MA256 Lesson 13 - Paired Data (7.1-7.3)

Review of methods so far...

	1-Categorical	1-Quantitative	2-Categorical	1-Cat/1-Quant
parameter				
statistic				
test				
sd				
stand. stat				
valid. conds.				
CI				

Q1) Are the examples above (and the ones we have considered throughout the semester) representative of *independent groups designs* or *paired designs*? Why?

Q2) What is a *paired design*? What is the difference between a *paired design using matching* and a *paired design using repeated measures*?

Q3) What are a few benefits from using a *paired design*?

Q4) Using the ACFT as an example, explain how you would set up 1) a *paired design using matching* and 2) a *paired design using repeated measures*.

Theory-based approach for paired samples

Parameter/statistic:

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Rounding First Base

Imagine you are at the plate in baseball and have hit a hard line drive. You want to try to stretch your hit from a single to a double. Does the path that you take to "round" first base make much of a difference? For example, is it better to take a "narrow angle" or a "wide angle" around first base? This example is based on an actual study reported in a master's thesis by W. F. Woodward in 1970. In Woodward's study, he used a stopwatch to time 22 different runners going from a spot 35 feet past home to a spot 15 feet before second. He had each runner use each method, with a rest period in between. Here are the summary statistics on the difference in times (narrow - wide) for the 22 players:

Difference	Sample Size	Sample Average (sec)	Sample SD (sec)
narrow-wide	22	0.075	0.088

a. Use an appropriate theory-based method to investigate whether the choice of angle affects the running time. Be sure to find and report a standardized statistic and p-value. Interpret the value of the standardized statistic in the context of the study.

```
# n <- XXXXX
# xdbar <- XXXXX
# se <- XXXXX
# t <- XXXXX
# pval <- XXXXX
# c(t, pval)
```

b. Use an appropriate theory-based method to find and report a 95% confidence interval for the parameter of interest.

```
# multiplier <- XXXXX
# se <- XXXXX
# c(xdbar - multiplier * se, xdbar + multiplier * se)
```

c. Provide a full (four-part) conclusion, including comments on significance, estimation, causation, and generalization.

Filtering Water in Cameroon

Students and professors in the Nursing and Engineering Departments at Hope College went to the central African country of Cameroon to help improve drinking water quality and community health in rural communities by installing water filters in or near homes in one village. The families living in this village had no electricity, had no water distribution system, and got their water from streams. The filters they installed contained a diffuser plate, fine sand, coarse sand, and gravel. Using the new filters, family members were expected to gather their water and filter it before drinking, instead of drinking directly from the stream as was common practice.

Students working on this project examined the quality of the filters by looking at many different variables, including general observations, filter observations, microbiology observations, household practice observations, user perceptions, and water source observations. It should be noted, when making inferences, the water filters in this dataset should be treated as a sample of all filters that could be constructed if this pilot project were expanded to other villages. Thus, inference is to an as-yet-unbuilt, larger population of filters.

STEP 1: Ask a research question.

There are several research questions we will ask in this investigation. The first one is: On average, is there a significant difference in the E. coli counts between the water that has just been filtered and water that is sitting in the bottom of a filter after it was filtered the previous day? The dataset we will use contains results from 14 water filters each giving E. coli counts (per 100 mL) on the first day and the second day after the water was filtered.

STEP 2: Design a study and collect data.

1. What are the observational units?

2. What variables are measured/recorded on each observational unit?

3. Identify the role of these variables (explanatory, response). Classify the variables in this study as categorical or quantitative.

4. Are the samples of the first E. coli count independent or dependent of the samples of the second E. coli count? Explain. Based on your answer, is this an independent samples or paired design?

5. Could the sample size of 14 be large enough to give a valid p-value from a theory-based test of significance?

6. State the null and alternative hypotheses to be investigated with this study in symbols or in words.

STEP 3: Explore the data.

7. What are the average E. coli counts for each day? Did the E. coli in the sample increase or decrease on average from Day 1 to Day 2? Explain. Also give the standard deviations for the E. coli counts for each day.

```
# library(tidyverse)
# ecoli.time <- read.csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Ecoli_time-1.csv")
#
# ecoli.time %>%
#   summarise(
#     mean1 = XXXXX,
#     sd1 = XXXXX,
#     mean2 = XXXXX,
#     sd2 = XXXXX )
```

8. What is the average of the difference (Day 1 – Day 2) in E. coli counts? Does the sign of this average correspond to your answer of increasing or decreasing between Day 1 and Day 2 in #7? Explain. Also give the standard deviation for the differences in E. coli counts.

```
# ecoli.time <- ecoli.time %>%
#   mutate(Difference = XXXXX - XXXXX)
#
# ecoli.time %>%
#   summarise(
#     mean = XXXXX,
#     s = XXXXX,
#     n = XXXXX )
```

STEP 4: Draw inferences beyond the data.

