Analysis of risk

2019-03-25

Data Prep

1. Dropped records with missing risk 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`

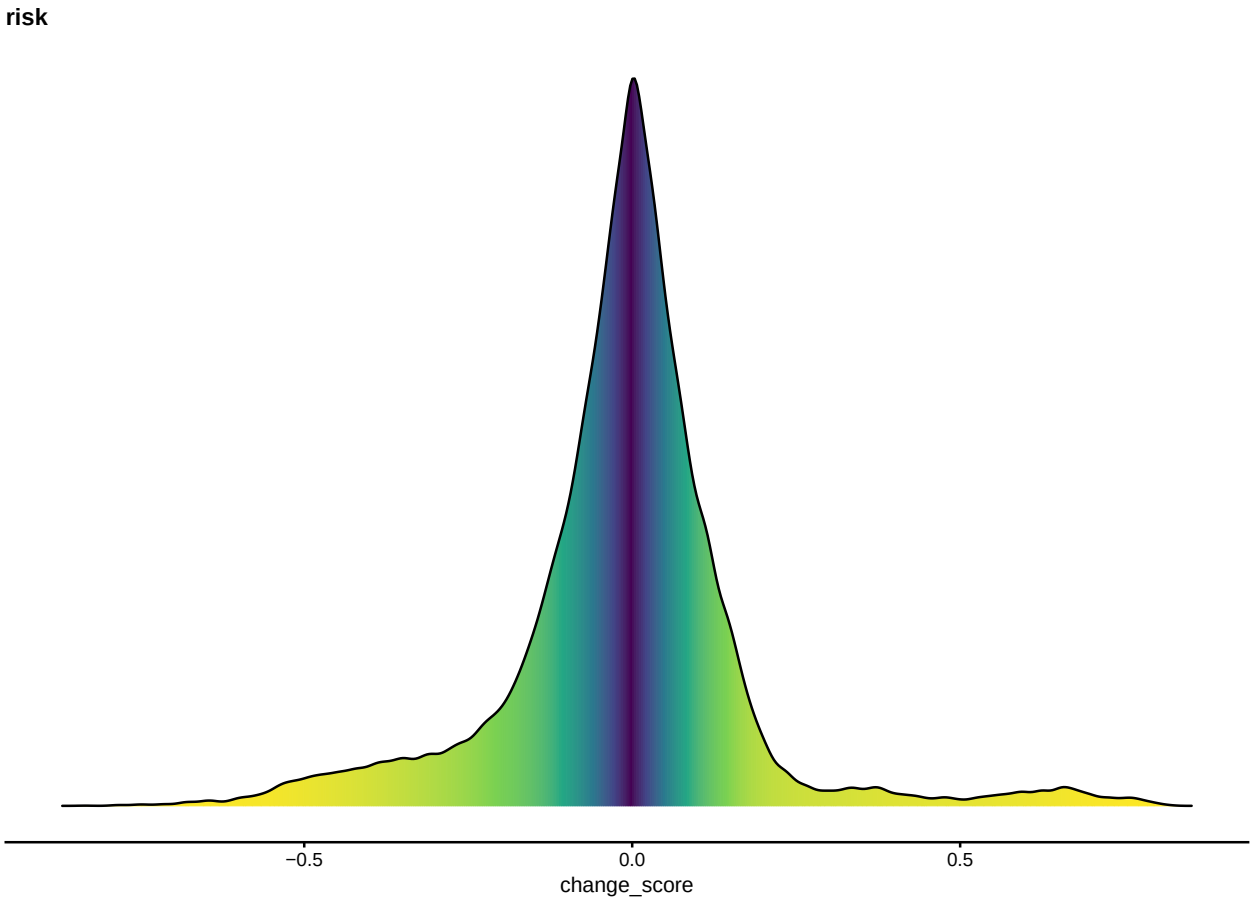
Table 1: Sample Summary Statistics of risk

