Paired Data - Class Activity

Vojtech Kejzlar

A group of researchers 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. Researchers 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 data set 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.

Research Question: There are several research questions we can ask with this data. 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 data set we will use contains results from 25 water filters each giving *E. coli* counts (per 100 mL) on the first day and the second day after the water was filtered.

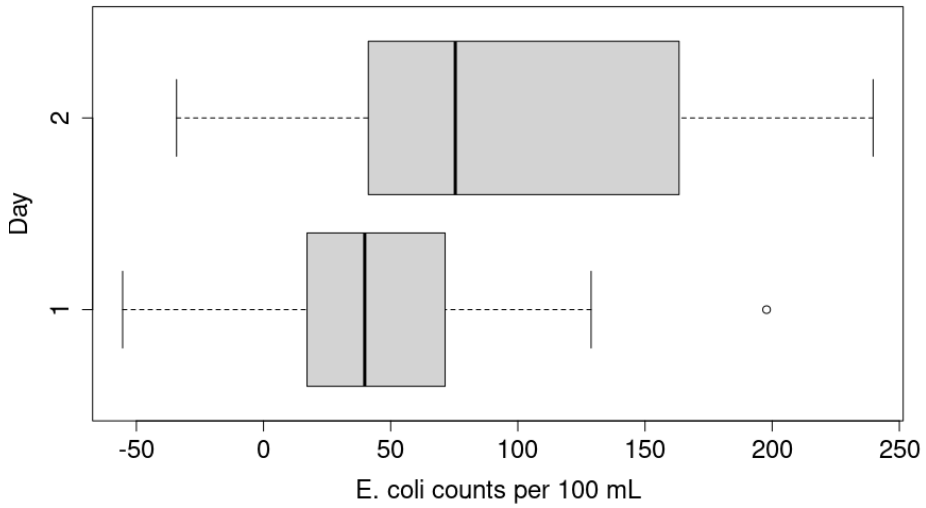
(1) What are the observational units?

(2) Identify the explanatory and response variables in this study. Also classify them as either categorical or quantitative.

(3) 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?

(4) Discuss the study design. Why does pairing make sense here?

(5) The boxplots below are based on the observed *E. coli* counts (per 100 mL) during the first and second day in the homes. Does it appear that there is a 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? Explain briefly.



(6) Estimate the median and IQR *E. coli* counts for each day. Is the median count higher for the measurements taken day 2?

(7) What are two possible explanations for the tendency for higher *E. coli* counts the next day?

(8) To investigate whether there is a difference in the *E. coli* counts, we need to conduct a hypothesis test using appropriate population parameters. In the case of quantitative response with paired data, the population parameter of interest is the mean difference. State the null and the alternative hypotheses using the population parameters and words:

(9) Describe how would you simulate this study so that any difference in the counts was just due to chance?

(10) We will now use technology to repeat the simulation process many times over.
   (a) Go to http://bit.ly/PairedSttSkid
   (b) Hit **Clear** and copy the study dataset from http://bit.ly/EcoliSttSkid
   (c) Hit **Use Data**, scroll to the right and check **Randomize**
   (d) Set the **Number of Shuffles** to 5000 and hit **Randomize**

   Now, use the simulated null distribution to compute the approximate p-value.

(11) Do these simulation analyses reveal that the researchers' data provide strong evidence that there is a 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? Explain the reasoning process by which your conclusion follows from the simulation analyses.

As in all the previous hypothesis testing scenarios, we can notice a familiar pattern when it comes to the shape of the null distribution. This is not a coincidence, but it is guaranteed by the Central Limit Theorem (CLT). CLT for paired data allows us to approximate the null distribution with t-distribution.

(12) Are the conditions for the theory-based approach satisfied? How do you know?

(13) Go to http://bit.ly/TestsSttSkid and carry out the theory-based test for paired data. Report your results here and compare the p-value with the one that you obtained using the simulation-based approach. You will need the summary statistics provided in the table below.

	n	$\bar{x}$	s
Day 1	25	46.032	59.528
Day 2	25	95.440	76.167
Day 1 - Day 2	25	-49.408	91.563

(14) Besides testing of hypotheses, the central limit theorem allows us to construct a confidence interval for the mean difference. Go to http://bit.ly/TestsSttSkid and construct a 95% confidence interval for the mean difference in the *E. coli* counts. Interpret the interval in the context of the problem.

(15) Based on your p-value and confidence interval, what conclusions can you draw from this test?

Additional exploration: Let's look at another 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.

	n	$\bar{x}$	s
Flow rate	23	913	582

(1) 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.