Week_2_Problem_Set

Nguyen Tien Anh Quach

2024-01-29

In this Problem Set, you will use R to conduct comparison of means tests (both parametric and non-parametric) to examine water quality in drinking wells in a fracking region of northeastern Pennsylvania. Please turn in both a write up (around two single-spaced pages of text (not including figures)) and .rmd file of your code. The write up must be submitted as a PDF that was knitted from your Rmd. Use tables to report your results in a clear and structured manner. While it is encouraged, you do not need to use In Line R coding to discuss your results. Your report (PDF) should not contain any R code or error messages. You will also need to submit ONE .Rmd file that contains all the code that you executed in RStudio (aka turn in a knitted version of the document and non-knitted version).

In this problem set, we will analyze water quality data from a study conducted by Molofsky et al. (2013) that examined methane levels in 1,701 drinking wells in Susquehanna County, Pennsylvania. Through our analysis we will seek to determine whether methane levels in drinking water are greater in water wells near fracking sites than in water wells farther away from these sites. The authors grouped the wells into two categories: (1) drinking wells within 1 km radius of a fracking site, and (2) wells located outside a 1km radius of a fracking site. The drinking water wells are also classified as either in a valley or in an upland area (see Molofsky et al., 2013). Prior to completing the problem set, please read with Molofsky article.

Our goal is to make inferences about methane concentrations across fracking group by conducting the following comparisons: (1) Methane levels near fracking sites vs. Methane levels far from fracking sites for ALL observations (2) Methane levels near fracking sites vs. Methane levels far from fracking sites for valley observations (3) Methane levels near fracking sites vs. Methane levels far from fracking sites for upland observations (4) Methane levels in the valley vs. Methane levels in the upland

Data analysis Instructions 1. Download the water quality data from Canvas, PAFracking.xlsx. Be sure to look over the data and then save as a .csv file before reading into RStudio.

```
## [1] "C:/Users/Ng Tien Anh Quach/OneDrive - University of North Carolina at Chapel Hill/UNC/Spring 20
```

2. Summarize and visualize the data by groups as outlined above (1-4). Present descriptive statistics in a professional table or tables. Include your graphics in a clearly labelled appendix.

Proximity Category	Mean Methane Conc.	Median Methane Conc.	Min Methane Conc.	Max Methane Conc.	Standard Deviation of Methane Conc.
Far	684.2574	0.6	0.050	39000	3132.928
Near	795.0171	5.9	0.078	43000	4086.957

Now with only valley sites:

Proximity Category	Mean Methane Conc.	Median Methane Conc.	Min Methane Conc.	Max Methane Conc.	Standard Deviation of Methane Conc.
Far	1186.406	1.3	0.087	39000	4058.772
Near	1225.604	19.0	0.078	43000	5172.061

Now with only upland sites:

Proximity Category	Mean Methane Conc.	Median Methane Conc.	Min Methane Conc.	Max Methane Conc.	Standard Deviation of Methane Conc.
Far	209.7305	0.4	0.05	32000	1753.1023
Near	133.8798	1.4	0.10	8300	799.4312

Now valley vs. upload, no matter the proximity:

Location	Mean Methane Conc.	Median Methane Conc.	Min Methane Conc.	Max Methane Conc.	Standard Deviation of Methane Conc.
Upland	198.2077	0.47	0.050	32000	1644.111
Valley	1195.2426	1.80	0.078	43000	4331.548

3. Conduct the appropriate comparison tests to determine whether methane concentrations vary across 1-4 (above). For each of the four comparisons above, conduct:

a. Parametric t-test
b. Non-parametric t-test
c. Parametric test on the log transformed data.

4. Interpret and discuss the results of each of the tests.
5. Examine and discuss the validity of the assumptions of your comparison tests. Remember to consider the transformation when interpreting the results of the transformed data set. Which of the tests are most valid?

Professional Report Format (1-2 pages, knitted PDF from your .rmd) hint - https://rmarkdown.rstudio.com/lesson-3.html

Your report should be structured with the following sections: 1. Introduction (5 points) a. Broad questions b. Data source c. Variables in the data set 2. Data Description (10 points) a. Descriptive statistics b. Discuss distributions of data c. Discuss censored observations d. Refer to figures e. Include all relevant figures 3. Statistical Analysis and Discussion (15 points) a. Comparison of means results b. Discussion of assumptions c. Discussion of transformations/non-parametric results d. Discuss which tests are most appropriate in this context. e. Include all relevant figures 4. Conclusion (5 points) a. Scope of inference (to what population can you infer?) b. Weaknesses of study/analysis c. Real-world implications/Comparison to Molofsky et al (2013) article 5. Professional writing (5 points) a. Clearly structured professional report b. Clearly labelled and professional tables and graphics c. Concise writing

Reference: Molofsky, L.J., Connor, J.A., Wylie, A.S., Wagner, T. and S.K. Farhat. (2013). Evaluation of Methane Sources in Groundwater in Northeastern Pennsylvania. Groundwater, 51(3): 333-349