pair-wise comparisons, particularly at a 95% confidence level?

```
# diet.mean <- dietdat %>% group_by(xxx) %>% summarise(mn = mean(xxx), num=n())
# diet.mean
```

```
# tstar <- qt(xxx)
# se.est <- summary(diet.lm)$sigma
# low <- xxx
# upp <- xxx
# c(low, upp)
```

(9) Calculate the 95% confidence intervals for each pair using TukeyHSD. For which groups can we conclude there is a statistically significant difference in the average number of kilograms lost?

```
# TukeyHSD(xxx)
```

B. t- Confidence Intervals for Each Population Mean We can also calculate a confidence interval for each treatment group (diet group) using the formula:

$$\bar{y}_i \pm (\text{Multiplier}) \times \frac{\text{SE of Residuals}}{\sqrt{n_i}}$$

$$\bar{y}_i \pm t^*_{df,\alpha/2} \times \frac{\text{SE of Residuals}}{\sqrt{n_i}}$$

(10) Which groups have statistically significant population mean weight loss?

```
# Fit the Multiple Means Model
# confint(xxx)
```

(11) Lets try to figure out how to do this "by hand":

```
# diet.mean
# kmean <- xxx
# nmean <- xxx
# vmean <- xxx
#
# tstar <- qt(0.025, 180-3,lower.tail = FALSE)
# c(kmean - tstar*se.est / sqrt(60), xxx)
# c(xxx,xxx)
# c(xxx,xxx)
```

(12) Interpret the confidence intervals.

C. Prediction Intervals for a New Observational Unit A prediction interval is a range of values that predicts the value of a new observation based on the existing model. We can calculate a prediction interval using the formula:

$$\bar{y}_i \pm (\text{Multiplier})\ (\text{SE of Residuals})\sqrt{1 + \tfrac{1}{n_i}}$$

Note the validity conditions for prediction intervals on the bottom of page 96

(13) Estimate a prediction interval for average kilograms lost for an individual who is on the Keto diet. Interpret the prediction interval.

```
# newobs.data <- data.frame(diet='Keto')
# predict(xxx)
# newobs.data <- data.frame(diet='xxx')
# predict(xxx)
# newobs.data <- data.frame(diet='xxx')
# predict(xxx)
```

(14) Do the same "by hand".

```
# tstar <- qt(0.025, 180-3,lower.tail = FALSE)
# c(kmean - tstar*se.est*sqrt(1+1/60), xxx)
# c(xxx, xxx)
# c(xxx, xxx)
```

(15) When we are making a prediction where is our 'uncertainty' coming from?

(16) What happens to confidence intervals when we increase the confidence level to 99%? What if we increase sample size?