MA256 Lesson 15 - Multiple Means (9.1-9.2)

Review: Comparing two means.

Hypotheses: When comparing two means our hypotheses (in symbols) were:

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

Multiple-means: Similar to our method used when considering multiple-proportions:

Hypothesis test:

H_0 : No association between the explanatory variable and the response variable. $\mu_1 = \mu_2 = \ldots = \mu_k$ for k groups.

H_0 : An association exists between the explanatory variable and the response variable. At least one mean is different from the others. At least one $\mu_i \neq \mu_j$ for $i, j \in \{1, \ldots, k\}$, for $i \neq j$.

Like in chapter 8 when we compared multiple proportions, we need a statistic other than the *Mean Group Difference* statistic to make the transition to theory-based a smooth one.

This new statistic is called an F-statistic and the theory-based distribution that estimates our null distribution is called an F-distribution. Unlike the Mean Group Diff statistic, the F-statistic takes into account the variability within each group.

The analysis of variance F-statistic is (see p. 658):

$$F = \frac{\textit{variability between groups}}{\textit{variability within groups}} = \frac{\sum_{i=1}^{I} n_i(\bar{x}_i - \bar{x})^2/(I-1)}{\sum_{i=1}^{I}(n_i - 1)s_i^2/(N-I)}$$

Validity conditions: $\geq$ 20 observations in each group without strong skewness or (i less than 20) be symmetric; & standard deviations of data within each group should be similar (within a factor of 2).

As sample size increases, strength of evidence... increases

As the means move farther apart, strength of evidence... increases. (This is the variability between groups.)

As the standard deviations increase, strength of evidence... decreases. (This is the variability within groups.)

Warm up:

If the balloons popped, the sound wouldn't be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.

The statement on the previous page is from a study that examined college students' comprehension of the ambiguous prose passage. Before the college students were tested to see whether they understood the passage, they were randomly assigned to one of three groups, and then each group was read the passage under one of the following cue conditions: 1) students were shown a picture before they read the passage; 2) students were shown a picture after they heard the passage; 3) students were not shown a picture before or after hearing the passage.

Did you understand what the passage was describing? Would it help to have a picture?

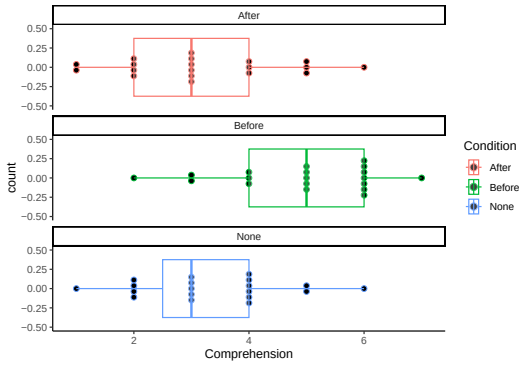
Q1. Fifty-seven students from a small Midwestern college were randomly assigned to be in one of the three groups with 19 in each group. After hearing the passage under the assigned cue condition, they were given a test and their comprehension of the passage was graded on a scale of 1 to 7 with 7 being the highest level of comprehension. What was the null hypothesis in symbols?

$\mu_{no\ picture} = \mu_{picture\ before} = \mu_{picture\ after}$

Q2. Download the data. Plot the data with a dotplot, overlaid with a boxplot (use `alpha = 0.25`), and `facet_wrap()` to separate the figures by treatment level. Does there appear to be a difference between the three treatments?

```
library(tidyverse)
comp <- read.table("http://www.isi-stats.com/isi/data/chap9/Comprehension.txt",
                   header = TRUE,
                   stringsAsFactors = TRUE)

comp %>% ggplot(aes(x = Comprehension, color = Condition)) +
  geom_dotplot(stackdir = "center", dotsize = 0.5) +
  geom_boxplot(alpha = 0.25) +
  facet_wrap(~Condition, ncol = 1) +
   theme_classic()
```



Q3. Calculate sample sizes, sample means, and sample standard deviations. Also calcuate the overall mean of the comprehension scores. Using these values, calculate the F-statistic.

```
comp.tab <- comp %>% group_by(Condition) %>% summarise(n = n(),
                                                       mn = mean(Comprehension),
                                                       SD = sd(Comprehension))
comp.tab
```

```
## # A tibble: 3 x 4
##   Condition     n    mn    SD
##   <fct>     <int> <dbl> <dbl>
## 1 After        19  3.21  1.40
```

```
## 2 Before       19  4.95  1.31
## 3 None         19  3.37  1.26
```

```
xbar <- mean(comp$Comprehension)
I <- 3
N <- length(comp$Condition)

xbar.i <- comp.tab$mn
sd.i <- comp.tab$SD
n.i <- comp.tab$n


VBG <- sum(n.i * (xbar.i - xbar)^2) / (I-1)
VWG <- sum((n.i - 1) * sd.i^2) / (N-I)
fstat <- VBG / VWG
fstat
```

```
## [1] 10.01225
```

```
c(VBG, VWG, fstat)
```

```
## [1] 17.526316  1.750487 10.012249
```

Q4. What does the value of the F-statistic mean in the context of the problem? Do you think you could calculate the p-value to determine the strength of evidence? Try it. What can we conclude about the study?

```
1 - pf(fstat, I-1, N-I)
```

```
## [1] 0.0002002136
```

The Fstatistic appears to be rather extreme. We can conclude that there is a difference between at least one mean and the others at any reasonable level of significance.

