Analysis of amen_rank

2019-04-10

Data Prep

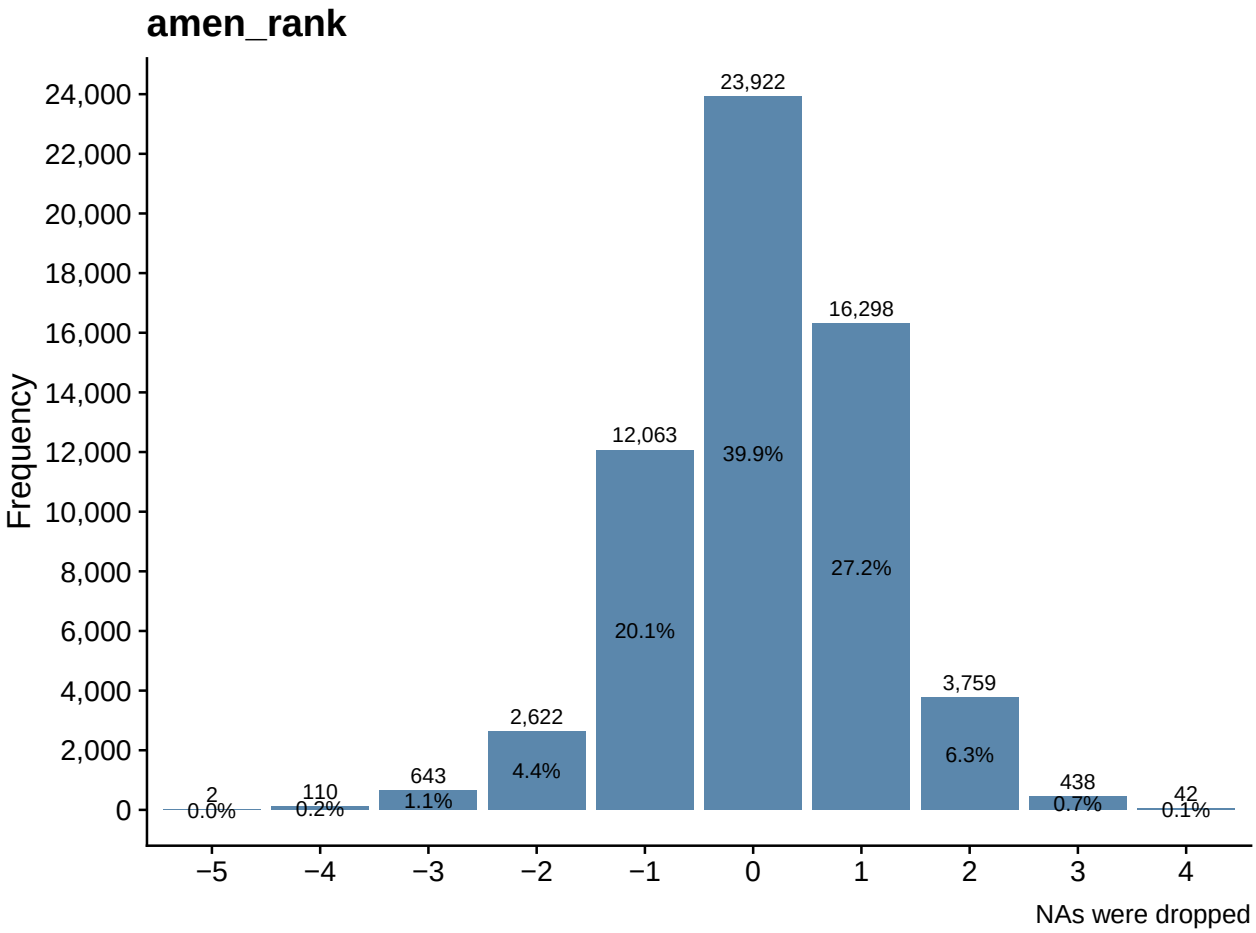
1. Dropped records with missing amen_rank values at t1 and t2. Call the resulting data set `df`.
2. Created long-format version:
    - `df_long`: long-format version of the full set `df`

