Week_2_Problem_Set

Nguyen Tien Anh Quach

2024-02-01

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:/GitHub Projects/enec-562/Week 2 Lab - t test"
```

Load data:

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	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	1379	684.2574	0.6	3132.928	15.830	6.230358
Near	322	795.0171	5.9	4086.957	25.575	6.467009

Range for near sites can be reported here: 0.08, 4.3×10^4.

Range for far sites: 0.05, 3.9×10^4.

Range for upland sites: 0.05, 3.2×10^4.

Range for valley sites: 0.08, 4.3×10^4.

Range for all sites: 0.05, 4.3×10^4.

Now with only valley sites:

Proximity Category	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	670	1186.406	1.3	4058.772	25.8075	4.510468
Near	195	1225.604	19.0	5172.061	25.4800	5.020043

Now with only upland sites:

Proximity Category	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	709	209.7305	0.4	1753.1023	2.25	12.798759
Near	127	133.8798	1.4	799.4312	25.69	8.959691

Now valley vs. upload, no matter the proximity:

Location	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Upland	836	198.2077	0.47	1644.111	3.565	13.326158
Valley	865	1195.2426	1.80	4331.548	25.780	4.790541

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.

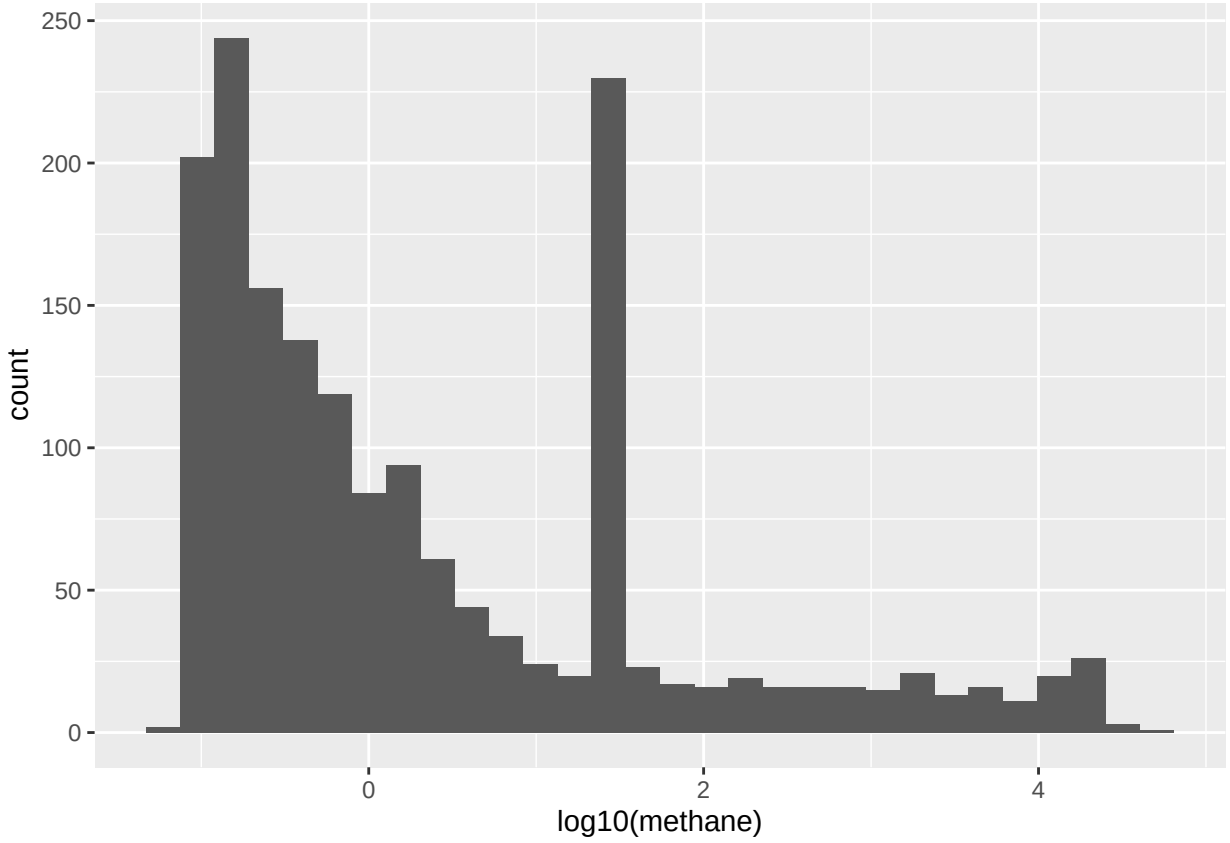
4. Interpret and discuss the results of each of the tests.
5. Examine and discuss the validity of the assumptions of your comparison tests. Remember to consider the transformation when interpreting the results of the transformed data set. Which of the tests are most valid?

Testing near vs far sites

