Computational Statistics - Seminar 4

Andrew Struthers

March 2023

Honor code: I pledge that I have neither given nor received help from anyone other than the instructor or the TAs for all work components included here. – Andrew

Introduction

The purpose of this seminar is to discuss how different resampling methods can be used as tools for data correction. Resampling methods are important tools for data correction and statistical analysis. We can use resampling methods to estimate parameters from a given sample, and to assess the reliability of statistical estimates. We will be focusing on two methods, the Bootstrap method and the Jackknife method. We will also be comparing and contrasting these two methods in terms of context sampling, including important statistical concepts such as medians, variances, and percentiles.

Bootstrap Method

The bootstrap method is a resampling technique that involves generating a large number of new samples from the original data set. Each new sample is created by randomly selecting data points from the original data set with replacement, which will mean that the new sample set has the same size as the original data. The bootstrap method can be used to estimate the variance, standard deviation, and confidence intervals of a sample. We can mathematically describe bootstraping like so: Let $x = \{x_1, x_2, ..., x_n\}$ be a sample of n observations. We want to estimate some property of the population distribution, such as mean or variance. We can generate B bootstrap samples by randomly selecting n observations from x and replacing the original with our new sample. Once we do this, we can then calculate the statistics of interest, such as mean or variance, on this new "resampled" data set. We do this many times, calculating the new statistics for each new sample we generate, ad we denote the b^{th} bootstrap sample statistic as $\bar{Y}_b$. We can then calculate the statistics generated during the bootstrap resampling by taking the average of the B estimates obtained from the B samples:

$$\bar{Y}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\bar{Y}_b \tag{1}$$

The bootstrap estimate will give us an approximation of the statistic we cared about, like the mean, inside this population, as well as giving us the measure of its uncertainty. We can also use the variance of the bootstrap estimate to estimate the standard error, which will allow us to build confidence intervals for the statistic we care about. The bootstrap method lets us also test our hypothesis about the population statistic by calculating a p-value from the bootstrap samples.

Jackknife Method

The jackknife method is another resampling technique where we create new samples from the original data set. Unlike bootstrapping, however, instead of creating new samples by randomly selecting data points with replacement, jackknife involves creating new samples by systematically removing one data point at a time. By systematically leaving one observation out of the sample, recalculating the estimator with the reduced sample, and repeating this process for each observation in the sample, we can estimate the bias and variance of an estimator, and make inferences about the population based on a sample. Again, assume that we have $x = \{x_1, x_2, ..., x_n\}$ which is a sample of n observations. We can let Θ be an estimator of a population statistic of interest. We can generate n jackknife samples by removing one observation from x, then calculating the estimator Θ_i from the reduced sample. We can then calculate the jackknife bias (equation 2) and variance (equation 3) from the following equations:

$$\text{bias} = (n-1)\cdot\frac{1}{n}\sum_{i=1}^{n}(\Theta_i - \Theta)^2 \tag{2}$$

$$\text{var} = (n-1)^2\cdot\frac{1}{n}\sum_{i=1}^{n}(\Theta_i - \bar{y})^2 \tag{3}$$

where $\bar{y}$ is the average of the jackknife estimates. We can calculate this $\bar{y}$ with the following:

$$\bar{y} = \frac{1}{n}\sum_{i-1}^{n}\Theta_i \tag{4}$$

We can use the jackknife variance to get an estimate of the variance of the estimator Θ as well. We can also calculate the jackknife estimate of the estimator Θ with the equation:

$$\Theta_{\text{jack}} = \Theta - \text{bias} \tag{5}$$

With the jackknife estimate, we can get an approximation of the population statistic we care about, adjusted for the bias of the estimator, as well as getting an understanding of statistics variance. We can also construct our confidence intervals and test hypotheses, just like the bootstrapping method, by calculating a p-value from the jackknife estimate and testing that against null hypothesis values.

Comparing Both Methods

Both the bootstrap and jackknife methods are resampling techniques that can be used to estimate statistics from a given sample. However, as we have seen, there are some distinct differences between the two. In terms of variance, the bootstrap method is generally more robust than the jackknife method due to bootstrapping generating new samples by randomly selecting data points using replacement. This can lead to a greater diversity of new samples, whereas the jackknife method systematically removes one data point at a time, which can lead to a lack of diversity in the new samples. With bootstrapping, we can look at different possible permutations of the data and see how that resampling impacts the statistics we care about, whereas with jackknife method we are removing data from the overall data set we have. As an example, lets suppose we want to estimate the median of a sample using both methods. We could generate 5000 new samples using both the bootstrap and jackknife methods, then calculate the median for each new sample. In this example, we would see the bootstrap method produces a larger range of medians, which suggests that bootstrapping is more robust in estimating the median of the original sample. Similarly, if we want to estimate the variance of a sample using both methods, we may find that the bootstrap method produces a more accurate estimate than the jackknife method. This is because the bootstrap method generates a greater diversity of new samples, which will lead to a more accurate estimate of the variance. One other difference between the two methods is that, when estimating the standard error of a statistic, bootstrapping will usually give slightly different results each time we run the resampling, but jackknife will give us the same results every time. This is due to bootstrapping using randomness to create a new sample. Jackknifing does not have any random element to it, so the results will be the same each time we run the resampling.

References

[1] Lorna Yen. An introduction to the bootstrap method. *Medium*, Jan 2019.

[2] Roelof Helmers. Bootstrap method. *Bootstrap method - Encyclopedia of Mathematics.*

[3] Jim Frost. Introduction to bootstrapping in statistics with an example. *Statistics By Jim*, Feb 2023.

[4] Stephanie Glen. Resampling methods: Bootstrap vs jackknife. *Data Science Central*, Jun 2019.

[5] C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261 – 1295, 1986.

[6] Robert Nisbet, Gary Miner, and Ken Yale. Chapter 11 - model evaluation and enhancement. In Robert Nisbet, Gary Miner, and Ken Yale, editors, *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*, pages 215–233. Academic Press, Boston, second edition edition, 2018.