Analysis of gas

2019-04-09

Data Prep

1. Dropped records with missing gas values at t1 and t2. Call the resulting data set `df`.
2. Separated `df` into two subsets:
   - `df_tiny`: gas < 1000 at t1 or t2.
   - `df_main`: gas >= 1000 at t1 and t2
3. Created long-format version:
   - `df_long`: long-format version of the full set `df`
   - `df_tiny_long`: long-format version of the subset `df_tiny`
   - `df_main_long`: long-format version of the subset `df_main`

Analyze the full set `df`

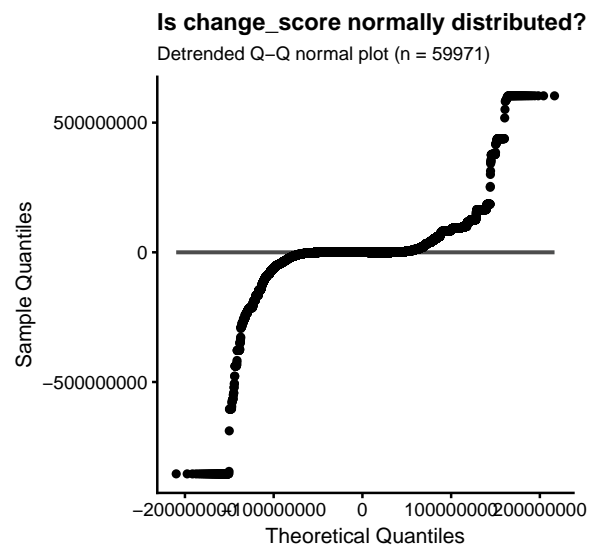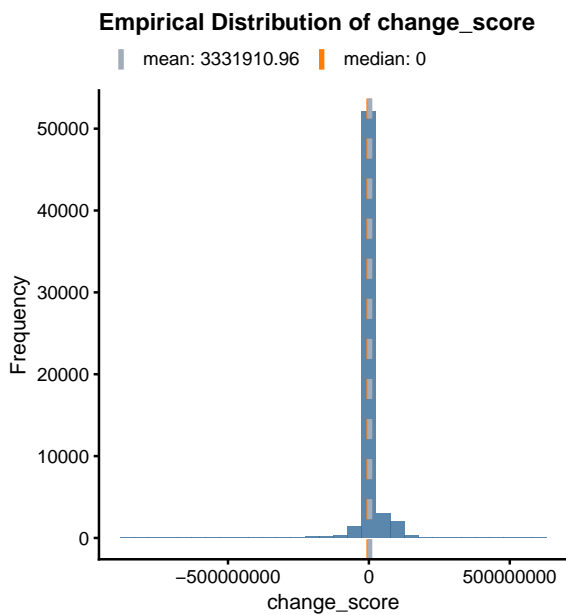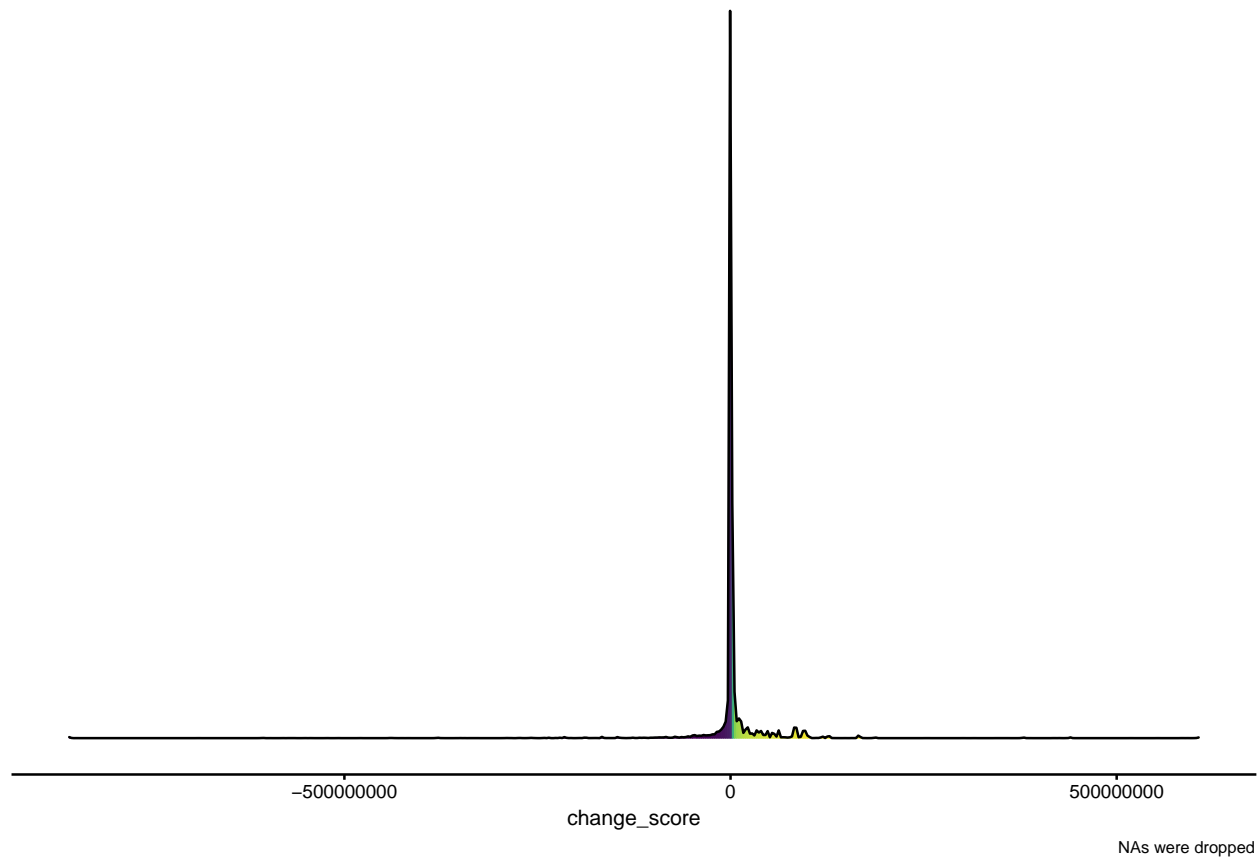Table 1: Sample Summary Statistics of gas

