Introduction to the Non-Parametric Bootstrap[1]

Daniel Berrar

Machine Learning Research Group
School of Mathematics and Statistics
The Open University, Milton Keynes, United Kingdom
Email: daniel.berrar@open.ac.uk
and
Department of Information and Communications Engineering, School of Engineering,
Tokyo Institute of Technology, Tokyo, Japan

Abstract

The bootstrap is a computationally intensive data resampling methodology for assessing the accuracy of statistical estimates and for making inferences about unknown population parameters. This article provides an introduction to the ordinary non-parametric bootstrap, with a focus on two applications that are of particular relevance to bioinformatics: bootstrap confidence intervals and bootstrap error estimates of predictive performance.

Keywords: .632 bootstrap; .632+ bootstrap; bootstrapping; bootstrap-t interval; bootstrap; confidence interval; percentile interval; Monte Carlo method; resampling; sampling distribution; resubstitution error; leave-one-out bootstrap estimate; inference

Key points

- This article explains the fundamental principles of bootstrapping, from sampling distributions to the .632+ bootstrap.
- Bootstrap confidence intervals and bootstrap error estimates of the predictive performance are covered as well.

1. Introduction

The *bootstrap* is a computationally intensive data resampling methodology for assessing the accuracy of statistical estimates and for making statistical inferences about unknown population parameters [2, 3, 5, 6, 7, 8]. A central question of inferential statistics is the following: how accurate is the sample statistic $\hat{\theta}$ as an estimator for an unknown population parameter θ? The key to this question is the *sampling distribution* of $\hat{\theta}$. Suppose that we are interested in the population mean, μ. It is well known that the sample mean,

[1]This article is the revised version of [1] for the 2nd edition of the *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier.

$\bar{x}$, is the best point estimate for μ, and its sampling distribution can be analytically derived. However, for more intricate statistics (for example, performance measures of predictive models, such as the area under the receiver operating characteristic curve, AUC), it may be very difficult or even impossible to find the theoretical sampling distribution. Here, the bootstrap provides a solution.

In theory, to obtain the sampling distribution of a statistic of interest, we would need to draw all or infinitely many random samples from the population and compute the statistic for each sample. In practice, we (usually) have only one sample from the population. At the heart of the bootstrap is the idea that the available sample is a good estimate of the population of interest. So instead of drawing random samples from the population, we draw random samples from our estimate. These "samples from a sample" are called *bootstrap sets* or *bootstrap samples*. From each bootstrap sample, we then compute our statistic of interest. The spread and the shape of the distribution of these *bootstrap statistics* tell us something about the sampling distribution of the sample statistic, which provides useful information for making inferences about the population parameter. Now the etymological roots of the methodology also become clear: the name is derived from the phrase "to pull oneself up by one's own bootstrap," which implies that a thing is being built by using the thing itself—by drawing samples from the sample, we are building a bootstrap distribution.

Without specific assumptions or a particular model for the population under investigation, the bootstrap is called *non-parametric*; otherwise, it is called *parametric* [9]. Like random permutation methods and the jackknife [4], the bootstrap belongs to the family of Monte Carlo resampling methods [10]. Essentially, these methods differ with respect to how the sampling is done. Both the jackknife and random permutation methods perform sampling *without* replacement, whereas sampling is done *with* replacement in the bootstrap.

This article provides an introduction to the ordinary non-parametric bootstrap, which is the most fundamental type. Here, we consider only the case that the population parameter and sample statistic are a scalars. The focus is on the key ideas and two applications that are particularly relevant for bioinformatics: how to derive basic bootstrap confidence intervals, and how to assess the prediction error of a statistical model (for example, a classifier). For more theoretical details, see [5, 7, 11]. For an excellent discussion of the role of the bootstrap in statistics education, see [12].

1.1. Basic concepts and notation

We begin with some basic terms, which largely follow the standard statistical notation [8]. Greek lower case letters usually denote population parameters, for example, μ denotes the population mean. Roman upper case letters, such as X, usually denote sample statistics. For instance, $X = x_i$ means that X assumes a particular value x_i. The exception is N, which refers to the population size. $\mathrm{E}(X)$ denotes the expected value or mean of X.

The probability distribution of X is called *sampling distribution of X*. The probability that a real-valued random variable X is smaller than, or equal to, a particular value x_0 is denoted by $P(X \leq x_0) = \mathcal{F}_X(x_0)$,

which is called the *cumulative distribution function* (CDF) of X. The unknown parameter of interest is commonly denoted by θ, which could represent the mean or any other parameter. The symbol $\hat{}$ indicates an estimate. For example, the sample statistic $\hat{\theta}$ is an estimate for θ; therefore, $\hat{\theta}$ is also called the *estimator* and θ is called the *estimand*. The standard deviation of the sampling distribution of the sample statistic $\hat{\theta}$ is called the *standard error* of $\hat{\theta}$. Finally, the symbol * indicates a bootstrap statistic.

Note that in frequentist statistics, a parameter has a fixed but (usually) unknown value and is therefore a constant. By contrast, a sample statistic is a random variable because it depends on the random makeup of the sample from which it was calculated. In the frequentist paradigm, sample statistics are random variables, and consequently, they have a sampling distribution; parameters, on the other hand, do not, since they are not random variables.

1.2. Sampling distribution of a sample statistic

Let us assume that we have a data set D, which is a random sample from the population of interest. The data set has n elements, $x_1, x_2, ..., x_n$. We will assume that the elements in our random sample are the real-valued outcomes of independent and identically distributed (iid) random variables $X_1, X_2, ..., X_n$. The *probability density function* (PDF) of these random variables is denoted by f, and their cumulative distribution function is denoted by $\mathcal{F}$. Both f and $\mathcal{F}$ are characteristics of the population and therefore unknown, just like the population parameter that we are interested in. We use the sample to make inferences about the unknown population parameter θ by calculating the statistic $\hat{\theta}$ from the sample data.

