Analysis of gas

2019-03-25

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 subset `df_main`

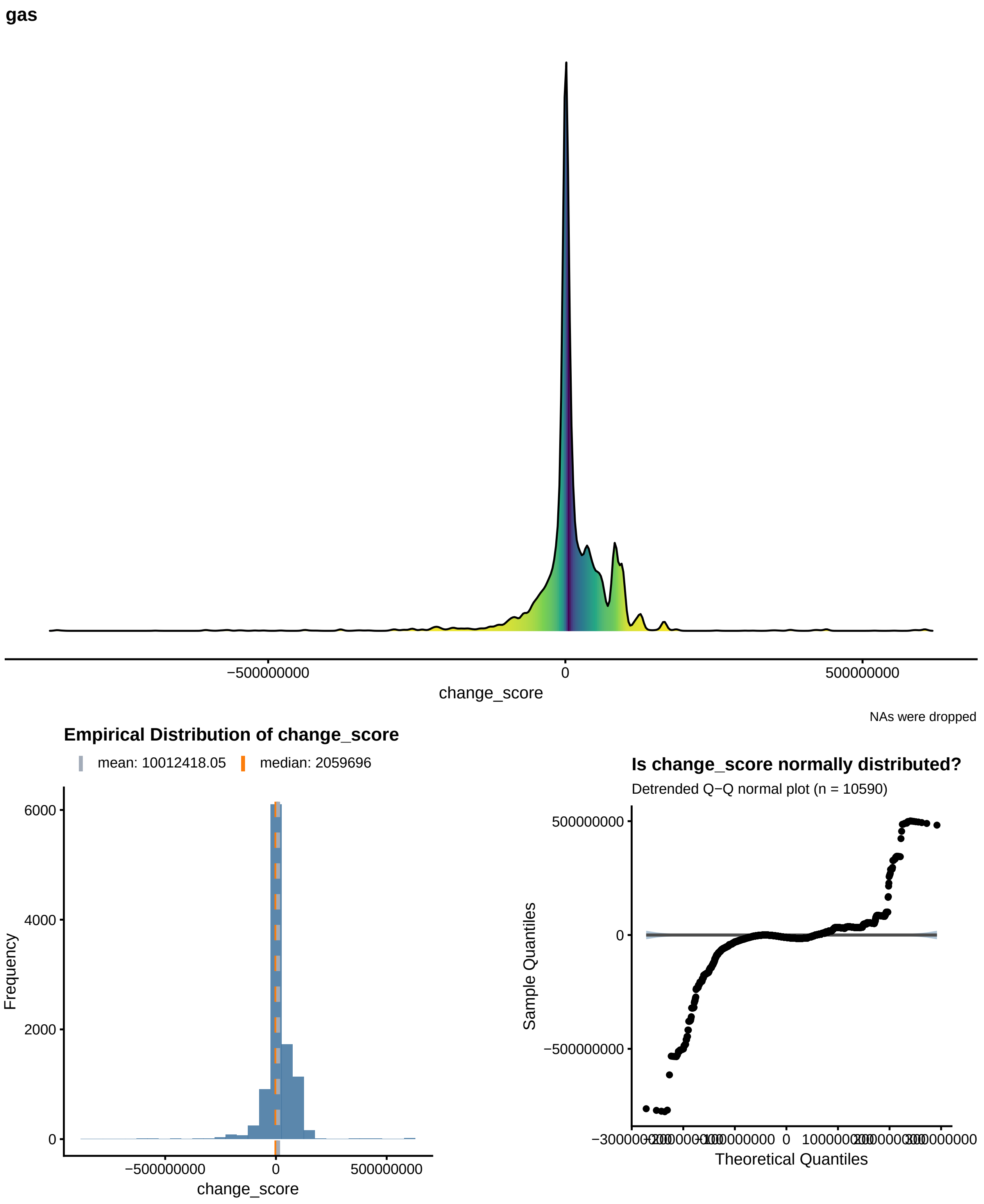
Table 1: Sample Summary Statistics of gas

time	n_tribes	n	mean	SEM
t1	59	10590	20798338	546722.2
t2	59	10590	30810756	498841.0