```
## # A tibble: 234 x 7
##    proximity    ID methane    dl location is.outlier is.extreme
##    <chr>     <dbl>   <dbl> <dbl> <chr>    <lgl>      <lgl>
##  1 Far           2   17000     0 Valley   TRUE       TRUE
##  2 Far          13    1400     0 Valley   TRUE       TRUE
##  3 Far          16    1300     0 Valley   TRUE       TRUE
##  4 Far          18     230     0 Upland   TRUE       TRUE
##  5 Far          29      68     0 Upland   TRUE       TRUE
##  6 Far          32   18000     0 Valley   TRUE       TRUE
##  7 Far          56    9500     0 Valley   TRUE       TRUE
##  8 Far          63     140     0 Valley   TRUE       TRUE
```

```
##  9 Far          67     610     0 Valley   TRUE       TRUE
## 10 Far          78     440     0 Valley   TRUE       TRUE
## # i 224 more rows
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 methane      0.220 2.39e-64
```




```
## # A tibble: 1 x 3
##   variable             statistic  p.value
##   <chr>                    <dbl>    <dbl>
## 1 log(pafrack$methane)     0.859 1.60e-36
```

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic        p p.signif
##   <chr>   <chr>  <chr>  <int> <int>     <dbl>    <dbl> <chr>
## 1 methane Far    Near    1379   322    171369 8.21e-11 ****
```

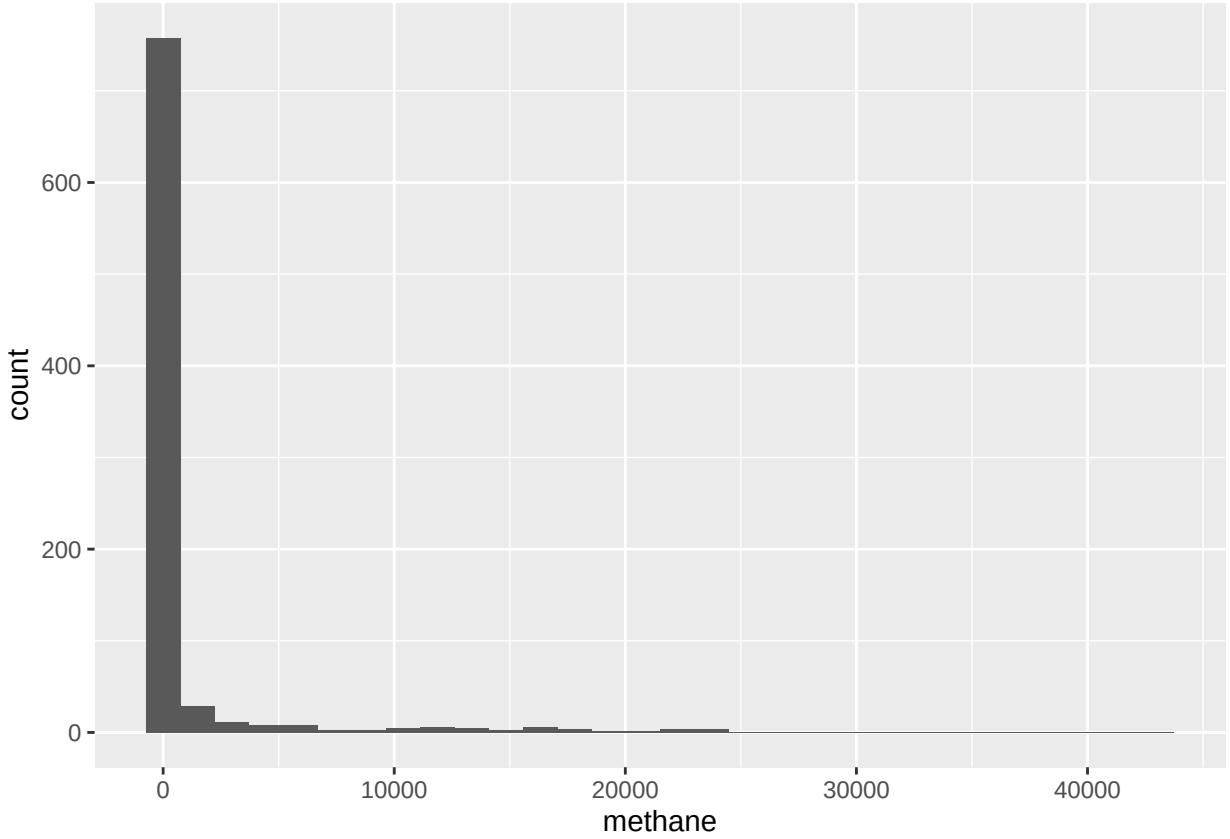
```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 methane Far    Near     0.155  1379   322 small
```