The population parameter can be thought of as a function of the population data, i.e., $\theta = t(x)$, where x denotes the population data and $t(\cdot)$ is a statistical function. This function tells us how to calculate θ from the data. For instance, let us assume that θ denotes the population mean, μ, and let us assume that the population elements are countably finite. Then $\mu = t(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$, where N is the number of elements in the population. The same statistical function may be used to calculate the sample statistic, $\hat{\theta} = t(x)$, where x now denotes the sample data. For example, the sample mean is calculated as $\bar{x} = t(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$, where n is the number of elements in the sample. In non-parametric analysis, the *empirical distribution function* (EDF), denoted by $\hat{\mathcal{F}}$, is an estimate of the cumulative distribution function, $\mathcal{F}$. Many simple statistics can be regarded as properties of the empirical distribution function [8]. For example, let us consider the discrete case, where each x_i has the probability $\frac{1}{n}$ of being sampled. The sample mean, $\bar{x}$, is the same as the mean of the empirical distribution function, $\hat{\mathcal{F}}$,

$$\hat{\mathcal{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - x_i), \quad \text{with} \tag{1}$$

$$H(u) = \begin{cases} 1, & \text{if } u \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $H(\cdot)$ is the Heaviside function. Both the population parameter and the sample statistic can be expressed as a function of their underlying distribution functions: $\theta = t(\mathcal{F})$ and $\hat{\theta} = t(\hat{\mathcal{F}})$. The reason is that the parameter can be considered a characteristic of the population, which is described by its cumulative distribution function, $\mathcal{F}$. Similarly, the value of the sample statistic is a function of the empirical distribution of the sample elements, $\hat{\mathcal{F}}$. Not all estimators, however, are exactly of the form $\hat{\theta} = t(\hat{\mathcal{F}})$ [8]; for example, the unbiased sample variance is $\frac{n}{n-1}t(\hat{\mathcal{F}})$. More precisely, therefore, it should be stated that $\hat{\theta} = t_n(\hat{\mathcal{F}})$ and $t_n \to t$ as $n \to \infty$. Here, we will ignore this detail, as it is irrelevant for the explanation of the non-parametric bootstrap.

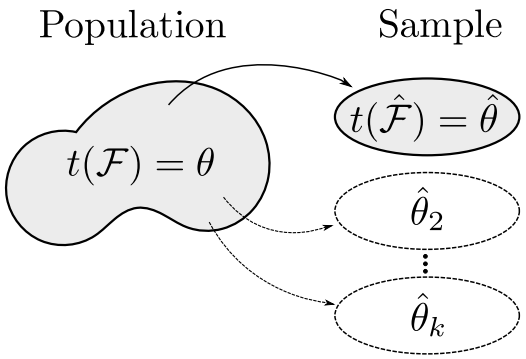

Figure 1: Random sampling from a population with unknown parameter θ. The sample statistic $\hat{\theta}$ is the estimator for the parameter (i.e., the estimand). Both the statistic and the parameter depend on their underlying distribution functions $\hat{\mathcal{F}}$ and $\mathcal{F}$, respectively.

The idea of random sampling from a population is illustrated in Figure 1. As we have seen, the sample statistic, $\hat{\theta}$, can be thought of as a function of the empirical distribution, $\hat{\mathcal{F}}$, of the data in that particular sample. Clearly, if we drew another random sample, then most certainly we would obtain a slightly different value of $\hat{\theta}$. This idea is illustrated by $\hat{\theta}_2$ and $\hat{\theta}_k$ in Figure 1. Assume that $t(\cdot)$ is the function for the arithmetic mean. It is plausible that sometimes, $\hat{\theta} = \bar{X}$ is a bit larger than $\theta = \mu$, sometimes a bit smaller, but on average, the sample mean and the population mean are the same. In fact, it can be proved that the expected value of the sample mean is the population mean, i.e., $\mathrm{E}(\bar{X}) = \mu$.

Two characteristics of an estimator are its bias, β, and variance, $\sigma^2_{\hat{\theta}}$. The bias is the systematic error, i.e., the difference between the estimator's expected value, $\mathrm{E}(\hat{\theta})$, and the true value of the estimand, θ, i.e., $\beta = \mathrm{E}(\hat{\theta}) - \theta$. An estimator is said to be *unbiased* if its bias is zero. Hence, the sample mean $\bar{X}$ is an unbiased estimator of μ.

Consider a sample statistic $\hat{\theta}$ with mean $\mu_{\hat{\theta}} = \theta + \beta$ and standard deviation $\sigma_{\hat{\theta}}$, and assume that $\hat{\theta}$ is normally distributed, $\hat{\theta} \sim \mathcal{N}(\mu_{\hat{\theta}}, \sigma^2_{\hat{\theta}})$. The probability that $\hat{\theta}$ takes on a value smaller than, or equal to, a particular value $\hat{\theta}_0$ is

$$P(\hat{\theta} \leq \hat{\theta}_0) = \Phi\left(\frac{\hat{\theta}_0 - \mu_{\hat{\theta}}}{\sigma_{\hat{\theta}}}\right) = \Phi\left(\frac{\hat{\theta}_0 - (\theta + \beta)}{\sigma_{\hat{\theta}}}\right) \tag{2}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Furthermore,

$$P\left((\theta + \beta) - z_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\theta}} \leq \hat{\theta} \leq (\theta + \beta) + z_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\theta}}\right) = 1 - \alpha \tag{3}$$

where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$ is the quantile of the standard normal distribution for probability $1 - \frac{\alpha}{2}$. For example, if $\alpha = 0.05$, then $z_{0.975} = 1.96$. Rearranging Eq. 3 leads to the following identity[2] for the population parameter θ,

$$P\left((\hat{\theta} - \beta) - z_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\theta}} \leq \theta \leq (\hat{\theta} - \beta) + z_{1-\frac{\alpha}{2}} \times \sigma_{\hat{\theta}}\right) = 1 - \alpha \tag{4}$$