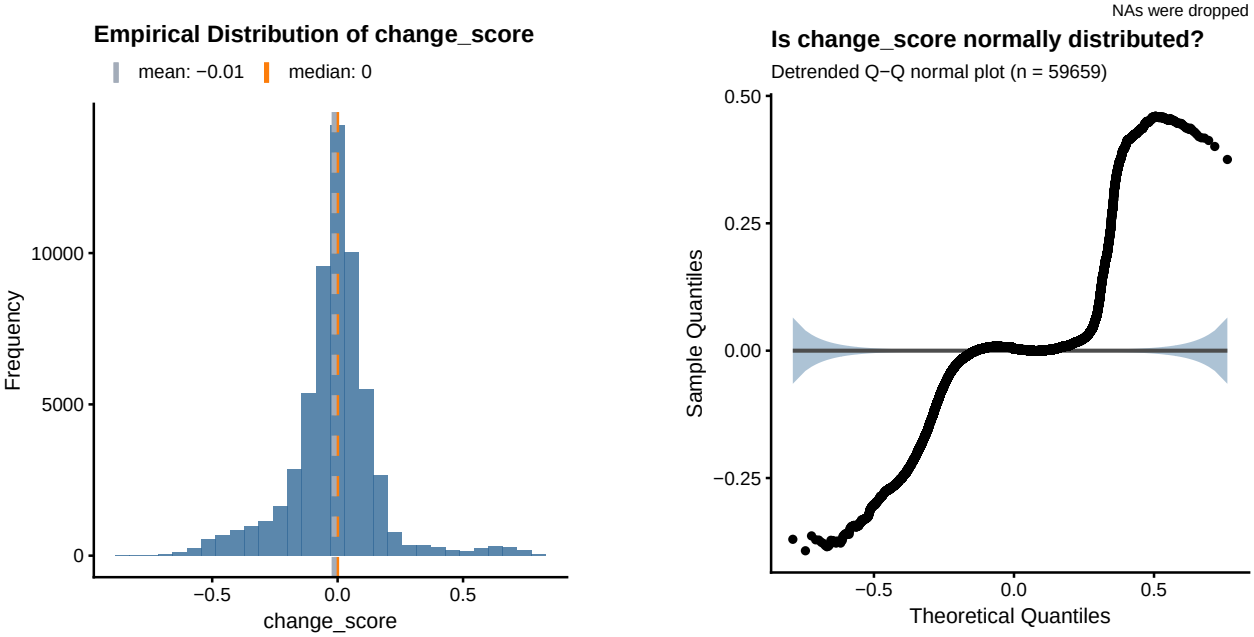
time	n_tribes	n	mean	SEM
t1	184	59659	0.2131825	0.0005526
t2	184	59659	0.1992809	0.0005339

Q1. Is there a difference between t1 and t2?