Testing near vs far for valley sites

```
## # A tibble: 158 x 7
##       ID methane    dl location proximity is.outlier is.extreme
```

```
##    <dbl>   <dbl> <dbl> <chr>    <chr>     <lgl>      <lgl>
##  1     2   17000     0 Valley   Far       TRUE       TRUE
##  2    13    1400     0 Valley   Far       TRUE       TRUE
##  3    16    1300     0 Valley   Far       TRUE       TRUE
##  4    32   18000     0 Valley   Far       TRUE       TRUE
##  5    56    9500     0 Valley   Far       TRUE       TRUE
##  6    63     140     0 Valley   Far       TRUE       TRUE
##  7    67     610     0 Valley   Far       TRUE       TRUE
##  8    78     440     0 Valley   Far       TRUE       TRUE
##  9    88   17000     0 Valley   Far       TRUE       TRUE
## 10    99   13000     0 Valley   Far       TRUE       TRUE
## # i 148 more rows
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 methane      0.309 8.20e-49
```




```
## # A tibble: 1 x 3
##   variable                  statistic  p.value
##   <chr>                         <dbl>    <dbl>
## 1 log(valley_sites$methane)     0.886 1.08e-24
```

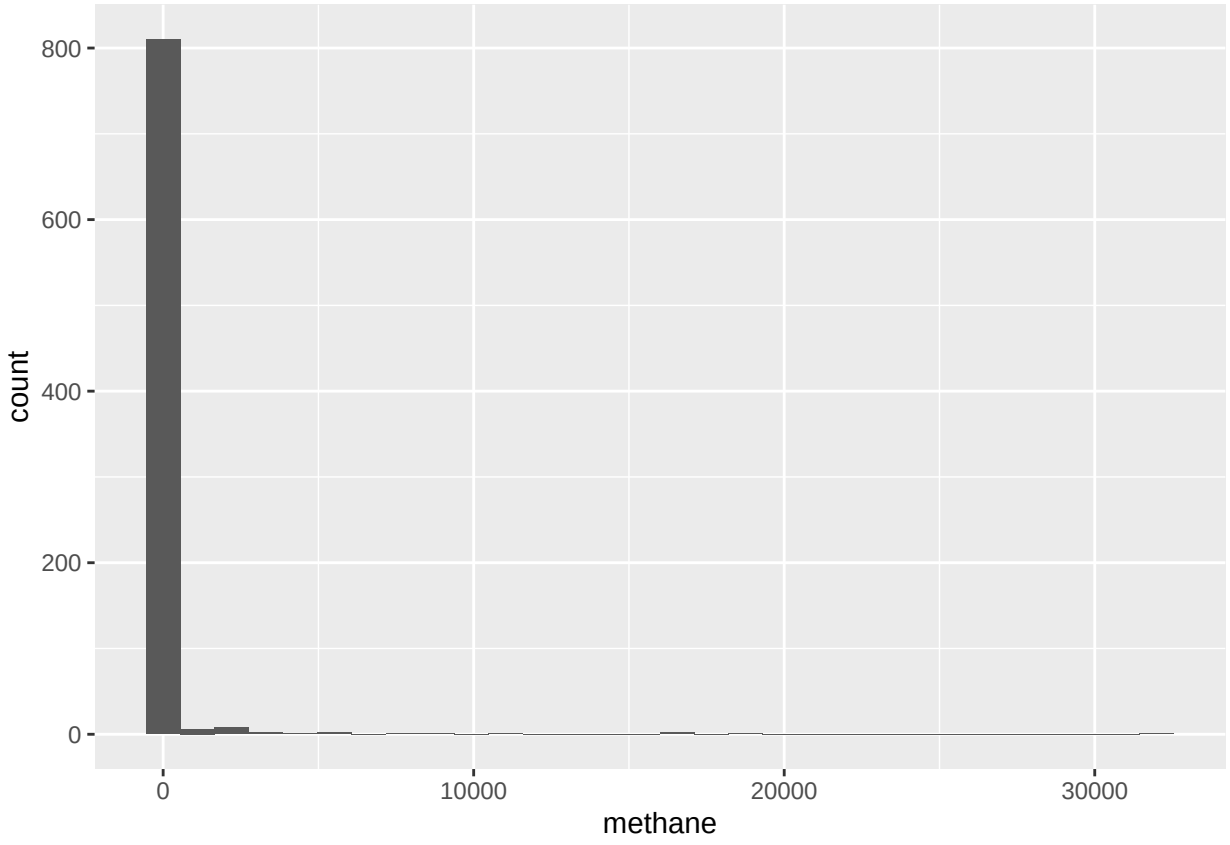
```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic        p p.signif
##   <chr>   <chr>  <chr>  <int> <int>     <dbl>    <dbl> <chr>
## 1 methane Far    Near     670   195    55564. 0.000727 ***
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 methane Far    Near     0.108   670   195 small
```

Testing near vs far for upland sites

```