Analyze the full set `df`

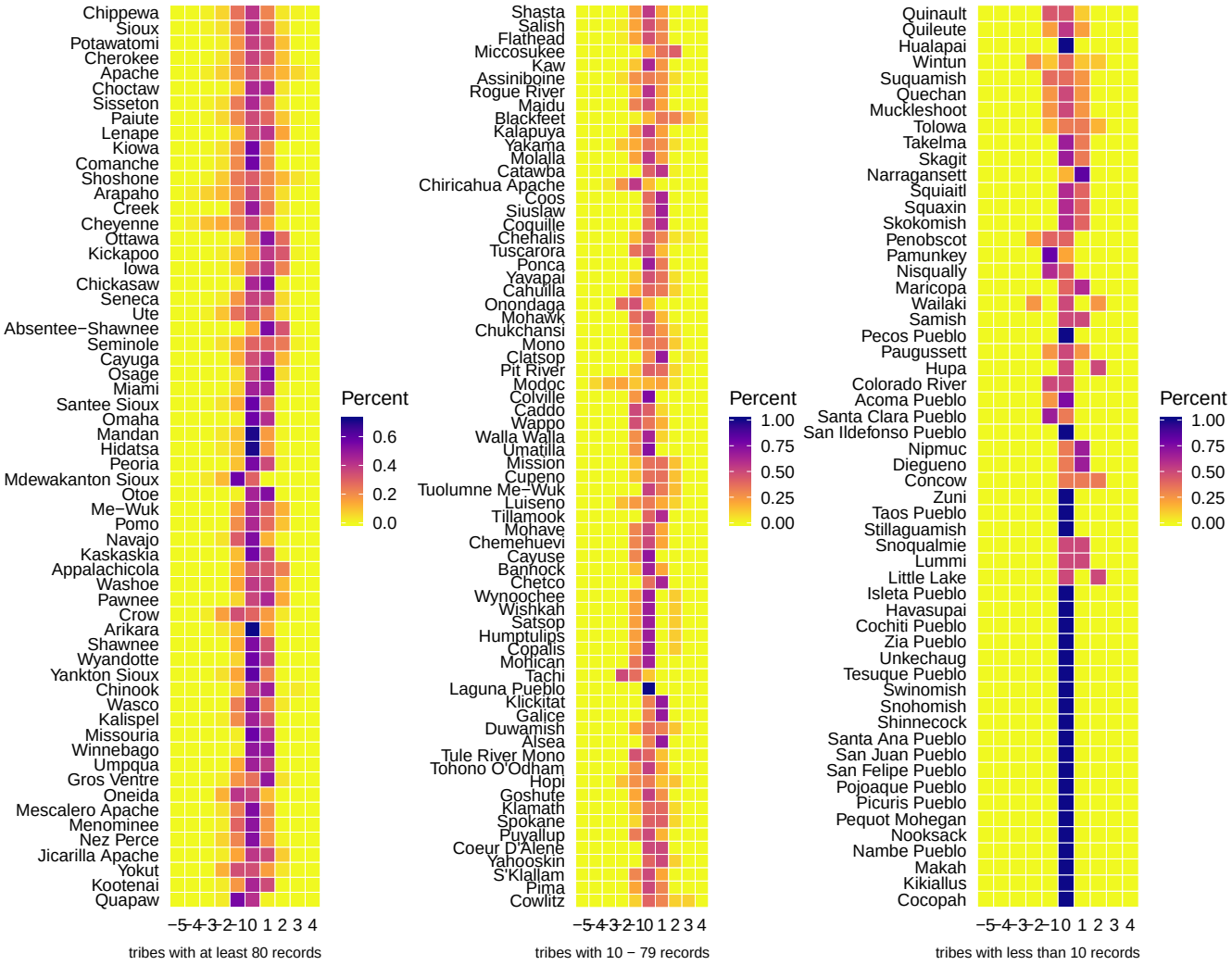
Q1. Is there a difference between t1 and t2?