| time	n_tribes	n	mean	median	std	skewness	kurtosis	SEM
t1	184	59971	6330002	0	39070534	14.284540	263.2441	159543.3
t2	184	59971	9661913	0	33655549	8.579463	115.1978	137431.4

Q1. Is there a difference between t1 and t2?

Descriptive Analysis

First we look at the distribution of the change scores between t2 and t1 (t2 - t1). The following density plot, histogram and detrended qqnormal plot show that its distribution is symmetric but not normal, with very long tails extending far in the positive and negative directions. A normal distribution would produce a detrended qqnormal plot with most of the data points randomly scattered around the line `y=0` and within the grayish blue confidence band. There's a second smaller peak along the right tail, but the left tail is longer than the right tail. The sample median is 0. The sample mean is a large positive number, but is relatively small in relation to the extreme tail values. This can be seen from the histogram where the silver dashed line (mean) almost overlap with the orange dashed line (median = 0). As a result, we cannot tell from these plots if the mean gas change from t1 to t2 is significantly different from zero.

gas

change_score

Empirical Distribution of change_score

▌ mean: 3331910.96   ▌ median: 0

Is change_score normally distributed?
Detrended Q–Q normal plot (n = 59971)

Next we look at the distributions of (log1p transformed) gas values at t1 and t2 side by side with regular boxplots (cannot draw notched boxplot because of the zero values). We see the two box bodies overlap, with t2 mean slightly above t1 mean. So we cannot tell if there's an significant difference in the population mean gas values between t2 and t1.
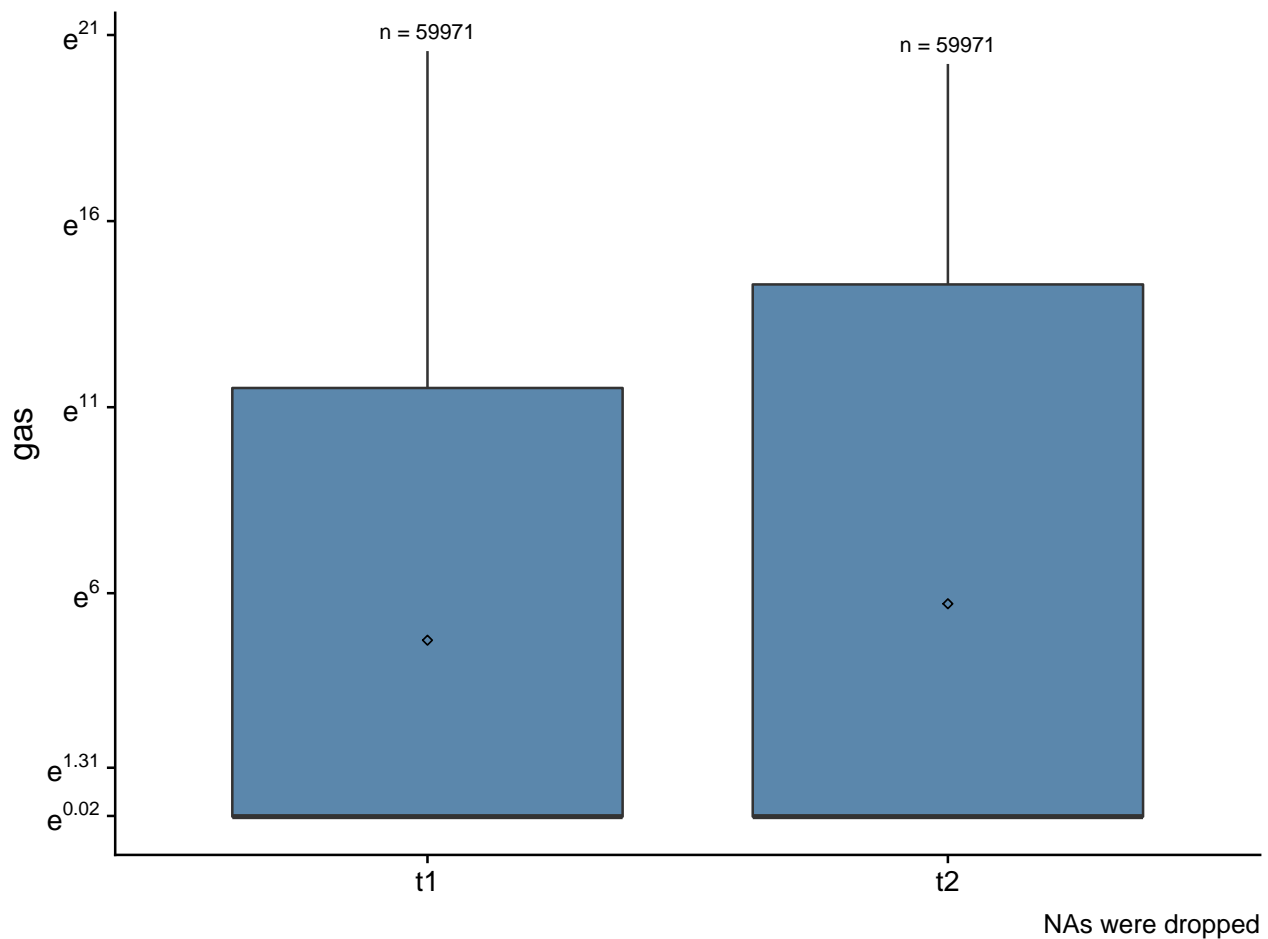
Statistical Analysis