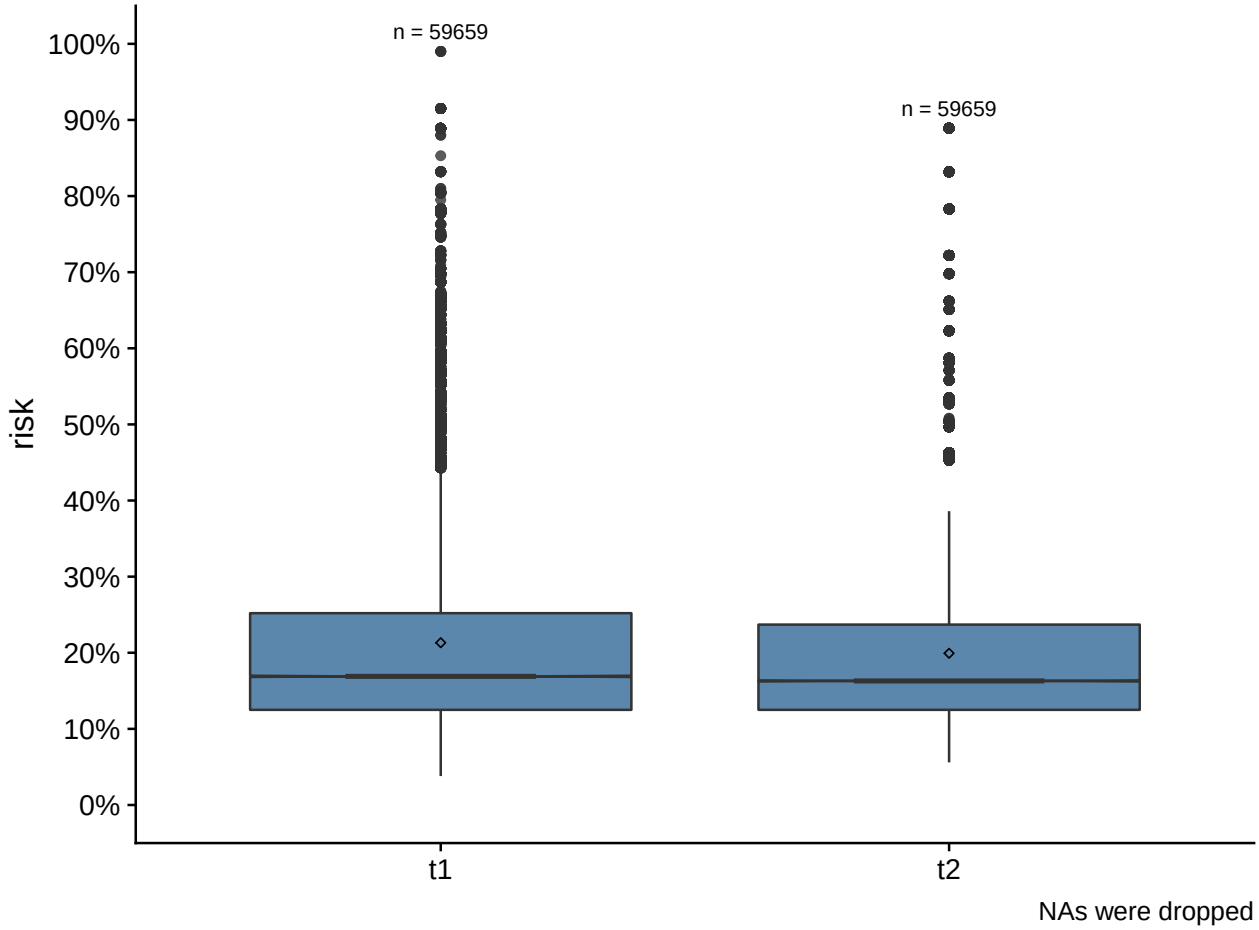
Descriptive Analysis

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 long tails extending in both 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. The left tail is fatter than the right tail, and this results a slightly negative sample average. It is unclear from these plots that if the mean risk change from t1 to t2 is significantly different from zero.





Next we look at the distributions of risk values at t1 and t2 side by side with notched boxplots. We see the two box bodies overlap, with t2 mean slightly below t1 mean. We also observe there are many outliers at t1 or t2, and these outliers inflate the sample averages. We cannot tell from these boxplots if there's a significant difference in the population mean risk values between t2 and t1.