Descriptive Analysis

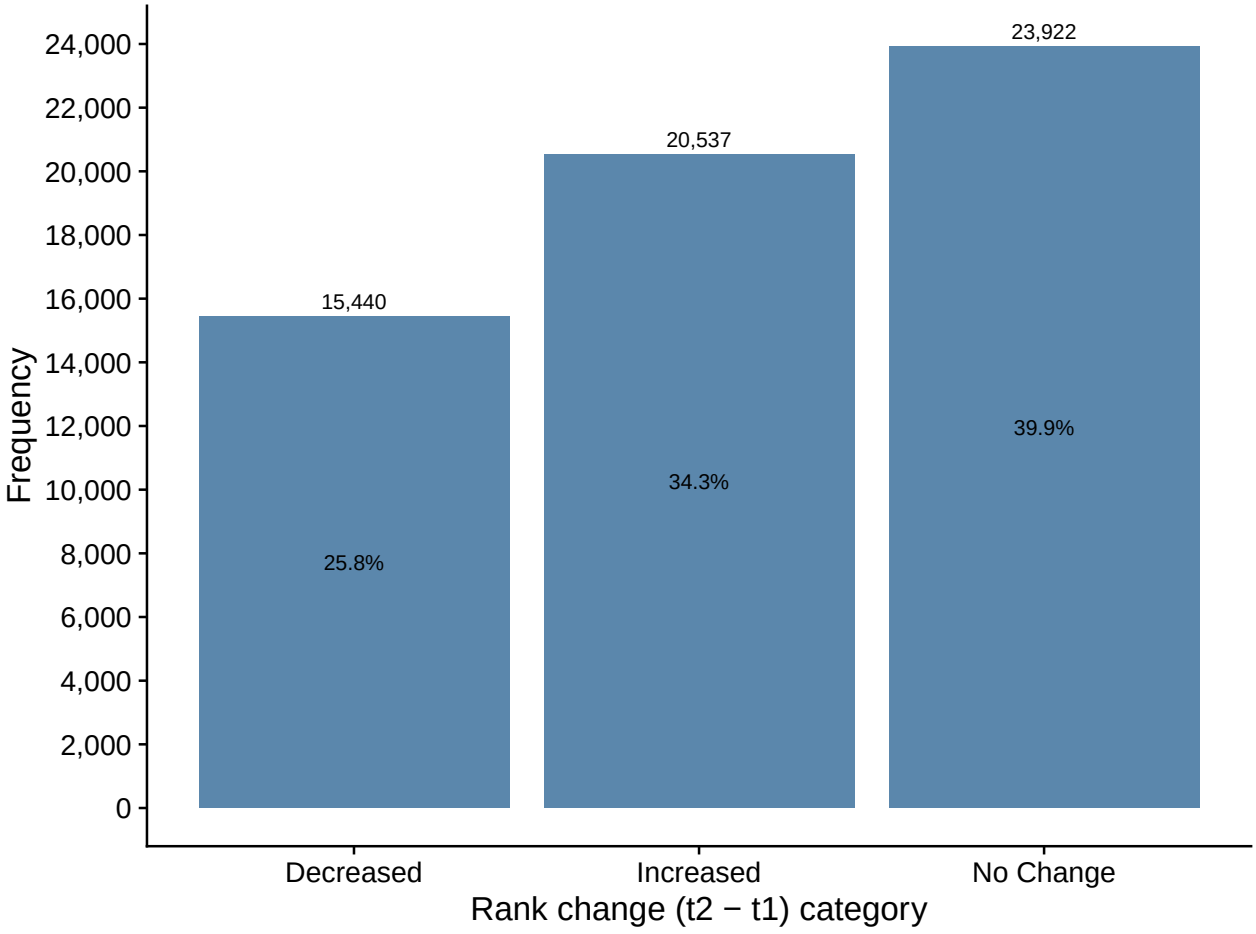
First we look at the overall distribution of the change scores (t2 - t1) without breaking down by tribes. Because the raw values at t1/t2 are rank scores, the change scores take on a finite number of values, namely -5, -4, -3, -2, -1, 0, 1, 2, 3, 4. The following bar chart shows their frequency counts. Our first impression is the chart is rather bell-curvish looking. We see that about 40% of the observations have a change score of 0, and about 87% of the observations have a change score of -1, 0 or 1. There're slightly (7%) more observations with rank increased by 1 than decreased by 1. There're slightly (2%) more observations with rank increased by 2 than decreased by 2. There're only small number of observations with rank increased or decreased by 3 or more.



Next we look at the distribution of the change scores within each tribe. The following heatmaps show that within each tribe, the bulk of the change scores also centers around -1, 0 and 1.



Finally, we group the sample records into three cohorts based on change score: “No Change”, “Increased” and “Decreased”. We then look at the distribution of these cohorts in a bar chart.



Statistical Analysis

The sample percentages of "No Change", "Increased" and "Decreased" are 39.94%, 34.29%, and 25.78%.

Let's test if there is some difference among the proportions of "No Change", "Increased" or "Decreased" in the population. Our null hypothesis is that there is an equal distribution (i.e., the proportions of "No Change", "Increased" and "Decreased" are 1/3 each).

We choose 0.05 significance level and run Chi-squared test. We conclude we don't have enough evidence to reject the null (Chi-squared test statistic = 3.05, pvalue = 0.22). In other words, it's perfectly likely that a population with equal proportions of "No Change", "Increased" and "Decreased" can produce a sample set like we have where the sample proportions are different.