Q5. ANOVA. Take a look at the output from the `aov()` command. What are these numbers?

```
aov.comp <- comp %>% aov(Comprehension ~ Condition, data = .)
summary(aov.comp)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Condition    2  35.05   17.53   10.01  2e-04 ***
## Residuals   54  94.53    1.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2- is the (number of groups -1); 54 is the (number of observations - number of groups); the SS-condition (35.05) is the numerator for VBG; SS-residuals is the denominator above for VWG; If you divide the SS values by the df's, you get the mean-squared values. Finally, you divide the MS-condition by MS-resid. you get the F-statistic. We can also notice that the p-value is the same that we calculated above.

The Sum Sq in the ANOVA output stands for sums of squares. It is a measure of variability. The Sum Sq Total is the variability in all the data. It is the sum of the squared distances each response is away from the overall mean of the responses. The Sum Sq Condition is a measure of variability between the groups (the sum of the square of the distance the group means are from the overall mean multiplied by the sample size). It is the variability that is explained by the treatment. The Sum Sq Residuals is the variability that is left over or not explained by the treatment. (the sum of the square of the distance each value is from its corresponding group mean).

Q6. Which groups are different? Use the `TukeyHSD()` command to find out. Explain what you see.

```
TukeyHSD(aov.comp)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Comprehension ~ Condition, data = .)
##
## $Condition
##                     diff        lwr        upr     p adj
## Before-After  1.7368421  0.7023389  2.7713453 0.0004836
## None-After    0.1578947 -0.8766084  1.1923979 0.9282350
## None-Before  -1.5789474 -2.6134505 -0.5444442 0.0015494
```

We can see that the differences between Before-After and None-Before are statistically different from each other. We can verify this with the boxplots we had above.

Diet Comparisons

Because about two-thirds of Americans are considered overweight, weight loss is big business. There are many different types of diets, but do some work better than others? Is low fat better than low carb or is some combination best? Researchers (Garnder et al. 2007) conducted a study involving four popular diets: Atkins (very low carb), Zone (40:30:30 ratio of carbs, protein, fat), LEARN (high carbohydrate, low fat), and Ornish (low fat). They randomly assigned women aged 25–50 with a body mass index (BMI) of 27-40 (overweight and obese) to one of the four diets. The 311 women who volunteered for the program were educated on their assigned diet and were observed periodically as they stayed on the diet for a year. At the end of the year, the researchers calculated the change in BMI (e.g., negative means reduction in BMI) for each woman and compared the results across the four diets.

```
diet <- read.csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Diet_Data.csv", stringsAsFa
```

D1. What is the overarching research question the researchers hoped to answer?

Is there evidence of an association between change in BMI and type of diet used?

D2. What are the observational units in this study? Identify the explanatory and response variables. Classify them as categorical or quantitative. For categorical variables, indicate how many categories are used.

Each of the 311 women in the study.
Explanatory variable: diet (Atkins, Zone, Ornish, LEARN); categorical with 4 categories, Response: change in BMI; quantitative.

D3. Does this study make use of random sampling, random assignment, both, or neither? What are the implications of your answer with regard to scope of inference? Did the researchers collect the data as paired data or as independent samples? In other words, according to the study design, are the responses from one treatment group paired with or independent of the responses from other treatment groups?

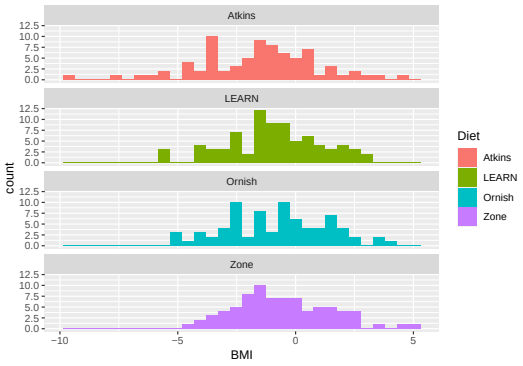
a.Random assignment (yes), random sampling (no), Cause and effect is possible but not generalization to a population, b. Responses from one treatment group are independent of the responses from another treatment group.

D4. State the null and alternative hypotheses, both in words and symbols, for testing the research conjecture. (Recall that the response variable is change in BMI; positive values indicate an increase in BMI and negative values indicate a decrease in BMI from the beginning to the end of the study.)

Null hypothesis: The long-run mean change in BMI is the same for all three diets ($\mu_A = \mu_O = \mu_Z = \mu_L$)
Alternative hypothesis: At least one of the means ($\mu_A, \mu_O, \mu_Z, \mu_L$) is different than the others

D5. Using the dataframe, produce a histogram separated by diet.

