**Statistical Models & Testing**

**Analysis Objective:** Is UCI's subreddit a more positive space compared to other UC universities?

**Groups:**

1. UC Irvine
2. UCLA
3. UC Berkeley
4. UC Davis
5. UC San Diego
6. UC Santa Barbara
7. UC Santa Cruz
8. UC Riverside
9. UC Merced
10. ANOVA Test
    1. Definition: ANOVA is a statistical test used to determine whether there are significant differences between the means of three or more groups.
    2. Our interest: Is there a difference in the means of the sentiment scores of the 9 UC subreddits?
    3. H\_0: There is no significant difference in the means of the sentiment scores of the 9 UC subreddits.
    4. H\_1: There is a significant difference in the means of the sentiment scores of the 9 UC subreddits.
    5. Significance Level: 0.05
    6. *Note:* If we reject the null, then we can move onto the two-paired sample t-test.
    7. *Note:* If we fail to reject the null, we should not move onto the two-paired sample t-test since there would be no convincing evidence that there is a significant difference in the means of the sentiment scores of the 9 UC subreddits.
11. Two-paired Sample T-test (Extended from ANOVA Test)
    1. Definition: Paired sample t-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated.
    2. Our interest: Is UCI more positive than other UC schools?
    3. H\_0: (pi\_uci) <= (pi\_ucla)
    4. H\_1: (pi\_uci) > (pi\_ucla)
    5. *Note:* The pi is the sample mean of the sentiment score of the corresponding UC school.
    6. *Note:* We need to conduct 8 paired sample t-tests.
       1. UCI vs. UCLA
       2. UCI vs. UC Berkeley
       3. …
12. Trend Test (Extended from ANOVA Test)
    1. ANOVA Trend Test: It can be used to analyze the data from longitudinal studies!
13. Longitudinal Study
    1. Definition: Longitudinal data focuses on multiple individuals at various time intervals.
    2. Advantage: Longitudinal studies allow researchers to follow their subjects in real time. It means we can better establish the real sequence of events, allowing us insight into cause-and-effect relationships. Longitudinal studies also allow repeated observations of the same individual over time.
    3. Our interest: Are sentiment scores of all UCs subreddits different over the time?
    4. Modeling Options:
       1. Simple Linear Regression Model
       2. Generalized Linear Regression Model
          1. EX. Y\_i = B\_0 + B\_1\*I(UCLA) + B\_2\*(UCB) + … + B\_9\*I(2011) + …
             1. UCI & 2010: Base Groups
       3. Generalized Linear Mixed Effect Regression Model (this would be the best)
          1. Advantage: Allow random intercept & random slopes
    5. Testing Options:
       1. T-test: To test the null hypothesis that the difference in the mean score between two time points has a mean value of zero
          1. Disadvantage: We can only see the “difference” between “two” time points in the “same” group.
          2. H\_0: (pi\_uci2019) >= (pi\_uci2020)
          3. H\_1: (pi\_uci2019) < (pi\_uci2020)
       2. Generalized Linear Regression Test on Generalized Linear Mixed Model
          1. We don’t have a covariate; We just have a covariate of “time.”
          2. H\_0: all betas and bs coefficients are equal to 0.
          3. H\_1: at least one of them is not equal to 0.
          4. *Note*: I am not sure if I approached correctly..! (from Ally)

ANOVA Extension