## # A tibble: 178 x 7
##       ID methane    dl location proximity is.outlier is.extreme
##    <dbl>   <dbl> <dbl> <chr>    <chr>     <lgl>      <lgl>
##  1     3      26     1 Upland   Far       TRUE       TRUE
##  2     4      26     1 Upland   Far       TRUE       TRUE
##  3    17      28     0 Upland   Far       TRUE       TRUE
##  4    18     230     0 Upland   Far       TRUE       TRUE
##  5    27      57     0 Upland   Near      TRUE       TRUE
##  6    29      68     0 Upland   Far       TRUE       TRUE
##  7    31      26     1 Upland   Far       TRUE       TRUE
##  8    97     610     0 Upland   Far       TRUE       TRUE
##  9   114    4300     0 Upland   Far       TRUE       TRUE
## 10   124     430     0 Upland   Far       TRUE       TRUE
## # i 168 more rows
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 methane     0.0978 1.95e-52
```



```
## # A tibble: 1 x 3
##   variable                  statistic  p.value
##   <chr>                         <dbl>    <dbl>
## 1 log(upland_sites$methane)     0.844 6.92e-28
```

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic           p p.signif
##   <chr>   <chr>  <chr>  <int> <int>     <dbl>       <dbl> <chr>
## 1 methane Far    Near     709   127     32805 0.000000527 ****
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 methane Far    Near     0.169   709   127 small
```

Testing upland vs valley sites

```
## # A tibble: 1 x 8
##   .y.     group1 group2    n1    n2 statistic        p p.signif
##   <chr>   <chr>  <chr>  <int> <int>     <dbl>    <dbl> <chr>
## 1 methane Upland Valley   836   865    284413 2.35e-14 ****
```

```
## # A tibble: 1 x 7
##   .y.     group1 group2 effsize    n1    n2 magnitude
## * <chr>   <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 methane Upland Valley   0.185   836   865 small
```

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