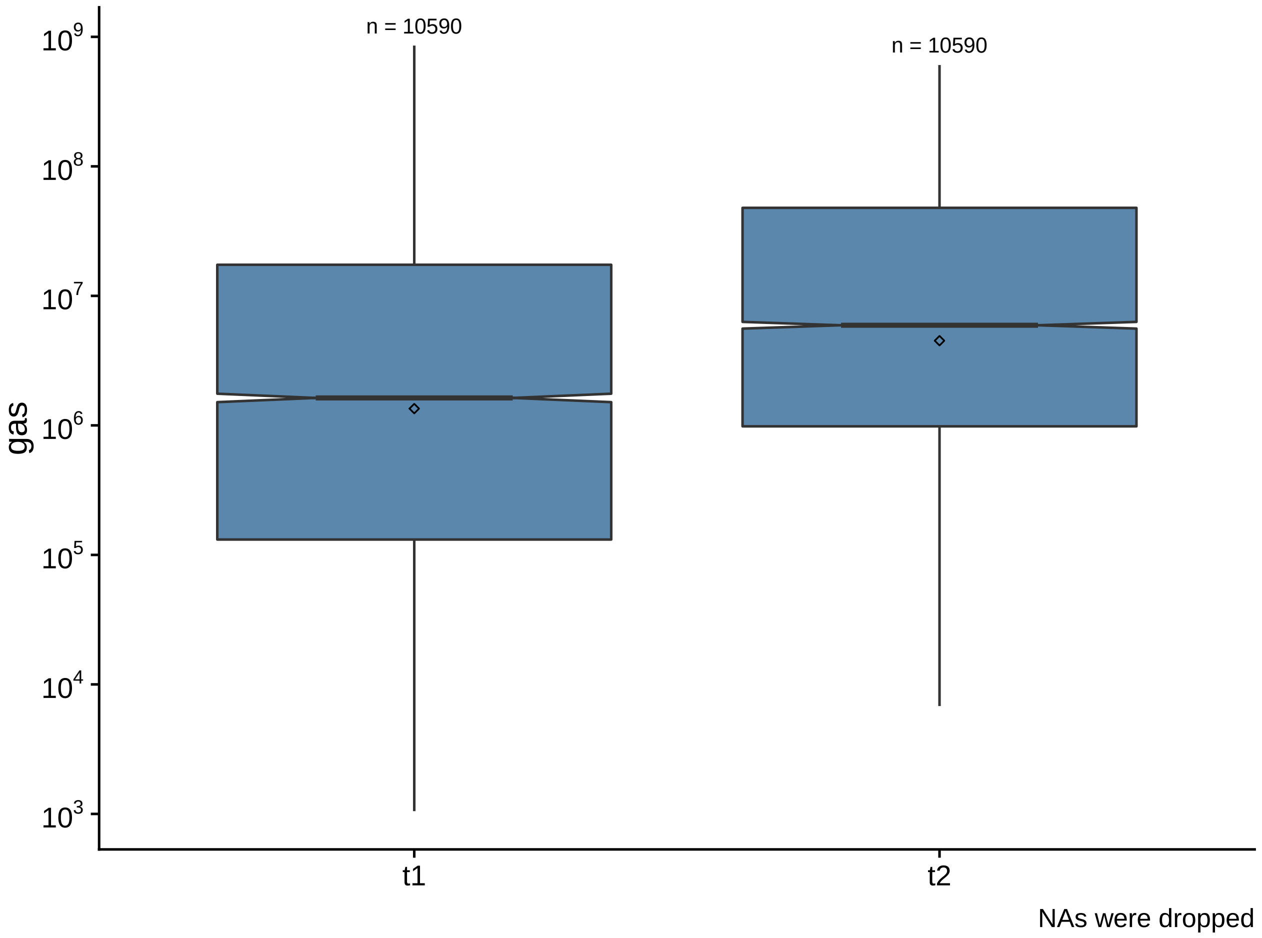
Q1. Is there a difference between t1 and t2?

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 big sample mean and median values, even when compared against the extreme tail values, suggest that the population mean gas change from t1 to t2 is significantly different from zero.



Next we look at the distributions of (log10 transformed) gas values at t1 and t2 side by side with notched boxplots. We see the two box bodies do not completely overlap. The t2 box body is taller, and has a bigger mean (represented by diamond shapes) than t1. Plus, the notches around the medians do not overlap at all. These all suggest a 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, as well as a linear mixed model with tribe as random effect. 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. We see that both methods give consistent result: a highly significant term `time` with a tiny p-value. 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 58 9134918442014501888 157498593827836224