Statistical Analysis

To test if there's a difference between the population mean risk 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 risk values for each tribe. These methods account for the within-tribe correlations. According to the ANOVA output, `time` is significant with a p-value of 0.00381. According to the linear mixed model output, `time` is significant with a p-value $<.0001$. Both methods show consistent results. So we conclude there's a significant difference between the mean risk 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
```

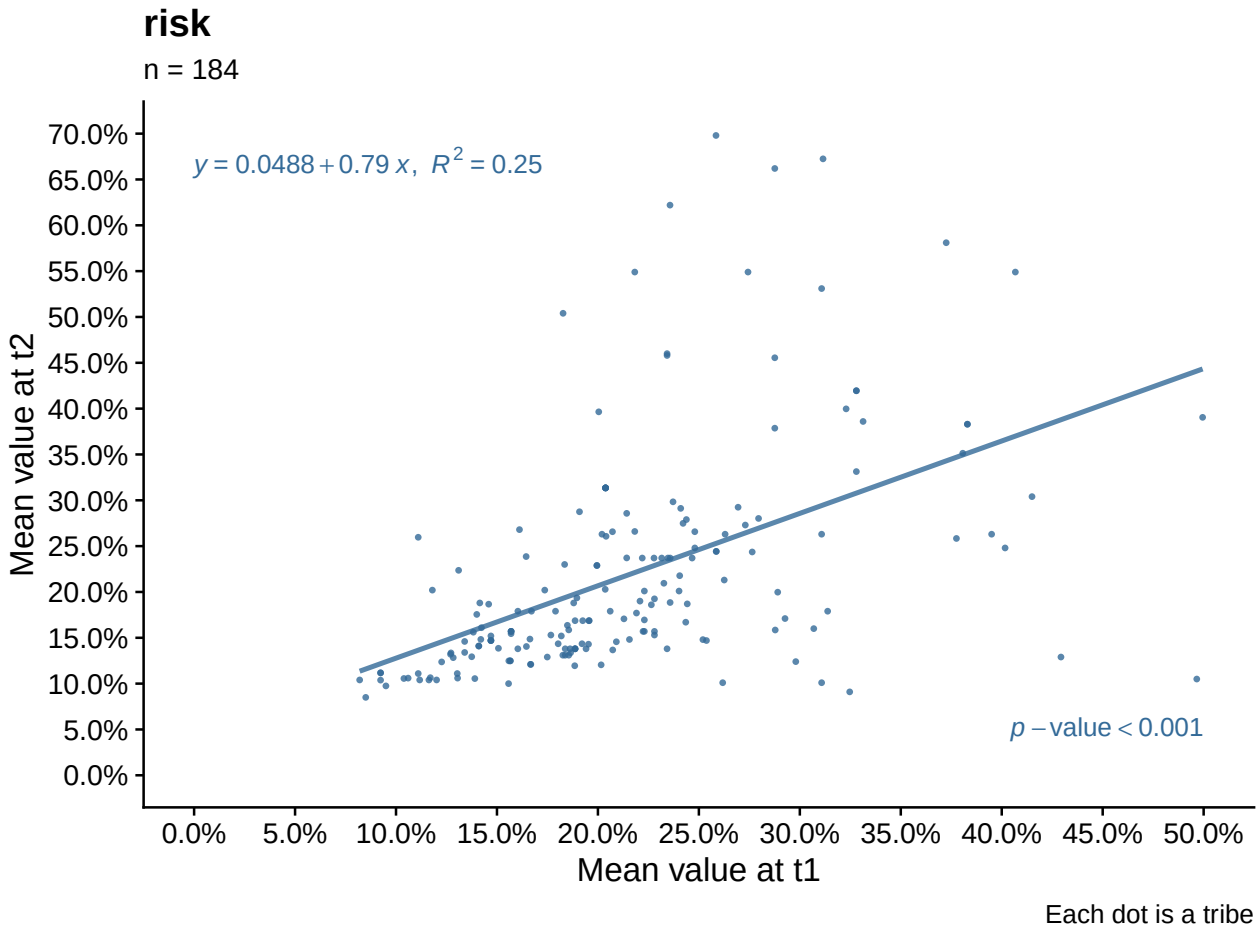
```
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 183  285.9   1.562

Error: tribe:time
           Df Sum Sq Mean Sq F value  Pr(>F)
time        1   5.76   5.765   8.592 0.00381 **
Residuals 183 122.78   0.671
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
              Df Sum Sq Mean Sq F value Pr(>F)
Residuals 118950   1692 0.01423