For example, let $\hat{\theta}$ be the sample mean, $\bar{X}$. It is well known that the sampling distribution of the sample mean is a normal distribution with mean $\mu_{\bar{X}} = \mathrm{E}(\bar{X}) = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, if the population has a normal distribution (with standard deviation σ) or if the sample size is sufficiently large (usually, $n \geq 30$) (cf. Figure 2). The sample mean is an unbiased estimator of the population mean, as $\beta = \mathrm{E}(\bar{X}) - \mu = \mu - \mu = 0$. A $(1 - \alpha) \times 100\%$ confidence interval for the population mean μ is now given by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \tag{5}$$

In practice, the population standard deviation is usually not known, and the confidence interval is calculated based on Student's t-distribution. The Student-t confidence interval for the population mean is calculated as

$$\bar{x} \pm t_{\nu,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \tag{6}$$

where $t_{\nu,1-\frac{\alpha}{2}}$ is the quantile of the t-distribution with ν degrees of freedom, with $\nu = n - 1$, and $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation.

So if we know the bias and variance of the sample statistic, then we can make inferences about the population parameter. For the sample mean, this is no problem, but what about more intricate statistics? Let us consider bias and variance as a function of the underlying distribution,

[2]Eq. 4 should be read carefully. Reading it as "the probability that the parameter θ is between ... and ... is $1 - \alpha$" might invite the misinterpretation that the parameter has a probability distribution. This is not so. The probabilistic interpretation must refer to the random variable, $\hat{\theta}$.

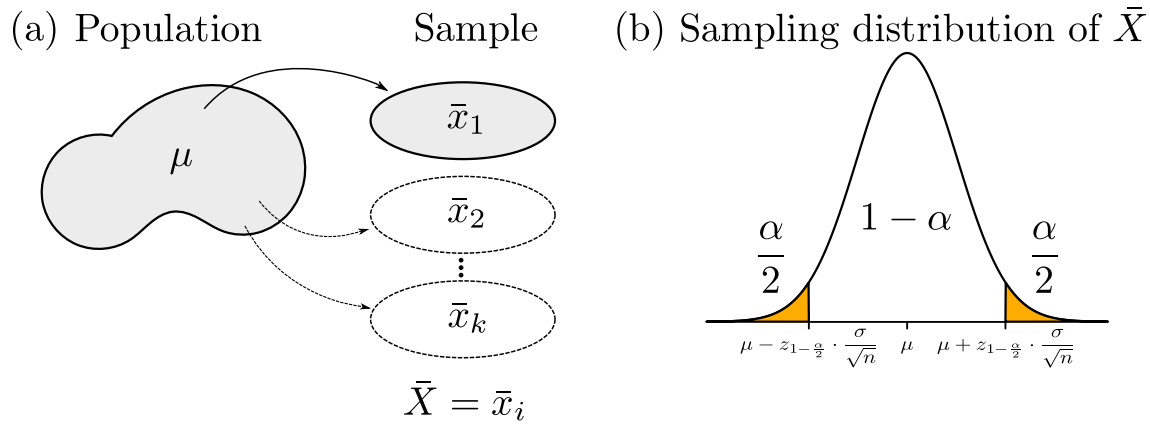

Figure 2: (a) Random sampling from a population with unknown mean μ. The estimator for the population mean is the sample mean, $\bar{X}$. The sample mean is a random variable and takes on the values $\bar{x}_i$, which depend on the concrete makeup of the random sample. (b) Sampling distribution of the sample mean, $\bar{X}$. The distribution is normal, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, i.e., the distribution is centered at the population mean, μ. The value $z_{1-\frac{\alpha}{2}}$ is the quantile of the standard normal distribution for probability $1 - \frac{\alpha}{2}$. For example, $z_{0.975} = 1.96$ for $\alpha = 0.05$.

$$\beta = \mathrm{E}(\hat{\theta}|\mathcal{F}) - t(\mathcal{F}) \tag{7}$$

$$\sigma^2_{\hat{\theta}} = \mathrm{Var}(\hat{\theta}|\mathcal{F}) \tag{8}$$

Using $\hat{\mathcal{F}}$ as an estimate for $\mathcal{F}$, we obtain estimates of bias and variance,

$$\hat{\beta} = \mathrm{E}(\hat{\theta}|\hat{\mathcal{F}}) - t(\hat{\mathcal{F}}) \tag{9}$$

$$\hat{\sigma}^2_{\hat{\theta}} = \mathrm{Var}(\hat{\theta}|\hat{\mathcal{F}}) \tag{10}$$

which are called *bootstrap estimates* [8]. This illustrates the *plug-in principle* [12] of the bootstrap: when something is unknown, an estimate is plugged in instead.

2. Ordinary non-parametric bootstrap

The *ordinary non-parametric bootstrap* procedure can be described as follows:

1. The available data set D is assumed to be a representative sample from the population of interest. This data set contains a total of n elements, $x_1, x_2, ..., x_n$.
2. From D, calculate the statistic $\hat{\theta}$, which is the estimate for the population parameter, θ.
3. Generate a bootstrap set, B, by randomly sampling n instances with replacement from D. The sampling is uniform, which means that each of the n elements in D has the same probability, $\frac{1}{n}$, of

being selected.[3]

4. Repeat step (3) b times to generate b bootstrap sets, $B_1, B_2, ..., B_b$ (cf. Figure 3).
5. Calculate the statistic $\hat{\theta}_i^*$ from the i^{th} bootstrap set B_i.
6. Repeat step (5) for all b bootstrap sets.