To test if there's a difference between the population mean gas values at t1 and t2, we ran an one-way repeated measure ANOVA. We also ran a linear mixed model with tribe as random effect, but it failed to converge due to the underlying data. We chose these methods because at each time point (t1 or t2), there are multiple gas values for each tribe. These methods account for the within-tribe correlations. From the ANOVA output, we see that `time` is significant with a p-value of 0.01. So we conclude there's a significant difference between the mean gas values at t2 and t1.

When reading the output from ANOVA and linear mixed model, you want to focus on the reported p-value of the term `time`. It tells you the probability of observing a difference between t1 and t2 as extreme as in the sample data due to chance or randomness. If it's small, it's more likely that the observed difference is not due to chance. To decide how small is "small", the convention is to compare the p-value with 0.05. (But you don't have to use 0.05, it's really your choice. For example, 0.01 or 0.1 are also commonly used in different applications). If it's less than 0.05, we say the observed difference between t1 and t2 in the sample is likely not due to chance and hence can be generalized to the entire population. In other words, the difference is (statistically) significant. Otherwise when the p-value is greater than 0.05, we say the observed difference between t1 and t2 in the sample is likely a fluke and cannot be generalized to the entire population. In other words, the difference is not (statistically) significant.

1-way Repeated Measure ANOVA Output:

```
Error: tribe
          Df              Sum Sq          Mean Sq F value Pr(>F)
```

```
Residuals 183 21639544130163552256 118248875028216128

Error: tribe:time
          Df              Sum Sq          Mean Sq F value Pr(>F)
time       1  332887946425953792 332887946425953792   6.621 0.0109 *
Residuals 183 9201458616311779328   50281194624654528
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
             Df                 Sum Sq          Mean Sq F value Pr(>F)
Residuals 119574 128631374282389798912 1075747020944267
```

Q2. How are t1 and t2 related?

We first calculated the mean and median gas values of each tribe at t1 and t2 respectively. The reason why we also looked at the median is because the median is extreme-value resistent while the mean is heavily influence by outliers. We then made a scatterplot of (log1p transformed) t2 means vs. t1 means, and another scatterplot of (log1p transformed) t2 medians vs. t1 medians. These scatterplots showed medium positive linear relationships, but the zero values weakened the strength of the linear relationships:

- **medium**: the tighter the dots, the stronger the correlation.
- **positive**: upward slanted trend from bottom left corner to upper right corner. Or y tends to increase as x increases.

Finally, we ran linear regressions to quantify these relationships. For the log1p transformed mean values, we obtained a r-squared value of 0.4, which translates to a correlation of 0.63 (the squared root of 0.4), i.e., the correlation between the log1p transformed mean values at t1 and t2 is 0.63. The slope of the line is 0.639, meaning that for every 1000-unit (or 6.909-unit in log1p scale) increase in gas production at t1, we can expect a 76.7-unit (or 4.353-unit in log1p scale) increase at t2. This is statistically signicant by the tiny p-value. A similar interpretation can be done for the median values.

gas

n = 184

$y = -0.547 + 0.639\, x,\ R^2 = 0.4$

$p-\text{value} < 0.001$

Mean value at t2

Mean value at t1

Each dot is a tribe

gas

n = 184

$y = 2.38 + 0.534\,x,\ R^2 = 0.21$

$p-\text{value} < 0.001$

Median value at t2

Median value at t1

Each dot is a tribe