Cheat Sheet: Bootstrap and Simulation-Based Statistics

Essential Concepts

- In statistics, bootstrapping is a procedure done by resampling a single data set to create a multitude of simulated samples.
- A bootstrap sample is a sample that is selected from the values in the original sample. The bootstrap sample is selected with replacement and the sample size is the same as the sample size of the original sample.
- A bootstrap distribution is constructed using the values of a sample statistic calculated from a large number of bootstrap samples. For example, the bootstrap distribution for a sample mean is formed by looking at sample means from a large number of different bootstrap samples.
- In bootstrapping, we do not need to worry about the assumptions and conditions of the sampling distribution. Additionally, we are no longer constrained to test only the means and proportions of the data set. Bootstrapping allows us to find accurate estimates of statistics, without access to the whole population.
- Bootstrap resampling is typically used to estimate confidence intervals.
- In simulation-based hypothesis testing, it allows us to conduct the hypothesis test with very few assumptions and have an intuitive result that can be easily interpreted. We can use our simulated data set to calculate the proportion of the simulated distribution above a certain value as the estimated P-value. We can then use the P-value to determine the strength of evidence the data provide against the null hypothesis.
- Randomization procedures use resampling techniques to construct a sampling distribution that can be used to make inferences about the population. Randomization is constructed given that the null hypothesis is true and its distribution will be centered on the null hypothesis value. This test procedure is only called a randomization test when the data arise from a randomized experiment, since the simulation mimics the random assignment in the study.

- When data arise from an observational study, the procedure is called a permutation test.
- The randomization resampling is typically used to test a hypothesis.
- All simulation-based hypothesis tests follow the same general steps:
    - Set up the null and alternative hypotheses based on the research question.
    - Simulate a large number of samples (usually 1,000 or more) under the assumption of the null hypothesis, calculating a sample statistic for each simulated sample.
    - Plot the simulated sample statistics with a histogram and compare the original observed statistic to the plot.
    - The proportion of simulated statistics as or more extreme than observed is the estimated P-value.

Glossary

bootstrapping

a procedure done by resampling a single data set to create a multitude of simulated samples

bootstrap distribution

the values of a sample statistic that have been calculated from a large number of bootstrap samples

bootstrap confidence interval

an interval estimate constructed using percentiles from a bootstrap distribution of the sample statistic.

bootstrap sample

a distribution of sample statistics calculated from many bootstrap samples, used to approximate the sampling distribution of a statistic.

permutation test

a hypothesis test that uses resampling without replacement to evaluate the likelihood of an observed statistic when data arise from an observational study.

randomization test

comparing the observed difference in proportions to this distribution to obtain an approximate P-value

simulation-based hypothesis test

a hypothesis test that relies on resampling techniques (such as bootstrapping or randomization) to approximate a sampling distribution and compute a P-value.