From a data set with n elements, we could theoretically generate $\binom{2n-1}{n}$ distinct bootstrap sets. If n is not too large, we could perform an *exhaustive bootstrap* by constructing all possible sets instead of random sampling. This approach is also referred to as the *theoretical bootstrap* [12]. In practice, however, n is usually too large, and we generate several hundreds or thousands of bootstrap samples. This approach is an example of *Monte Carlo sampling*. Hesterberg suggests $b = 1000$ for a rough approximation and $b \geq 10\,000$ for more accuracy [12].

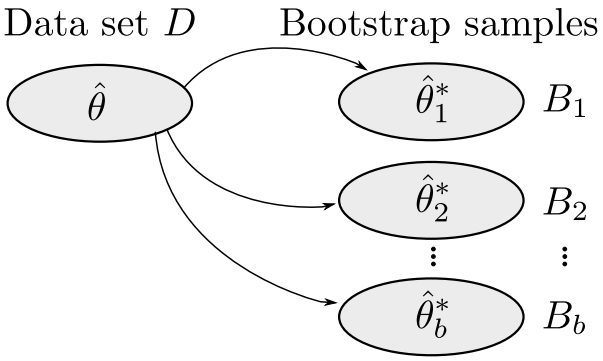

Figure 3: Ordinary non-parametric bootstrap. The available data set (i.e., original sample) is repeatedly sampled (with replacement) to generate b bootstrap sets B_i, $i = 1..b$. Each bootstrap set has the same number of elements, n, as the original data set D. Because of the sampling with replacement, some elements can occur more than once in one bootstrap set, while others do not occur at all.

After the described procedure, we obtain the *empirical* or *bootstrap distribution* of all $\hat{\theta}_i^*$, from which we can infer the *bootstrap cumulative distribution function*, $\hat{\mathcal{F}}^*$. The mean, bias, and variance of $\hat{\theta}^*$ are:

$$\bar{\hat{\theta}}^* = \frac{1}{b}\sum_{i=1}^{b}\hat{\theta}_i^* \tag{11}$$

$$\text{bias}\{\hat{\theta}^*\} = \text{E}(\hat{\theta}_i^*) - \hat{\theta} = \bar{\hat{\theta}}^* - \hat{\theta} \tag{12}$$

$$\hat{\sigma}^2_{\hat{\theta}^*} = \frac{1}{b-1}\sum_{i=1}^{b}\left(\hat{\theta}_i^* - \bar{\hat{\theta}}^*\right)^2 \tag{13}$$

It is important to note that the bootstrap distribution is centered at the observed statistic, $\hat{\theta}$, not at the population parameter, θ. Our best estimate for the population parameter (say, the mean μ) is therefore still

[3]More intricate sampling approaches, such as sampling with noise, are beyond the scope of this article.

our original sample statistic (i.e., $\bar{x}$ for μ), not the average of all bootstrap estimates, $\bar{\hat{\theta}}^*$. As the bootstrap distribution is not centered at the population parameter, the quantiles of the bootstrap distribution are usually different from the quantiles of the theoretical sampling distribution of the sample statistic, $\hat{\theta}$. This means that the bootstrap quantiles are not meaningful estimates for the quantiles of $\hat{\theta}$. However, the bootstrap quantiles are useful to estimate the quantiles and CDF of $\hat{\theta} - \theta$ and to estimate the standard deviation of $\hat{\theta}$. This idea is illustrated in Figure 4 using the sample mean $\bar{X}$ as an example. Note that in Figure 4d, the standard deviation of the bootstrap statistic is only slightly larger than the standard deviation of the theoretical sampling distribution of the sample statistic. The example in Figure 4 is just an illustration of the principle of bootstrapping. For the sample mean, the bootstrap is actually not needed because we already know the theoretical sampling distribution of $\hat{\theta}$. For many other statistics, however, that is not the case.

3. Basic bootstrap confidence intervals

We now describe two basic bootstrap confidence intervals: the *bootstrap percentile interval* and the *bootstrap-t interval* [7]. More advanced intervals are described in [6, 13].

3.1. Bootstrap percentile confidence interval

The bootstrap percentile confidence interval is arguably the simplest and most intuitive bootstrap interval. It is derived as follows.

1. Generate b bootstrap sets by repeatedly sampling with replacement from the available data set (i.e., original sample) D. For each bootstrap sample B_i, calculate the sample statistic $\hat{\theta}_i^*$.
2. Find the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles of the distribution of $\hat{\theta}_i^*$. These percentiles are denoted by $\hat{\theta}^*_{\frac{\alpha}{2}}$ and $\hat{\theta}^*_{1-\frac{\alpha}{2}}$, respectively.[4]
3. A $(1-\alpha)\times 100\%$ bootstrap percentile confidence interval for θ is given by

$$\left[\hat{\theta}^*_{\frac{\alpha}{2}},\ \ \hat{\theta}^*_{1-\frac{\alpha}{2}}\right] \tag{14}$$

3.2. Bootstrap-t confidence interval

The bootstrap-t interval is constructed as follows.

1. Draw b bootstrap samples by repeatedly sampling with replacement from the available data set D. For each bootstrap sample B_i, calculate

$$Z_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{\widehat{\mathrm{SE}}_i^*} \tag{15}$$

[4] In R, the function `quantile(Y, c(0.025, 0.975))` gives the 2.5% and 97.5% percentile for the data in `Y`.