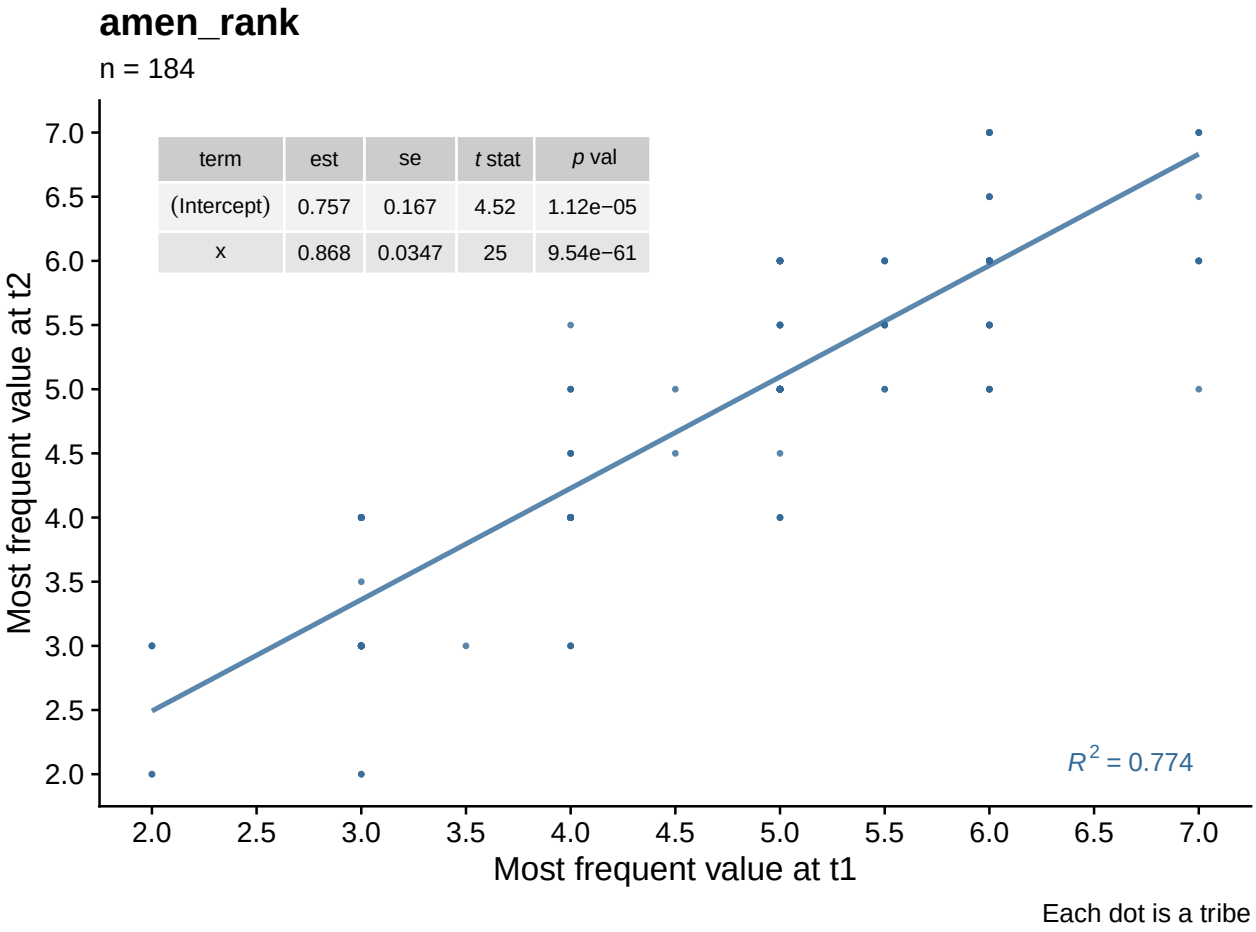
Next we want to test if the population mean of the change scores is different from zero. Our null hypothesis is that it is zero. We take 1000 bootstrapped samples and estimate the 95% confidence interval of the mean to be (0.086, 0.102). Because the 95% confidence interval sits entirely above zero, we conclude the mean change score is significantly different from zero (slightly bigger than zero).

Q2. How are t1 and t2 related?

We first calculate the most frequent (mode) amen_rank values of each tribe at t1 and t2. We use the mode instead of the mean or median because amen_rank is a ranking. We then make a scatterplot of the t2 vs. t1 modes, which showed strong positive linear relationships:

- **strong**: 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 use linear regression to quantify this relationship.



We get a r-squared value of 0.774, which translates to a correlation of 0.88 (by taking the squared root). This big positive correlation indicates the most frequent values at t1 and t2 tend to rise and fall together. The slope of the line is 0.868 with a 95% confidence interval of (0.799, 0.936). (Both the confidence interval and the tiny p-value indicate the slope is statistically significant and hence not zero.) This implies that for every unit increase in amen_rank at t1, we can expect an 0.868-unit jump give or take 0.069 at t2.