9. Carry out an appropriate test of significance to see whether, on average, there is a genuine difference between the first E. coli count and the second E. coli count. Report your p-value and a 95% confidence interval.

```
# xbar <- XXXXX
# n <- XXXXX
# sd <- XXXXX
# null <- XXXXX
# t <- XXXXX
# pvalue <- XXXXX
# pvalue
#
# multiplier <- XXXXX
# CI <- c(xbar - multiplier * sd, xbar + multiplier * sd)
# CI
```

STEP 5: Formulate conclusions.

10. Based on your p-value and confidence interval, what conclusions can you draw from this test?

11. Can generalizations be made to a larger population? Is there a cause-effect relationship between the variables?

STEP 6: Look back and ahead.

12. Discuss the study design. Why does pairing make sense here? What would you do differently to improve upon this study? What further research would you propose based on your findings from this study?

12-REDUX. What if we did not use a paired test? Conduct a different test, but compare the difference in the measurements from Day 1 and Day 2. What is our conclusion? (assume validity conditions are met for a theory based test) Are there any issues with conducting this test?

```
# mysummary <- ecoli.time %>% summarise(XXXXX) # optional - has been calculated above
# mysummary  # optional - has been calculated above
#
# xbar1 <- XXXXX
# xbar2 <- XXXXX
# s1 <- XXXXX
# s2 <- XXXXX
# n1 <- XXXXX
# n2 <- XXXXX
# sd <- XXXXX
# null <- XXXXX
# statistic <- XXXXX
# t <- XXXXX
# n <- XXXXX
# pvalue <- XXXXX
# pvalue
```

PART II

As mentioned earlier, sand is used in the filter. Different filters had different amounts of sand. The students split the filters into two groups: those that had more than 2 inches of sand and those that had less. We will explore whether this difference results in different E. coli counts when filtering the water. In this sample, we have results from 19 filters, 14 with more than 2 inches of sand and 5 with less than 2 inches of sand.

13. Identify the explanatory and response variables in this study. Classify the variables in this study as categorical or quantitative.

14. What are the average E. coli counts for each sand level? Are the counts higher or lower with the higher level of sand in the filter? Also give the standard deviations for the E. coli counts for each sand level.

```
# ecoli.sand <- read.csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Ecoli_sand-1.csv")
#
# ecoli.sand %>%
#   group_by(XXXXX) %>%
#   summarise(
#     mean_sand = XXXXX,
#     sd_sand = XXXXX,
#     n=n())
```

15. Are the samples of the first E. coli count independent or dependent of the samples of the second E. coli count? Explain.

16. Is the sample large enough to use a theory-based test?

17. (Assume you can conduct a theory-based test and...) Carry out an appropriate test of significance to see whether on average there is a difference in the E. coli counts between filters with more than 2 inches of sand and those with less. Report your p-value, a 95% confidence interval, and your conclusion.

```
# xbar_above <- XXXXX
# xbar_below <- XXXXX
# s_above <- XXXXX
# s_below <- XXXXX
# n_above <- XXXXX
# n_below <- XXXXX
# se <- XXXXX
# null <- XXXXX
# statistic <- XXXXX
# t <- XXXXX
#
# n <- XXXXX
# pvaluesand <- XXXXX
# multiplier <- XXXXX
# CIsand <- c(statistic - multiplier * se, statistic + multiplier * se)
# c(t, pvaluesand, CIsand)
```

PART III

Let's look at one last research question. As a general rule, the flow rate of a water filter in good working condition should be around 1,000 mL/min. Let's test to see whether these water filters had an average flow rate that is significantly different than 1,000 mL/min.

18. List some similarities and differences between this research question and our first research question looking at the average difference in first and second E. coli counts.

19. Create a dotplot of the flow rates. Include average and standard deviation for the flow rate. What do you observe?

```
# ecoli.flow <- read.csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Ecoli_flow-1.csv")
#
# ecoli.flow %>%
#   ggplot(aes(x=Flow_ratio)) + XXXXX
#
# ecoli.flow %>%
#   summarise(
#     meanflow = XXXXX,
#     sflow = XXXXX,
#     nflow = n()   )
```

20. Carry out an appropriate test of significance to see whether the Cameroon water filters (population) tend to flow at a rate different than 1,000 mL/min, or is an average of 1,000 mL/min plausible? State your hypotheses, 95% confidence interval, p-value, and conclusions.

```
# xbar <- XXXXX
# s <- XXXXX          # sample standard deviation
# n <- XXXXX           # sample size
# se <- XXXXX  # standard deviation of the null distribution
# null <- XXXXX
# t <- XXXXX
# pvalue_flow <- XXXXX
#
# multiplier <- XXXXX  # 95% CI
# CI <- c(xbar - multiplier*se, xbar + multiplier*se)
# c(t, pvalue_flow, CI)
```

21. You should have found that 1,000 mL/min is a plausible value for average flow rate for the population of all similar water filters. Does this mean that all the filters have flow rates of about 1,000 mL/min? Is testing a single mean here really telling us what we need to know about these filters? Why or why not?