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:
weight loss (in kgs)	diet groups: keto, vegetarian, none	exercise level
		Age
Inclusion criteria: healthy adults		"genetics"
Design: randomized, zero-blind		Unknown

(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$diet.cont) = contr.sum
diet.lm <- lm(wloss~0+diet, data=dietdat)
SST <- sum((dietdat$wloss - mean(dietdat$wloss))^2); #SST
SSE <- sum((dietdat$wloss - diet.lm$fitted.values)^2); #SSE
SSM <- sum((diet.lm$fitted.values - mean(dietdat$wloss))^2); #SSESST-SSE; #SSM
n <- nrow(dietdat)
dfmod <- 2
dfred <- n - dfmod - 1
Fstat <- (SSM/dfmod) / (SSE/dfred); Fstat
```

```
## [1] 10.68527
```

```
pf(Fstat, dfmod, dfred, lower.tail = FALSE)
```

```
## [1] 4.158189e-05
```

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

```
# one-way ANOVA
diet.aov <- aov(wloss~diet, data = dietdat)
summary(diet.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## diet          2  363.9  181.95   10.69 4.16e-05 ***
## Residuals   177 3014.0   17.03                     
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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

```
## Analysis of Variance Table
## 
## Response: wloss
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## diet        2  363.9 181.952  10.685 4.158e-05 ***
## Residuals 177 3014.0  17.028                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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?

Which population means are different? How much do they differ by?

(5) 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?

Type I error is the decision to reject the null hypothesis when the null hypothesis is true. The experiement-wise Type I error rate will increase as the number of comparisons increases.

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:

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

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

The Empirical Rule (Page 89)! For a symmetric mound-shaped distribution approximately 68% of the values lie within one standard deviation of the mean; approximately 95% lie within two standard deviations of the mean. (and it is the t-distribution).

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.

(7) 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(diet) %>% summarise(mn = mean(wloss), num=n())
diet.mean
```

```
## # A tibble: 3 x 3
##   diet          mn   num
##   <fct>      <dbl> <int>
## 1 Keto        4.31    60
## 2 None        2.84    60
## 3 Vegetarian  6.31    60
```

```
tstar <- qt(0.95, 120-2)
se.est <- summary(diet.lm)$sigma
low <- (diet.mean$mn[2] - diet.mean$mn[1]) - tstar * se.est * sqrt(2/60)
upp <- (2.84 - 4.31) + tstar * se.est * sqrt(2/60)
c(low, upp)
```

```
## [1] -2.7183711 -0.2209623
```

The Empirical Rule (Page 89)! For a symmetric mound-shaped distribution approximately 68% of the values lie within one standard deviation of the mean; approximately 95% lie within two standard deviations of the mean. (and it is the t-distribution).

(8) 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(diet.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

```
## Fit: aov(formula = wloss ~ diet, data = dietdat)
##
## $diet
##                      diff        lwr       upr     p adj
## None-Keto       -1.469333 -3.2500664 0.3113998 0.1278282
## Vegetarian-Keto  2.000000  0.2192669 3.7807331 0.0234375
## Vegetarian-None  3.469333  1.6886002 5.2500664 0.0000233
```

Vegetarian-Keto, Vegetarian-None.

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

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

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

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

Since zero is not in the confidence intervals for any of the diet groups, average weight loss is statistically significant for every diet.

```r
# Fit the Multiple Means Model
confint(diet.lm)
```

```
##                   2.5 %   97.5 %
## dietKeto       3.259339 5.361994
## dietNone       1.790006 3.892661
## dietVegetarian 5.259339 7.361994
```

(10) Lets try to figure out how to do this “by hand”:

```r
diet.mean
```

```
## # A tibble: 3 x 3
##   diet          mn   num
##   <fct>      <dbl> <int>
## 1 Keto        4.31    60
## 2 None        2.84    60
## 3 Vegetarian  6.31    60
```

```r
kmean <- diet.mean$mn[1];
nmean <- diet.mean$mn[2];
vmean <- diet.mean$mn[3];

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

```
## [1] 3.244669 5.376664
```

```
c(nmean - tstar*se.est / sqrt(60),nmean + tstar*se.est / sqrt(60))
```

```
## [1] 1.775336 3.907331
```

```
c(vmean - tstar*se.est / sqrt(60),vmean + tstar*se.est / sqrt(60))
```

```
## [1] 5.244669 7.376664
```

(11) Interpret one of the significant 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:

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

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

(12) 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(diet.lm, newobs.data, interval = 'prediction')
```

```
##        fit       lwr     upr
## 1 4.310667 -3.900462 12.5218
```

```
newobs.data <- data.frame(diet='None')
predict(diet.lm, newobs.data, interval = 'prediction')
```

```
##        fit       lwr      upr
## 1 2.841333 -5.369795 11.05246
```

```
newobs.data <- data.frame(diet='Vegetarian')
predict(diet.lm, newobs.data, interval = 'prediction')
```

```
##        fit       lwr     upr
## 1 6.310667 -1.900462 14.5218
```

We are 95% certain that a (future/new) individual who is on the Keto diet will lose between -3.9 and 12.5 kilos.

(13) Do the same "by hand".

```
c(kmean - tstar*se.est*sqrt(1+1/60), kmean + tstar*se.est*sqrt(1+1/60))
```

```
## [1] -4.015042 12.636376
```

```
c(nmean - tstar*se.est*sqrt(1+1/60), nmean + tstar*se.est*sqrt(1+1/60))
```

```
## [1] -5.484376 11.167042
```

```
c(vmean - tstar*se.est*sqrt(1+1/60), vmean + tstar*se.est*sqrt(1+1/60))
```

```
## [1] -2.015042 14.636376
```

(14) When we are making a prediction where is our ‘uncertainty’ coming from?

Prediction intervals take into account individual variations in weight loss and variations in weight loss between diets.

This is why prediction intervals tend to be wider than confidence intervals which only take into account the sample to sample variation in the sample mean.

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

*The predition interval will widen if we increase the confidence level to 99% since the multiplier will increase (to about 3). If we incrase the sample size, it will not have a significant affect as the square root value will only get closer to 1. *