Randomization Addendum to Population Analysis

Greg Pollock

10/9/2020

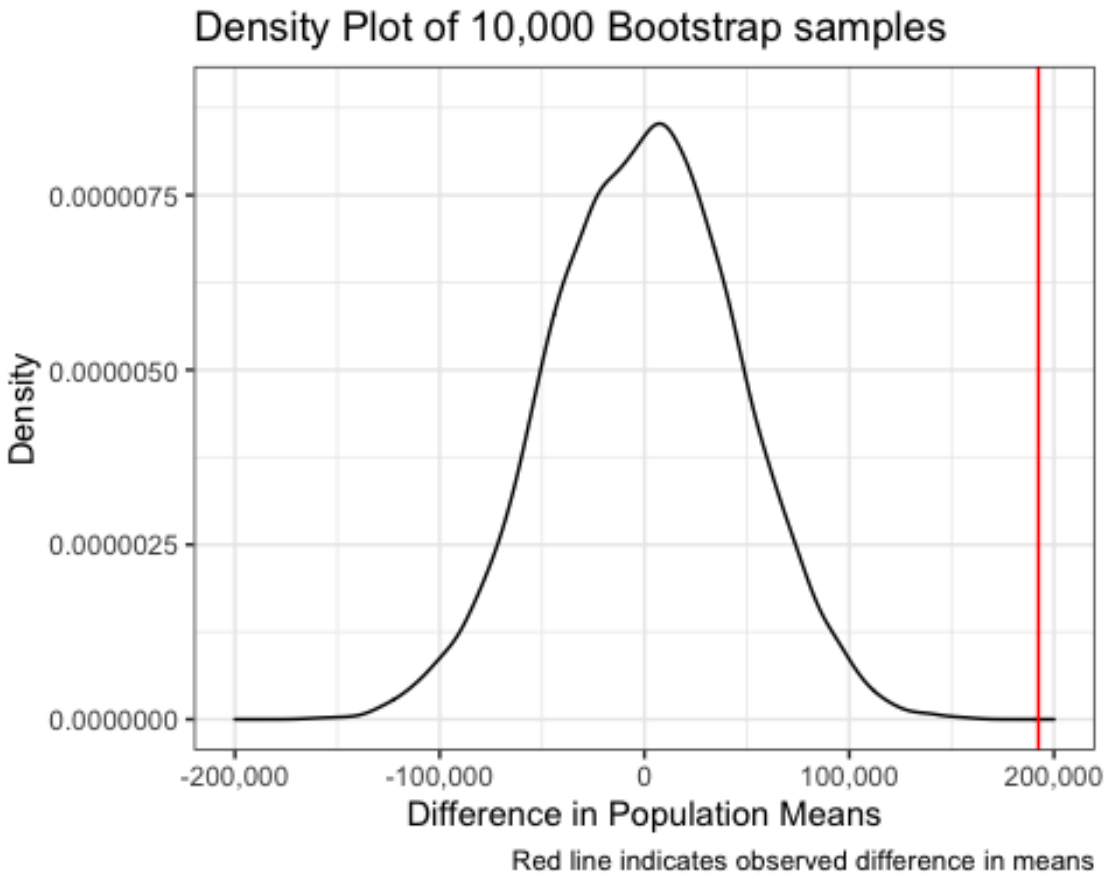
Introduction and Methodology

To take an alternate approach in analyzing this data in another way that doesn't rely on any assumptions of normality, we will conduct a randomized resampling analysis which will restrict the sample space to the observed values. In the case of this dataset there are $\binom{26}{13} = 10{,}400{,}600$ ways to group the population measurements into two groups. One resampling method is a permutation test which would be to consider every possible arrangement of the data values in two groups and find where the observed data fits in relation to this set of all possible arrangements. Since this entire sample space can't be easily expressed, stored or accessed given its spatial complexity, we are forced to either take random samples from the set of all possible groupings and end up with an approximation of a p-value, or we can use a bootstrapping method.
Bootstrapping is a resampling method when random samples are taken with replacement. Given that bootstrapping methods don't need any assumption of mutual exclusivity between observations, we will use it rather than approximating the sample space. For these reasons, the bootstrapping method of sampling is better than parametric tests and permutation tests in this situation and is comparable to the nonparametric Wilcoxon test because of the few required assumptions that need to be met, namely that the observations are independent and identically distributed.

Statistics

As was previously explained, a bootstrapping analysis will be done to test for the significance of the observed data. The statistic we will use is the difference in mean because we're interested in whether or not there is a difference in the prefectures' population counts. This means that we will generate samples of size 26 with replacement and then split these into two groups where a difference in means will be calculated. At this point we can compare our observed difference in mean to the generated distribution.

Results



As we can see from the density plot of 10,000 bootstrap samples along with the red line indicating the observed difference in means from the data, there is a significant difference in means. Another result that was found was that the difference between the 99^{th} percentile and our observed difference in means was positive meaning that our observed difference in means was more extreme than the 99^{th} percentile and thus significant.

Conclusion

The bootstrapping technique used in the end further justifies the claim that there exists a difference in average population counts between the two Japanese prefectures studied. Should greater computing resources be attained, an exact p-value could be found through a permutation analysis on the entire sample space.