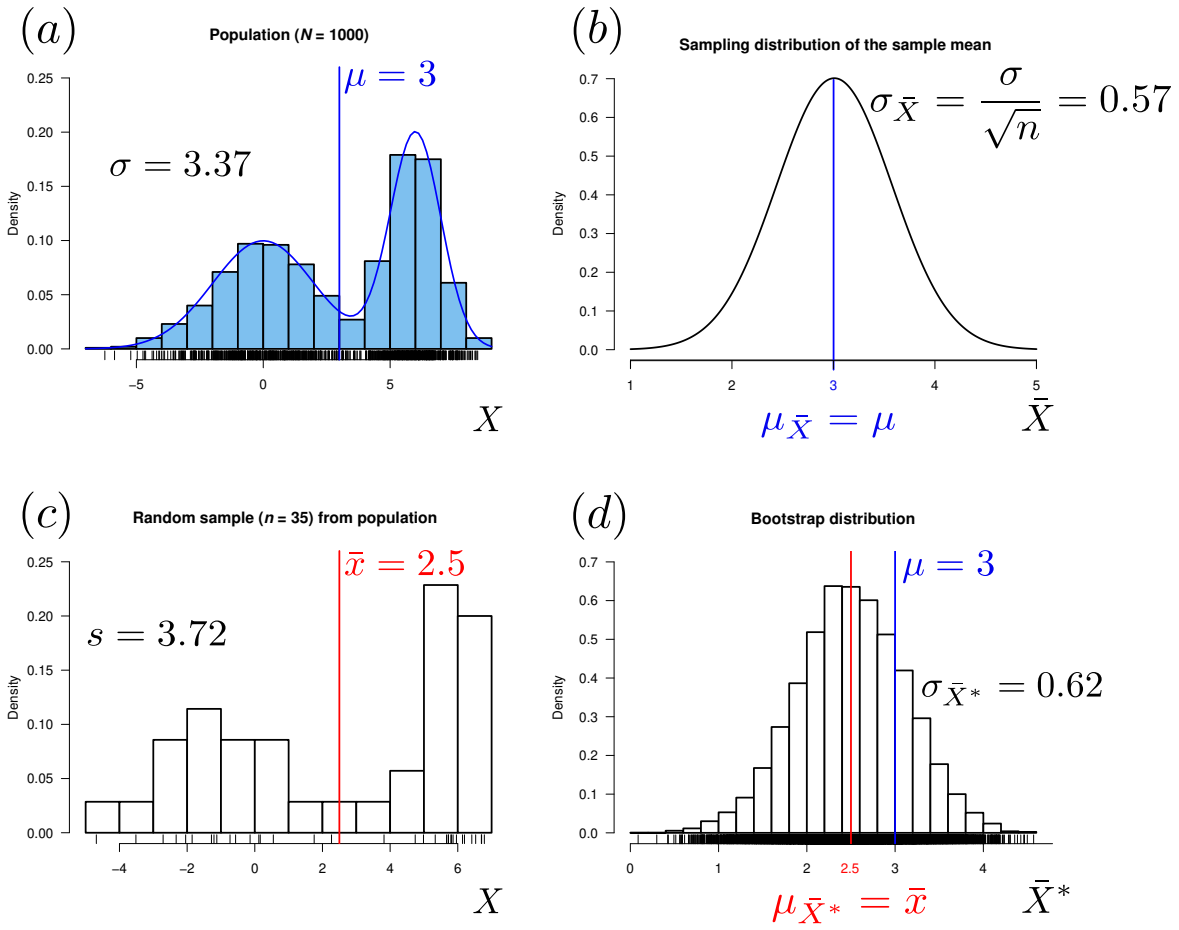

Figure 4: (a) An example of a population of interest, representing $N = 1000$ instances x_i from a normal mixture of two distributions, $\mathcal{N}_1(0, 4)$ and $\mathcal{N}_2(6, 1)$. The population mean and standard deviation are $\mu = 3$ and $\sigma = 3.37$, respectively. (b) The sampling distribution of the sample mean is normal, with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 0.57$. (c) A random sample of $n = 35$ instances from the population. The sample mean is $\bar{x} = 2.5$ and the standard deviation is $s = 3.72$. (d) Bootstrap (empirical) distribution of the bootstrap estimate $\bar{X}^*$ based on $b = 10\,000$ bootstrap sets sampled from the data in (c). The bootstrap distribution has mean $\mu_{\bar{X}^*} = \bar{x} = 2.5$ and standard deviation $\sigma_{\bar{X}^*} = 0.62$.

where $\widehat{\mathrm{SE}}_i^*$ is the estimate of the standard error of $\hat{\theta}^*$ based on the data in B_i. This estimate is calculated as the standard deviation of all values in B_i divided by the square root of the sample size, $\widehat{\mathrm{SE}}_i^* = \frac{\hat{\sigma}_i^*}{\sqrt{n}}$. The random variable Z_i^* is called an *approximate pivot* [7].

2. Find the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ percentiles of the distribution of Z_i^*. Denote these percentiles by $\hat{t}_{\frac{\alpha}{2}}^*$ and $\hat{t}_{1-\frac{\alpha}{2}}^*$, respectively.

3. A $(1-\alpha)\times 100\%$ bootstrap-t interval for the parameter θ is given by

$$\left[\hat{\theta} - \hat{t}_{1-\frac{\alpha}{2}}^* \times \frac{s}{\sqrt{n}},\ \ \hat{\theta} - \hat{t}_{\frac{\alpha}{2}}^* \times \frac{s}{\sqrt{n}}\right] \tag{16}$$

where s denotes again the standard deviation of the sample (i.e., the original data set D).

