CS 567 Seminar 4

Name: Kollin Trujillo

Date: March 4, 2023
Student ID: 31047365

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.

Introduction: For seminar four we are tasked with looking into resampling methods. Resampling methods are methods commonly employed in statistics in order to be better attain values of interest, which I will describe later on in this seminar called estimators.

Bootstrap falls under a class of methods called robust methods. These class of methods are known as robust because they are better constructed to deal with outliers in data analysis [5]. Outliers are when certain observations are different from the majority of the other ones [5]. There are multiple ways to quantify this deviation mathematically, some equations are highlighted in [5]:

$$z_i = (x_i - \bar{x})/s \tag{1}$$

with s = standard deviation. A typical metric when $|z_i| \geq 2.5$ or so. There are modifications to this that Rousseeuw and Hubert highlight that include the use of the median of all absolute deviations from the median (MAD) which is given by the equation:

$$MAD = 1.483 * \underset{i=1,\ldots,n}{\text{median}}\left|x_i - \underset{j=1,\ldots,n}{\text{median}}(x_j)\right| \tag{2}$$

Which when modifying equation 1 with equation 2 results in:

$$\frac{\left(x_i - \underset{j=1,\ldots,n}{\text{median}}(x_j)\right)}{MAD} \tag{3}$$

which Rousseeuw and Hubert mention give uses the same metric of 2.5 and is a more robust description [5]. These more robust methods are more capable of describing data that is to be an issue with building a model from the mean which the median tends to be more accommodating of outliers which is what stems the creation of this robust method.

Now that I have talked a little bit about what robust statistical methods are, I can elaborate some more on what Bootstrap is. Bootstrap is a statistical method that we can use without having to know the shape of the sampling distribution but allows us to infer a normal one. We are not, in general, guaranteed a normal sampling distribution unless we have a big sample and bootstrap is useful in this case because it allows us to circumvent this problem by utilizing properties from smaller subsamples of the population. These bootstrapped samples are where we have $1/n$ probability that each point can be chosen. With some empirical distribution function (taken from [3]) $\hat{F}$ defined as:

$$\hat{F} = (x_1, \cdots, x_n) \tag{4}$$

we can then choose n random samples, which we refer to as our bootstrapped samples as (taken from [3]):

$$\hat{F} = (x_1^*, \cdots, x_n^*) \tag{5}$$

where x_i can be any point from the population zero-to-many times with replacement to make up the bootstrap sample. With this definition, we can bootstrap our samples and in our case we can calculate the mean with each of these bootstrapped samples and with a lot of these samples we can start to estimate the sampling distribution [4]. From this we can then estimate the standard error of the samples of the population and this helps us to also construct metrics like confidence interval and significance tests [4]. This is shown for our estimator $\hat{\theta} = s(\mathbf{x})$ or our regular and $\hat{\theta}^* = s(\mathbf{x}^*)$ for our bootstrapped estimator where we apply some function $(\cdot)$ to our dataset [3].

The Quenouille jackknife is another nonparametric method [1]. The jackknife method can be seen as a linear expansion approximation ("delta method") to the bootstrap method [1]. The jackknife method differs from the bootstrap in a few key places. This is notated as such:

$$\hat{\theta}_{(i)} = s(x_{(i)}) \tag{6}$$

indicating the i^{th} sample is removed [3]. One of the most different aspects is that the jackknife method is a leave-one-out method as opposed to bootstrap's full utilization. Jackknife also does not utilize replacement. The jackknife method typically involves a factor of $\frac{n}{n-1}$ as opposed to $\frac{1}{n}$ in bootstrap. This isn't a definitive thing but just an observation. It can be seen in the variance calculation of x as well (taken from [2]):

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{7}$$

With our bootstrap and jackknife, we have a few different types of methods that we can use with our variance sampling. For bootstrap, when we talk about variance sampling of our medians. This starts by being given data of the original population. From there we can sample n with replacement. Then, from there, we can calculate the variance of the medians that we calculated. We can do similar with the jackknife method. However, as a whole jackknife is generally good as an approximation to the bootstrap but will fail for non-smooth data, of which median is one [3].

For the percentiles aspect. It follows a similar process as any but involves the calculation of a confidence interval as well, however, it implicates an additional filtering step. The general procedure is that one needs to needs needs to define a range of

$$\left(\theta^*_{(\frac{\alpha}{2})}, \theta^*_{(1-\frac{\alpha}{2})}\right)$$

to get the percentile ranges. Typically we will choose 95% with a ±2.5 percentile, so an α of 0.05. We can then look to take the mean of the bootstrapped samples and use the criterion above to calculate the percentiles. For jackknife, we know that percentile data, such as median, is not typically smooth and is not employed as it is not a good estimation method.

For the variance aspects, it tends to vary highly. For things like linear statistics like the mean, the estimators tend to be higher for jackknife but not nearly as high as things like nonlinear statistics like a correlation coefficient [3]. There are ways to eliminate the necessity of smoothness of data for reliable jackknife resampling data. One example of this is the delete-d method which is an extension of deleting more than just one in the original jackknife method [6].

Overall, one of the larger concerns with the jackknife method was initially during it founding it provided a less computationally expensive method to evaluate estimators than bootstrap. When computational power started to get cheaper and we started to be able to worry less and less about the computational cost of bootstrap. Overall, there are corrective modifications to both jackknife and bootstrap that seek to correct deficient portions of the theory. Bootstrap is widely used when one isn't easily able to attain values of estimators of interest of one's population.

1. References

[1] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[2] Bradley Efron. *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

[3] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[4] A Field, J Miles, and Z Field. Discovering statistics using r| sage publications ltd, 2012.

[5] Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, 2011.

[6] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap.* Springer Science & Business Media, 2012.