An Overview on the Bootstrap

Yufeng Gu, Huajin Yu

Abstract

In this paper, we try to illustrate the basic ideas and methods of the bootstrap by means of which people have access to estimation of the sampling distribution of almost any statistic using random sampling methods. The first two sections show its origin and theoretical background, and then provide a general process to apply the bootstrap. A simple case is shown for you to understand the ideas. In the end, we briefly discuss the advantages and disadvantages of the bootstrap.

1 Introduction

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. It allows assigning measures of accuracy (defined in terms of bias, variance,confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of re-sampling methods.

The idea of bootstrapping was initially introduced by Bradley Efron (1979)[1]. In his book published in 1993[2], Efron gives a complete framework of statistical inference using bootstrap, including estimate of standard errors, bias, establishment of regression models and confidence intervals, hypothesis testing, etc.

In general, the bootstrap can be classified into two groups:

i. Parametric bootstrap. While applying a parametric method, people usually make some assumptions on the probability density (mass) function, $f(x|\theta)$ or $p(x|\theta)$, of population. The estimated function $f(x|\hat{\theta})$ is called a plug-in distribution.

ii. Non-parametric bootstrap. In contrast to parametric ways, non-parametric bootstrap estimates the population properties without assuming any function forms. It just re-sampling from a given sample and get the inferences about population.

In the next section, we will provide some statistical process and a simple practice of the bootstrap.

[1] The bootstrap was published in Bradley Efron's article "Bootstrap methods: Another look at the jackknife", *The Annals of Statistics*. 7(1): 1-26.

[2] *An Introduction to the Bootstrap*, Springer-Science+Business Media, B.V.

2 Theoretical Background

The bootstrap is based on *Edgeworth Expansion*, which is denoted as the expansion of the distribution function around normal distribution. It's similar to the *Taylor Series*.

For instance, if we have n *i.i.d* random variables $X_1, X_2, \cdots, X_n$ with density f, mean μ and variance σ^2. An Edgeworth Expansion for the cumulative distribution function of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ can be written as

$$P\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq \omega\right) = \Phi(\omega) + \phi(\omega)\left[\frac{-1}{6\sqrt{n}}\kappa\left(\omega^2-1\right) + R_n\right] \tag{1}$$

Where nR_n has a boundary, Φ and ϕ denote the c.d.f. and p.d.f. for standard normal distribution $N(\mu, \sigma^2)$, and $\kappa = E\left(X_1 - \mu\right)^3$ is the skewness.

Usually, $\Phi(\omega)$ is a normal converge, and sometimes the bootstrap is proved to converge to the second item, which is called *second-order accurate*[3].

3 General Process to Obtain Bootstrap Estimator

Suppose we select (with putting back) N random variables from a given sample of size N and we get n different re-samples. Since each sample has the same probability to be obtained, they can be regarded as random samples. If $\bar{x}_i^\star$ denotes the mean of the i^{th} re-sample, the variance of $\bar{X}$, mean of the initial sample, can be written as

$$Var^\star(\bar{X}) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\bar{x}_i^\star - \bar{\bar{x}}^\star)^2 \tag{2}$$

Where $\bar{\bar{x}}^\star = \frac{1}{n^n-1}\sum_{i=1}^{n^n}\bar{x}_i^\star$ is the mean of all re-samples (variables with * denote the value of re-sampling).

Furthermore, it's worth noting that equation (1) can be apply to any estimator. i.e. for any estimator $\hat{\theta}(x) = \hat{\theta}$, the estimated variance can be written as

$$Var^\star(\hat{\theta}) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\hat{\theta}_i^\star - \bar{\hat{\theta}}^\star)^2 \tag{3}$$

Where $\hat{\theta}_i^\star$ is the estimator calculated from the i^{th} re-sample, and $\bar{\hat{\theta}}^\star = \frac{1}{n^n-1}\sum_{i=1}^{n^n}\hat{\theta}_i^\star$ is the mean of estimators for re-sampling.

More specifically, the computation of bootstrap estimator can be divided into two different part[4]:

i. Establish B bootstrap samples and calculate

$$Var_B^\star(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^\star - \bar{\hat{\theta}}^\star)^2 \tag{4}$$

[3] For further discussion on Edgeworth Expansion, you can refer to Hall (1992) and Shao and Tu (1995).

[4] This process was introduced in Casella and Berger's book, *Statistical Inference*, 2nd edition, Cengage Learning Press.

ii. Identify the convergence of variance, i.e.

$$Var^{\star}(\hat{\theta}) = \lim_{B\to\infty} Var^{\star}_{B}(\hat{\theta}) \tag{5}$$

Sometimes we can identify the convergence using *Law of Large Number*.

iii. Identify the consistency of equation (2), i.e.

$$Var(\hat{\theta}) = \lim_{n\to\infty} Var^{\star}(\hat{\theta}) \tag{6}$$

A special case is *i.i.d.* random variables, from which we can always get consistent samples.

4 A Simple Practice on Parametric Bootstrap

Here we provide a simple practice of parametric bootstrap. Now we have a sample

$$-1.81, 0.63, 2.22, 2.41, 2.95, 4.16, 4.24, 4.53, 5.09$$

where $\bar{x} = 2.71$ and $s^2 = 4.82$. If we assume that the population follows a normal distribution $N(\mu, \sigma^2)$, the bootstrap should drawn a new sample

$$X^{\star}_1, X^{\star}_2, \cdots, X^{\star}_n \sim N(2.71, 4.82).$$

Using $B = 1000$ sets of observed sample, we calculate $Var^{\star}_{B}(S^2) = 4.33$. Since we have assumed a normal distribution, the theoretical variance of S^2 is $Var(S^2) = 2(\sigma^2)^2)/8$. Then we can apply MLE method to get $2(4.82)^2/8 = 5.81$, i.e. $\hat{\sigma} = 4.82$, while the real population variance $Var(S^2) = 4.00$, implying that the re-sampling method gives a reasonable estimation.

5 Summary

The bootstrap provides a straightforward way for researchers to derive estimates of critical values for estimators of complicated distributions. People can check the stability of the results and reject some hypotheses without knowing the accurate confidence interval. However, since it only uses finite sample, we should suspect the external validity, i.e. the result may not be reliable when it expands to the population.

Hence, it's of great significance that we should strictly check if the data satisfies all the basic requirements of the bootstrap before we use it.

References

1. Casella, G. and R. L. Berger (2002). *Statistical Inference* (2nd), Thomson Learning.

2. DiCiccio TJ, Efron B (1996). "Bootstrap confidence intervals (with Discussion)". *Statistical Science*. 11: 189–228.

3. Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". *The Annals of Statistics*. 7 (1): 1–26.

4. Efron, B. (2003). "Second thoughts on the bootstrap". *Statistical Science*. 18(2): 135-140.

5. Efron, B. Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

6. Varian, H.(2005). "Bootstrap Tutorial". *Mathematica Journal*. 9: 768–775.

7. Weisstein, Eric W. "Bootstrap Methods." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BootstrapMethods.html