Eq. 16 is similar to the common Student-t interval, except that the t-value is substituted by the bootstrap estimates $\hat{t}_{\frac{\alpha}{2}}^*$ and $\hat{t}_{1-\frac{\alpha}{2}}^*$. These estimates are not necessarily symmetric about 0, in contrast to the percentiles of the standard normal and t-distribution. Whereas the common z- and t- intervals are always symmetric about θ, bootstrap-t intervals (and bootstrap percentile intervals) are not necessarily symmetric.

It may be surprising that Eq. 16 does not include a term for the bias (cf. Eq. 4). Indeed, we could calculate a bias-corrected estimate as $\hat{\theta} - \mathrm{bias}\{\hat{\theta}^*\} = \hat{\theta} - (\bar{\hat{\theta}}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\hat{\theta}}^*$. However, bias estimates can have a high variability [7, 12], so it is not advisable to perform this correction.

Table 1 shows the two bootstrap 95% confidence intervals for the population mean μ of the example population (Figure 4a). For comparison, the common z- and t-intervals are also shown.

Table 1: 95% confidence intervals for the population mean based on the random sample in Figure 4a.

Method	95% confidence interval	Width
Bootstrap percentile interval	$[1.238, 3.696]$	2.458
Bootstrap-t interval	$[1.184, 3.769]$	2.585
Student-t interval	$[1.224, 3.780]$	2.556
Normal-z interval	$[1.387, 3.617]$	2.230

For the example data in Figure 4, $n = 35$ seems to be sufficient for a reliable bootstrap confidence interval. The width of both bootstrap intervals is comparable to that of the Student-t interval. The z-interval is of course the most accurate interval in this example because it uses the population standard deviation, σ. For real-world data sets, however, it is usually not possible to derive a z-interval, as σ is almost never known.

Both the bootstrap-t interval and the bootstrap percentile interval are only first-order accurate, which means that the actual one-sided coverage probabilities differ from the nominal values by $\mathcal{O}(\frac{1}{\sqrt{n}})$. Hence, both intervals tend to be too narrow for small n [12]. For small sample sizes, the bootstrap intervals offer therefore no improvement over the Student-t interval. On the other hand, when n is large, bootstrap intervals usually perform better, particularly for skewed distributions [12]. The reason is that common confidence intervals

based on the standard normal and t-distribution are symmetric about zero, whereas bootstrap percentile intervals and bootstrap-t intervals may be asymmetric, which can improve coverage [7]. In fact, the Student-t interval was shown to be surprisingly inaccurate when the population distribution is skewed, even if the sample size is much larger than 30 [14]. For large samples from skewed populations, bootstrap-t intervals tend to have a better coverage than the Student-t intervals and are therefore preferable. The bootstrap-t interval is particularly suitable to location statistics, such as the mean or median [7].

When the bootstrap distribution is approximately normal and the bias is small, the bootstrap-t interval and the percentile bootstrap interval will agree closely. If that is not the case, Hesterberg et al. caution against the use of either interval [15].

4. Bootstrap estimates of prediction error

When only one data set is available to develop and test a statistical model (for example, a classifier or regression model), the bootstrap is an effective technique to estimate the true prediction error [9]. This true prediction error represents the unknown population parameter. Let us assume that a data set D contains n instances (or cases) $\mathbf{x}_i$, $i = 1..n$, and each instance is described by a set of p features, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Let us further assume that with each instance, exactly one target label y_i is associated. In the case of classification, y_i is a discrete class label, $y_i \in \{y_1, y_2, ..., y_k\}$, where k indicates the number of distinct classes. In the case of regression, the target is a real value, $y_i \in \mathbb{R}$. A statistical model $f(\cdot)$ estimates the target y_i of the case $\mathbf{x}_i$ as $f(\mathbf{x}_i) = \hat{y}_i$. The estimation error is quantified by a *loss function*, $\mathcal{L}(y_i, \hat{y}_i)$. For example, in the case of classification and the 0-1 loss function, the loss is 1 if $y_i \neq \hat{y}_i$ and 0 otherwise. In the case of a regression problem, the loss function generally involves a form of squared error. The *bootstrap estimate of the prediction error* is calculated as follows.

1. Generate b bootstrap sets B_j, $j = 1..b$, by repeatedly sampling (with replacement) n cases from D.
2. Use the bootstrap set B_j as training set to build the model f_j^*.
3. Calculate $f_j^*(\mathbf{x}_i) = \hat{y}_i$, and then compute the loss function $\mathcal{L}(y_i, f_j^*(\mathbf{x}_i))$.
4. The *bootstrap estimate of the prediction error*, $\hat{\epsilon}^*$, is the average over all cases and all bootstrap sets,

$$\hat{\epsilon}^* = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} \sum_{j=1}^{b} \mathcal{L}(y_i, f_j^*(\mathbf{x}_i)) \tag{17}$$

As the training sets B_j partially overlap with D, the error estimate $\hat{\epsilon}^*$ is biased downward, which means that it is smaller than the true prediction error. A simple approach to alleviate the optimistic bias is to exclude those cases from the evaluation that served already as training cases. Let S_{-i} denote the set of indices of the bootstrap samples that do not contain the case $\mathbf{x}_i$, and let $|S_{-i}|$ denote the number of these indices. The *leave-one-out bootstrap estimate*, $\hat{\epsilon}^*_{loob}$, is defined as follows [16].

$$\hat{\epsilon}^*_{loob} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|S_{-i}|} \sum_{j \in S_{-i}} \mathcal{L}(y_i, \hat{f}_j^*(\mathbf{x}_i)) \tag{18}$$

It is possible that all bootstrap samples contain the case $\mathbf{x}_i$, and consequently that $|S_{-i}| = 0$. To avoid a division by zero, the number of bootstrap samples, b, has to be sufficiently large to ensure that at least one of them does not contain $\mathbf{x}_i$, or those cases that appear in all bootstrap samples should be omitted from Eq. 18.