Error: tribe:time
          Df              Sum Sq            Mean Sq F value  Pr(>F)
time       1  530815887710806144 530815887710806144   10.49 0.00198 **
Residuals 58 2933604202025125376  50579382793536648
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
             Df               Sum Sq          Mean Sq F value Pr(>F)
Residuals 21062 49354516266428915712 2343296755599132
```

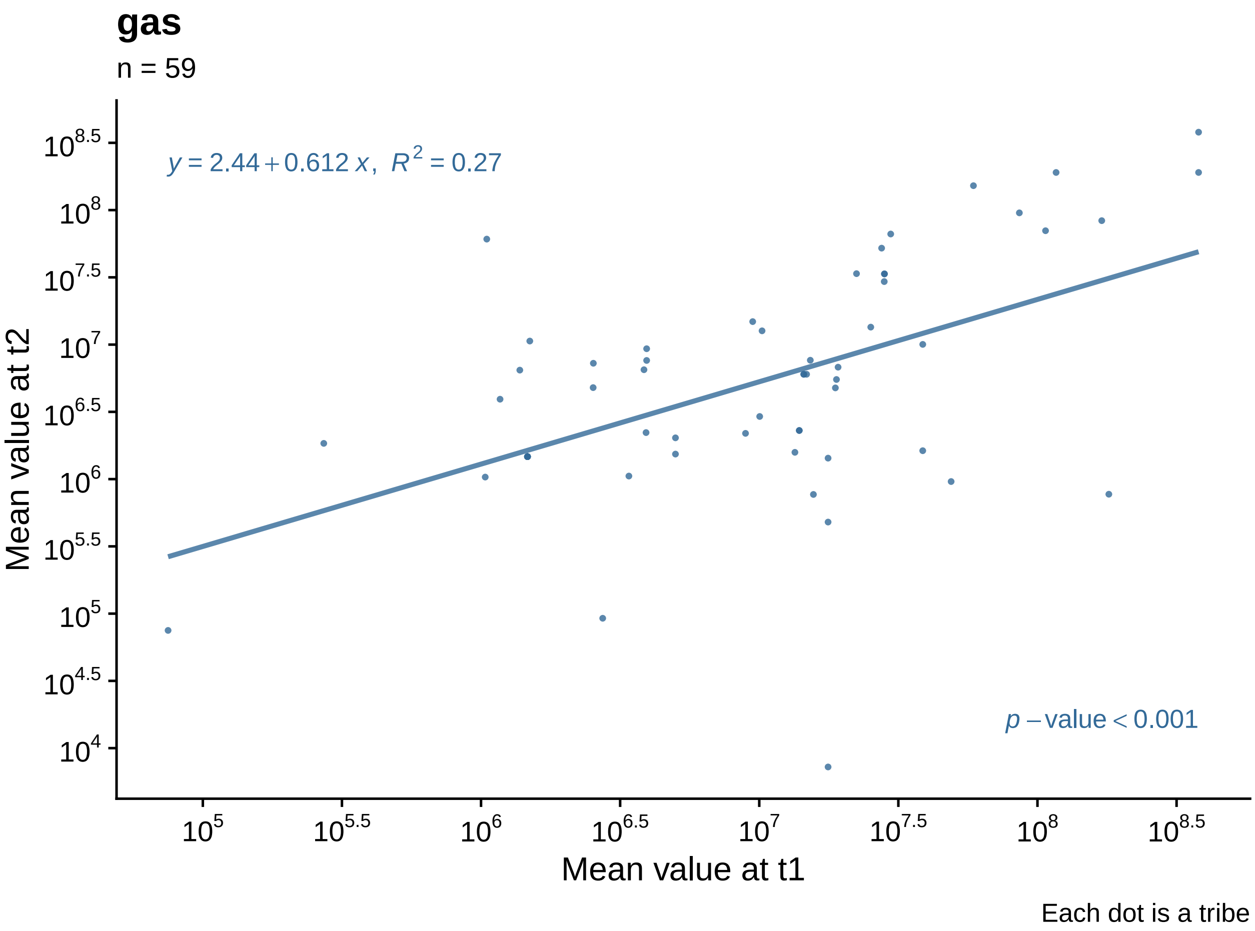
```
Linear Mixed Model Output:
            numDF denDF   F-value p-value
(Intercept)     1 21120  17.06441  <.0001
time            1 21120 214.32729  <.0001
```

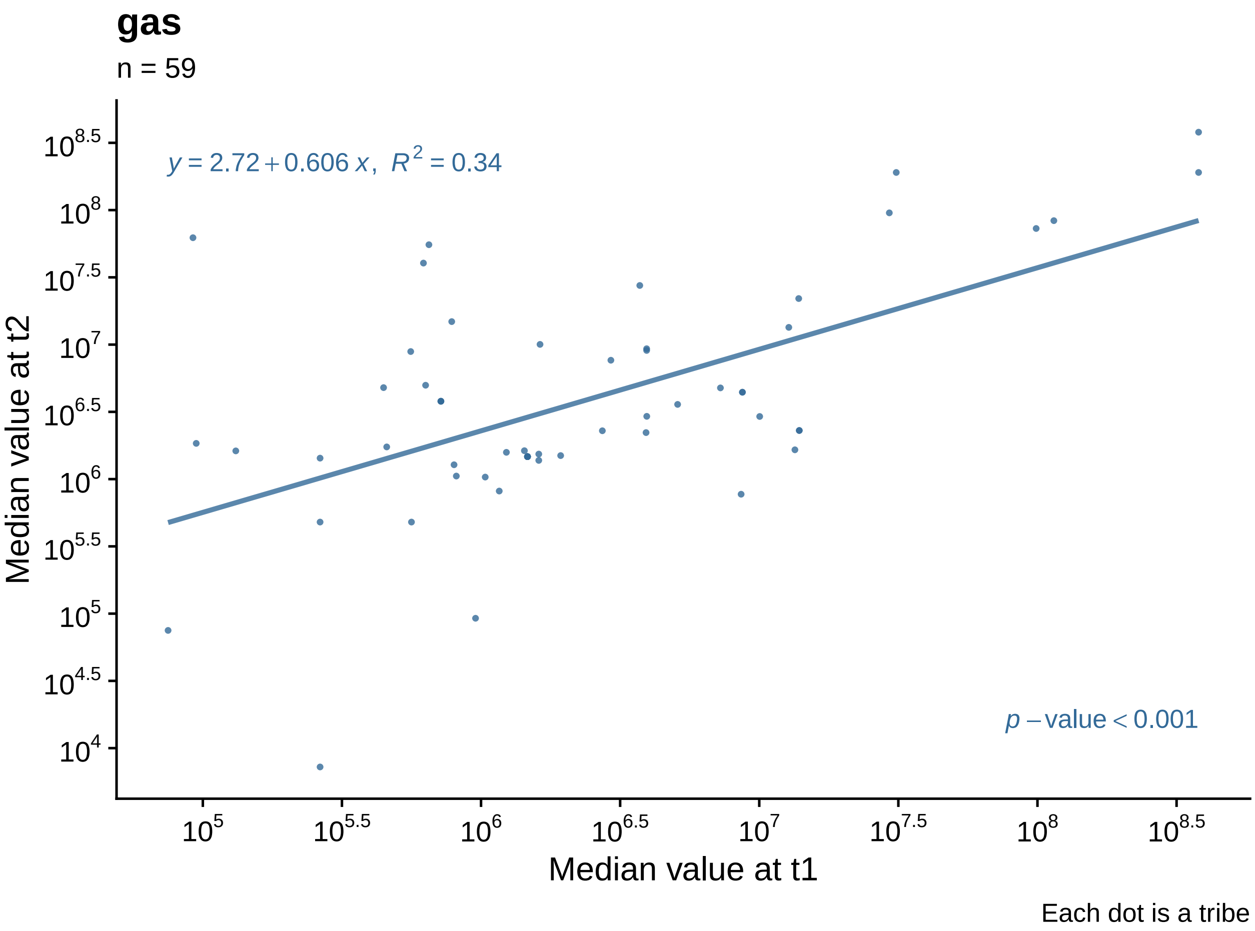
Q2. How are t1 and t2 related?

We first calculated the mean and median gas values of each tribe at t1 and t2. 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 the (log10 transformed) t2 means vs. t1 means, and another scatterplot of the (log10 transformed) t2 medians vs. t1 medians. These scatterplots showed medium positive 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 log10 