

## Lesson 7: Do Diets Really Work? (continued #2)

LTC James K Starling

Last compiled on 24 January, 2022

### Background

Recall that in Lesson 5 and 6 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 data types appropriately.
dietdat <- dietdat %>% mutate(diet = as.factor(diet)) %>%
  mutate(exercise = as.factor(exercise))
```

- (1) What was our null and alternative hypotheses (in words and symbols, assuming we are going to conduct ANOVA)?

The null hypothesis is there is no association between weight loss and diet type, the alternative hypothesis is that there is an association between weight loss and diet type.

$$H_0 : \mu_{Vegetarian} = \mu_{Keto} = \mu_{None}$$

$$H_A : \text{at least one of the } \mu \text{ is different from the others.}$$

- (2) For the diet study, what is a Type I error/Type I error rate? Type II error/Type II error rate?

Type I error (false positive)- rejecting  $H_0$  when  $H_0$  is TRUE. In the diet study, a type I error will occur when we say there is an association between diet and weight loss when there actually is no association. The type I error rate is the probability of saying the above (also known as the significance level ( $\alpha$ )).

Type II error (false negative)- failing to reject  $H_0$  when  $H_0$  is FALSE. In the diet study a Type II error will occur when say there is no association between diet and weight loss when in fact there is an association. The Type II error rate is the probability of saying the above ( $\beta$ ).

- (3) What is statistical power?

Statistical power is the probability that the researchers find evidence against the null hypothesis and in favor of the alternative hypothesis when the alternative hypothesis is TRUE.

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

```
multi.lm <- lm(wloss~0+diet, data=dietdat)
SST <- sum((dietdat$wloss - mean(dietdat$wloss))^2); #SST
SSE <- sum((dietdat$wloss -multi.lm$fitted.values)^2); #SSE
SSM <- SST-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
```

(5) 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
```

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.

(6) What else might we want to know?

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

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 experiment-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:

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

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

`\textcolor{blue}`{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}.

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% 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?

[Vegetarian-Keto, Vegetarian-None.](#)

```
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
```

- (9) Interpret one of the significant confidence intervals.

**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?

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

```
# Fit the Multiple Means Model
multi.lm <- lm(wloss ~ 0 + diet, data = dietdat)
summary(multi.lm)
```

```
##
## Call:
```

```
## lm(formula = wloss ~ 0 + diet, data = dietdat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.271  -2.809  -0.576   2.651  10.449
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## dietKeto         4.3107     0.5327   8.092 9.08e-14 ***
## dietNone         2.8413     0.5327   5.333 2.91e-07 ***
## dietVegetarian   6.3107     0.5327  11.846 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.127 on 177 degrees of freedom
## Multiple R-squared:  0.5696, Adjusted R-squared:  0.5623
## F-statistic: 78.08 on 3 and 177 DF,  p-value: < 2.2e-16
```

```
confint(multi.lm)
```

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

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

```
dietdat %>% group_by(diet) %>% summarise(n=n(), mn=mean(wloss))
```

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

```
kmean <- 4.31;
nmean <- 2.84;
vmean <- 6.31;
```

```
tstar <- qt(0.025, 26, lower.tail = FALSE)
SEres <- summary(multi.lm)$sigma
c(kmean - tstar*SEres / sqrt(60), kmean + tstar*SEres / sqrt(60))
```

```
## [1] 3.21495 5.40505
```

```
c(nmean - tstar*SEres / sqrt(60), nmean + tstar*SEres / sqrt(60))
```

```
## [1] 1.74495 3.93505
```

```
c(vmean - tstar*SEres / sqrt(60), vmean + tstar*SEres / sqrt(60))
```

```
## [1] 5.21495 7.40505
```

(12) 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:

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

**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(multi.lm, newobs.data, interval = 'prediction')
```

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

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

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

```
newobs.data <- data.frame(diet='Vegetarian')
predict(multi.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.*

- (14) Do the same “by hand”.

```
c(kmean - tstar*SEres*sqrt(1+1/60), kmean + tstar*SEres*sqrt(1+1/60))
```

```
## [1] -4.242613 12.862613
```

```
c(nmean - tstar*SEres*sqrt(1+1/60), nmean + tstar*SEres*sqrt(1+1/60))
```

```
## [1] -5.712613 11.392613
```

```
c(vmean - tstar*SEres*sqrt(1+1/60), vmean + tstar*SEres*sqrt(1+1/60))
```

```
## [1] -2.242613 14.862613
```

- (15) 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.**

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