```
diet %>% ggplot(aes(x=BMI, fill=Diet)) +
  geom_histogram() + facet_wrap(~Diet, ncol = 1)
```



D6. Calculate the F-statistic and p-value.

```
diet.tab <- diet %>% group_by(Diet) %>% summarise(n = n(),
                                                  mn = mean(BMI),
                                                  SD = sd(BMI))
xbar.d <- mean(diet$BMI)
I.d <- 4
N.d <- length(diet$BMI)

xbar.d.i <- diet.tab$mn
sd.d.i <- diet.tab$SD
n.d.i <- diet.tab$n

VBG.d <- sum(n.d.i * (xbar.d.i - xbar.d)^2) / (I.d-1)
VWG.d <- sum((n.d.i - 1) * sd.d.i^2) / (N.d-I.d)
fstat.d <- VBG.d / VWG.d
pval.d <- 1 - pf(fstat.d, I.d - 1, N.d - I.d)

c(fstat.d, pval.d)
```

```
## [1] 3.80009726 0.01063369
```

```
aov.bmi <- diet %>% aov(BMI ~ Diet, data = .)
summary(aov.bmi)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Diet          3   54.2  18.056     3.8 0.0106 *
## Residuals   307 1458.7   4.752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D7. Calculate the 95% CI for the following differences using `TukeyHSD()`:

```
TukeyHSD(aov.bmi)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BMI ~ Diet, data = .)
##
## $Diet
##                     diff         lwr      upr     p adj
## LEARN-Atkins  0.7291304 -0.17258536 1.630846 0.1590577
## Ornish-Atkins 0.8809125 -0.02954526 1.791370 0.0619904
## Zone-Atkins   1.1202696  0.21855388 2.021985 0.0079830
## Ornish-LEARN  0.1517821 -0.75293279 1.056497 0.9726913
## Zone-LEARN    0.3911392 -0.50477761 1.287056 0.6726684
## Zone-Ornish   0.2393571 -0.66535784 1.144072 0.9033681
```

D8. Which of the confidence intervals calculated in #17 indicate a significant difference between groups? How are you deciding? Based on your analysis, would you conclude that there is one diet that works significantly better than the other three? If so, which one?

Atkins vs. Zone, Atkins vs. Ornish, and Atkins vs. LEARN because they do not include zero. The Atkins diet is associated with a significantly greater decrease in BMI than the LEARN, Ornish, and Zone diets.

D9. Do the sample data for the diet study appear to satisfy the validity conditions for conducting an ANOVA F-test? How are you deciding?

There is not strong skewness in any of the four groups. The standard deviations are: 2.00, 2.14, 2.00 and 2.54. These standard deviations are similar—none of them is more than twice the size of any other.

HALO kills.

Statistics students at Hope College, Holland, MI, wanted to determine whether the level on which the game was played affects how well a person plays, and so data were collected from 90 different games of Halo, 30 from each of the three different levels.

H1. Use a theory-based approach to run an ANOVA test to investigate whether the average number of kills is different for at least one of the groups (or levels). Use the data file HaloKills. Report the value of the F-statistic and the corresponding p-value. (Note: As done earlier, pretend for purposes of this study that each of the 90 games was played by a different player.)

```
halo <- read.table("http://www.isi-stats.com/isi/data/chap9/HaloKills.txt", header = TRUE)
halo$Level <- as.factor(halo$Level)
```

```
halo.tab <- halo %>% group_by(Level) %>% summarise(n = n(),
                                                    mn = mean(Kills),
                                                    SD = sd(Kills))
xbar.h <- mean(halo$Kills)
I.h <- 3
N.h <- length(halo$Kills)

xbar.h.i <- halo.tab$mn
```

```r
sd.h.i <- halo.tab$SD
n.h.i <- halo.tab$n

VBG.h <- sum(n.h.i * (xbar.h.i - xbar.h)^2) / (I.h-1)
VWG.h <- sum((n.h.i - 1) * sd.h.i^2) / (N.h-I.h)
fstat.h <- VBG.h / VWG.h
pval.h <- 1 - pf(fstat.h, I.h - 1, N.h - I.h)

c(fstat.h, pval.h)
```

```
## [1] 0.04521099 0.95581774
```

```r
aov.halo <- halo %>% aov(Kills ~ Level, data = .)
summary(aov.halo)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Level        2      4    1.78   0.045  0.956
## Residuals   89   3500   39.33
```

H2. Based on the p-value, state your conclusion in the context of the study.

We do not have evidence of an association between kills and level.