transformed mean values, we obtained a r-squared value of 0.27, which translates to a correlation of 0.52 (the squared root of 0.27), i.e., the correlation between the log10 transformed mean values at t1 and t2 is 0.52. The slope of the line is 0.612, meaning that for every 1000-unit (or 3-unit in log10 scale) increase in gas production at t1, we can expect a 68.5-unit (or 1.836-unit in log10 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 = 59
$y = 2.44 + 0.612\,x,\ R^2 = 0.27$
$p - \text{value} < 0.001$
Mean value at t2
$10^{8.5}$
10^{8}
$10^{7.5}$
10^{7}
$10^{6.5}$
10^{6}
$10^{5.5}$
10^{5}
$10^{4.5}$
10^{4}
10^{5}
$10^{5.5}$
10^{6}
$10^{6.5}$
10^{7}
$10^{7.5}$
10^{8}
$10^{8.5}$
Mean value at t1
Each dot is a tribe


gas
n = 59
$y = 2.72 + 0.606\,x,\ R^2 = 0.34$
$p - \text{value} < 0.001$
Median value at t2
$10^{8.5}$
10^{8}
$10^{7.5}$
10^{7}
$10^{6.5}$
10^{6}
$10^{5.5}$
10^{5}
$10^{4.5}$
10^{4}
10^{5}
$10^{5.5}$
10^{6}
$10^{6.5}$
10^{7}
$10^{7.5}$
10^{8}
$10^{8.5}$
Median value at t1
Each dot is a tribe