The leave-one-out bootstrap estimate tends to overestimate the true prediction error; in other words, it has an upward bias. The reason is that the model is trained on only a subset of the available data because each bootstrap sample contains, on average, only about 63% of the data: the probability that a case $\mathbf{x}_i$ is selected for a bootstrap set in one random sampling is $\frac{1}{n}$, and the probability of not being selected is therefore $1 - \frac{1}{n}$. When the sampling is done n times, the probability that the case is not selected at all is $(1 - \frac{1}{n})^n$. Consequently, the probability that it *is* selected is $1 - (1 - \frac{1}{n})^n$. Note that $(1 - \frac{1}{n})^n \to e^{-1}$ for $n \to \infty$. Therefore, for sufficiently large n, the probability that a case $\mathbf{x}_i$ appears in a bootstrap set is approximately $1 - e^{-1} = 0.632$. This means that each bootstrap set has, on average, about 63% of distinct cases only—so about 37% of the available data are not used for training, which results in an overestimation of the prediction error. This overestimation can be corrected by considering the resubstitution error. The *bootstrap resubstitution error* is defined as follows [16].

$$\hat{\epsilon}^*_{resub} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|S_{+i}|} \sum_{j \in S_{+i}} \mathcal{L}(y_i, \hat{f}^*_j(\mathbf{x}_i)) \tag{19}$$

where S_{+i} is the set of bootstrap set indices that contain the case $\mathbf{x}_i$. The resubstitution error is also referred to as the *training error*. It has a downward bias, which means that it underestimates the true prediction error.

Figure 5 illustrates the relation between the leave-one-out error and the resubstitution error. From a data set with $n = 10$ cases, three bootstrap samples are randomly drawn. Here, only the case indices are shown. A model f^* is fitted on each sample. Consider now the contribution of case #1 to the leave-one-out bootstrap estimate: only the prediction of f^*_1 is relevant, since it is fitted on the bootstrap sample that does *not* include case #1. By contrast, to calculate the contribution of case #1 to the resubstitution error, only the predictions of f^*_2 and f^*_3 are relevant, because these models were fitted on samples that do contain case #1.

5. The .632 bootstrap and .632+ bootstrap

The *.632 bootstrap* error, $\hat{\epsilon}^*_{.632}$, is a weighted average between the overly optimistic resubstitution error and the overly pessimistic leave-one-out bootstrap error,

$$\hat{\epsilon}^*_{.632} = 0.368 \times \hat{\epsilon}^*_{resub} + 0.632 \times \hat{\epsilon}^*_{loob} \tag{20}$$

The error estimate $\hat{\epsilon}^*_{.632}$ still has a downward bias [17], which the *.632+ bootstrap* corrects by adjusting the weights for the resubstitution error and leave-one-out bootstrap error,

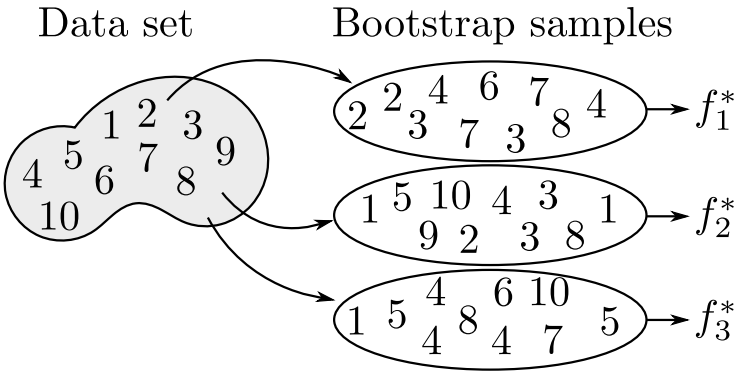

Figure 5: Simplified example of a data set of 10 cases and 3 bootstrap samples. The models f_1^*, f_2^*, and f_3^* are fitted on the respective samples. Cases are represented by their indices.

$$\hat{\epsilon}^*_{.632+} = (1 - w) \times \hat{\epsilon}^*_{resub} + w \times \hat{\epsilon}^*_{loob}, \quad \text{with} \tag{21}$$

$$w = \frac{0.632}{1 - 0.368\hat{R}} \tag{22}$$

$$\hat{R} = \frac{\hat{\epsilon}^*_{loob} - \hat{\epsilon}^*_{resub}}{\hat{\epsilon}_{random} - \hat{\epsilon}^*_{resub}} \tag{23}$$

The quantity $\hat{\epsilon}_{random}$ in Eq. 23 is an estimate of the no-information error rate, i.e., the expected error when the cases (more precisely, the attributes or features describing the cases) are statistically independent of their labels. $\hat{R}$ is a measure of the relative overfitting rate, which ranges from 0 if $\hat{\epsilon}^*_{loob} = \hat{\epsilon}^*_{resub}$ (i.e., there is no overfitting) to 1 if $\hat{\epsilon}^*_{loob} = \hat{\epsilon}_{random}$. The weight w ranges from 0.632 (if $\hat{R} = 0$) to 1 (if $\hat{R} = 1$), and consequently, $\hat{\epsilon}^*_{.632+}$ ranges from $\hat{\epsilon}^*_{.632}$ to $\hat{\epsilon}^*_{loob}$. Hence, the weights in the .632+ bootstrap error are not fixed, but they depend on the estimated degree of overfitting.

