STUDENT RESEARCH CONFERENCE APPLICATION

KAYLA HORAK
MARCH 26, 2014
ADVISOR: CHRISTOPHER DANFORTH

We examine the difference in means between two samples larger than 100,000,000 from discrete distributions, in an attempt to determine whether there is a statistically significant difference. In particular, our samples consist of average happiness scores for words taken from the social network Twitter. We first use subsampling methods on one sample of words to evaluate the sensitivity of the average happiness measure to a change in the number of unique words, finding that the Spearman correlation between the original sample and the subsamples only drop below 0.7 once we remove 75% of the unique words. We then perform a randomization test on the difference in average happiness between the two samples (words used on Wednesdays versus words used on Fridays) and find a highly statistically significant difference between the two groups ($p < 0.0001$). Thus, a reduction in the number of unique words does not change the characteristics of the happiness distributions and the randomization test is a successful tool for determining statistically significant difference in two discrete distributions at large sample size.