Linear Mixed Model Output:
            numDF  denDF   F-value p-value
(Intercept)     1 119133 1167.6515  <.0001
time            1 119133  378.3412  <.0001
```

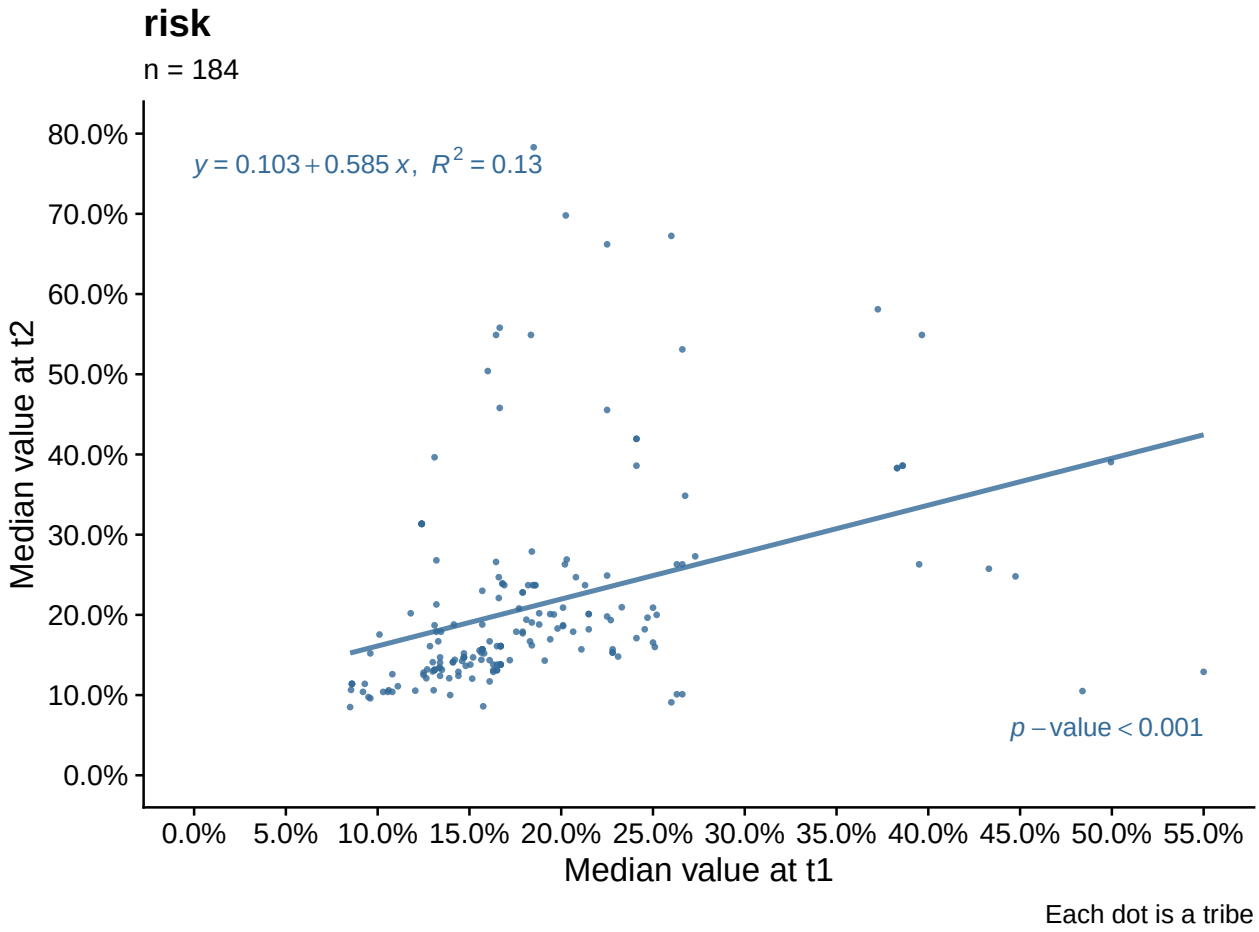
Q2. How are t1 and t2 related?

We first calculated the mean and median risk 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 t2 means vs. t1 means, and another scatterplot of the 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 mean values, we obtained a r-squared value of 0.25, which translates to a correlation of 0.5 (the squared root of 0.25). The slope of the line is 0.79, meaning that for every 100% increase in risk at t1, we can expect a 79% increase at t2. This is statistically signicant by the tiny p-value. A similar interpretation can be done for the median values.


risk
n = 184
$y = 0.0488 + 0.79\,x,\ R^2 = 0.25$
p – value < 0.001
Mean value at t2
Mean value at t1
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
30.0%
35.0%
40.0%
45.0%
50.0%
55.0%
60.0%
65.0%
70.0%
Each dot is a tribe


risk
n = 184
y = 0.103 + 0.585 x, R^2 = 0.13
p – value < 0.001
Median value at t2
Median value at t1
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0.0% 5.0% 10.0% 15.0% 20.0% 25.0% 30.0% 35.0% 40.0% 45.0% 50.0% 55.0%
Each dot is a tribe