6. Discussion

The bootstrap is a powerful methodology that allows statistical inferences for a wide range of problems that are extremely difficult or even impossible to tackle by other means. The bootstrap therefore plays an important role in the sciences, both in research and education.

The bootstrap is an extremely versatile methodology for statistical inference. It can be applied to problems that are intractable for standard approaches. For example, consider the problem of assessing the discriminatory power of a classification model based on genomic profiling. Typically, only a relatively small data set is available for both training and testing the model. The observed classification performance may be measured by balanced accuracy, area under the ROC curve, or any other metric that is deemed suitable

for the problem at hand. But the observed performance value is only an estimate of the true discriminatory power for new, unseen cases from the same population. How accurate is this estimate? The bootstrap enables us to address this question.

The bootstrap is not the only data resampling methodology for this purpose, though. Molinaro et al. compared different resampling techniques for estimating the prediction error in the context of the small-n-large-p problem, i.e., in problems where the number of cases (n) is much smaller than the number of features (p) [18]. Such problems are not at all uncommon in the life sciences, for example, in classification studies involving gene expression data. The bias of the .632+ bootstrap was found to be comparable to that of leave-one-out cross-validation and 10-fold cross-validation [18, 19]. According to Isaksson et al., however, both cross-validation and bootstrapping are unreliable for estimating the true prediction error when the data set is small, and they recommend Bayesian intervals based on a holdout test set [20].

The bootstrap spares us theoretical analysis, but at the expense of considerable computational costs due to the required repeated sampling. On the other hand, computational costs become increasingly negligible with the availability of relatively cheap computing power. Most statistical software tools nowadays include functions for bootstrapping; for example, the widely used language and environment R [21] provides the package `boot` [22] for parametric and non-parametric bootstrapping. In particular, a variety of bootstrap confidence intervals can be calculated with the function `boot.ci`. The R package `resample` provides various resampling functions for bootstrapping, jackknifing, and random permutation testing [23].

The bootstrap can also serve as a useful teaching tool in introductory statistics courses [12]. Concepts such as repeated random sampling, standard errors, etc. might be explained in a more accessible way if the instructor complements statistical theory and formulas with histograms of sampling distributions.

At the heart of the bootstrap is the idea that the available data set is an estimate of the population of interest, and that we can repeatedly take random samples from that estimate. But if the available data set is too small, it is unlikely to be a good estimate. Researchers need to keep in mind that no amount of resampling will ever be a panacea for the lack of data. Also, error estimates based on resampling are no substitute for independent validation studies using new data.

References

[1] D. Berrar, Introduction to the non-parametric bootstrap, in: S. Ranganathan, K. Nakai, C. Schönbach, M. Gribskov (Eds.), Encyclopedia of Bioinformatics and Computational Biology, 1st edition, Elsevier, 2018, pp. 766–773.

[2] B. Efron, Bootstrap methods: Another look at the jackknife, The Annals of Statistics 7 (1) (1979) 1–26.

[3] B. Efron, Nonparametric standard errors and confidence intervals, Canadian Journal of Statistics 9 (2) (1981) 139–158.

[4] B. Efron, C. Stein, The jackknife estimate of variance, The Annals of Statistics 9 (3) (1981) 586–596.

[5] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Statistical Science 1 (1) (1986) 54–75.

[6] B. Efron, Better bootstrap confidence intervals, Journal of the American Statistical Association 82 (397) (1987) 171–185.

[7] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, 1993.

[8] A. Davison, D. Hinkley, Bootstrap Methods and Their Applications, Cambridge University Press, 1997.

[9] D. Berrar, W. Dubitzky, Bootstrapping, in: W. Dubitzky, O. Wolkenhauer, K.-H. Cho, H. Yokota (Eds.), Encyclopedia of Systems Biology, Springer, 2013, pp. 158–163.

[10] T. Baguley, Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences, Palgrave Macmillan, 2012.

[11] P. Dixon, Bootstrap resampling, in: A. El-Shaarawi, W. Piegorsch (Eds.), Encyclopedia of Environmetrics, Wiley, 2006, pp. 212–220.

[12] T. Hesterberg, What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum, The American Statistician 69 (4) (2015) 371–386.

[13] T. DiCiccio, B. Efron, Bootstrap confidence intervals, Statistical Science 11 (3) (1996) 189–228.

[14] T. Hesterberg, Bootstrap, Wiley Interdisciplinary Reviews: Computational Statistics 3 (6) (2011) 497–526.

[15] T. Hesterberg, D. Moore, S. Monaghan, A. Clipson, R. Epstein, B. Craig, G. McCabe, Bootstrap methods and permutation tests, in: D. Moore, G. McCabe, B. Craig (Eds.), Introduction to the Practice of Statistics, 7th edition, W.H. Freeman, N.Y., 2010.

[16] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd edition, Springer, New York/Berlin/Heidelberg, 2008.

[17] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[18] A. Molinaro, R. Simon, R. Pfeiffer, Prediction error estimation: a comparison of resampling methods, Bioinformatics 21 (15) (2005) 3301–3307.

[19] R. Simon, Resampling strategies for model assessment and selection, in: W. Dubitzky, M. Granzow, D. Berrar (Eds.), Fundamentals of Data Mining in Genomics and Proteomics, Springer, 2007, pp. 173–186.

[20] A. Isaksson, M. Wallman, H. Göransson, M. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, Pattern Recognition Letters 29 (14) (2008) 1960–1965.

[21] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2017).
URL https://www.R-project.org/

[22] A. Canty, B. Ripley, boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20 (2017).
URL https://CRAN.R-project.org/package=boot

[23] T. Hesterberg, resample: Resampling Functions, R package version 0.4 (2015).
URL https://CRAN.R-project.org/package=resample