Introduction to the bootstrap

Outline

1. Introduction and motivation
2. Empirical cdf
3. The plug-in principle
4. The bootstrap principle
5. Estimating standard errors
6. The parametric bootstrap
7. Bias estimation
8. Jackknifing

Introduction and motivation

The bootstrap is a computer-intensive resampling method that was introduced by Efron in 1979 to overcome the problems of classical statistical inferential techniques that rely either on distributional assumptions or asymptotic theory.

The bootstrap is used for estimating standard errors and bias and for inferential purposes: obtaining confidence intervals and tests.

The bootstrap does **not** serve for obtaining better estimates.

It tries to analyse data that does not come from a known distribution model, or from some known distribution model, with techniques that are otherwise not theoretically supported.

The term bootstrap comes from the phrase to pull oneself up by one's bootstraps. It is based on the 18th century book *The Adventures of Baron Munchausen* by Rudolph E. Raspe.

The Baron had fallen into the bottom of a deep lake and came up with the idea of escaping by pulling on the laces of his own boots...

Introductory example

An *aspirin* study tracked strokes as well as heart attacks, with the following results.

	Stroke	**Subjects**
Aspirin	119	11037
Placebo	98	11034

$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1.21$$

An asymptotic 95% interval for the true value is θ is

$$0.93 < \theta < 1.59$$

This interval includes the *neutral* value 1, deducing that aspirin is not significantly better or worse than placebo.

Odds Ratio and Asymptotic Distribution

Given a 2×2 contingency table,

	C	**D**
A	n_{11}	n_{12}
B	n_{21}	n_{22}

The estimator used for the odds ratio is

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The distribution of this estimator is very **asymmetric**, so to consider an approximation to normal, it is better to take the transformation $\log(\widehat{\theta})$

Applying the *delta method*

`https://en.wikipedia.org/wiki/Delta_method`

An estimate of the standard error of $\log(\widehat{\theta})$ is

$$\widehat{\sigma}_{\log(\widehat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

so that the corresponding Wald interval is

$$\log(\widehat{\theta}) \pm z_{\frac{\alpha}{2}}\widehat{\sigma}_{\log(\widehat{\theta})}$$

If the exponential function (antilogarithm) of the ends is taken, the corresponding interval for θ is obtained. The test is somewhat **conservative** (the coverage probability is somewhat higher than the nominal level).

The bootstrap approach in the stroke example

We create two populations: the first consisting of 119 ones and $11037 - 119 = 10918$ zeros, and the second consisting of 98 ones and $11034 - 98 = 10936$ zeros.

We draw with replacement a sample of 11037 items from the first population, and a sample of 11034 items from the second population. Each of these is called a *bootstrap sample*.

From these we derive the bootstrap replicate of $\widehat{\theta}$:

$$\widehat{\theta}^* = \frac{\text{Proportion de } \textbf{ones} \text{ in bootstrap sample 1}}{\text{Proportion of } \textbf{ones} \text{ in bootstrap sample 2}}$$

This process is repeated many times (say $B = 1000$) to obtain a sample of 1000 values of $\widehat{\theta}^*$.

This sample of 1000 values of $\widehat{\theta}^*$ contains information that can be used to make inferences from the real data.

For example, Efron and Tibshirani obtain a sample standard deviation of around 0.17 and a confidence interval based on sample quantiles of around $(0.93; 1.60)$.

In this simple example, the confidence interval derived from the bootstrap agrees very closely with that derived from statistical theory.

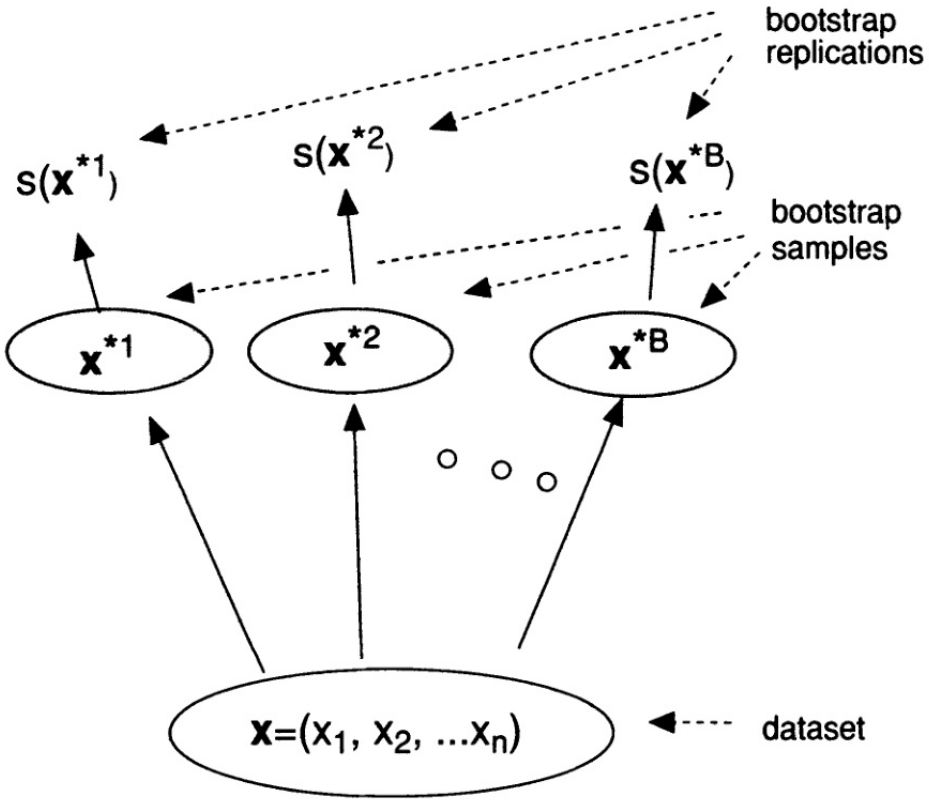
The bootstrap steps

Bootstrap methods are intended to simplify the calculation of classical inferences, producing them automatically even in situations much more complicated than the aspirin study.

In more detail, the bootstrap method works as follows:

1. A sample of size n is drawn from a population.
2. The statistic of interest is calculated for the sample.
3. The sample is resampled with replacement, meaning that some data points may be included more than once in the resampled data set.
4. The statistic of interest is calculated for the resampled data set.
5. Steps 3 and 4 are repeated a large number of times, typically 1000 or more.
6. The distribution of the statistic of interest from the resampled data sets is used to make inferences about the population parameter.

From *Efron and Tibshirani*, 1993



Empirical Distribution

In statistics, understanding the distribution of a data set is crucial for analysis and inference. The empirical distribution provides a straightforward way to describe the distribution of a sample.

It assigns equal probability to each observed value, making it a discrete distribution that corresponds to the observed data.

The empirical distribution is the discrete distribution that gives equal weight to each data point in a sample. This approach allows for the construction of a distribution function that closely follows the sample data, making no assumptions about the underlying population distribution.

The empirical cumulative distribution function offers a cumulative perspective on the distribution of data points. For a random sample $X_1, \ldots, X_n$, the empirical cdf is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X_i,\infty)}(x)$$

where $\mathbf{1}_{[X_i,\infty)}(x)$ is an indicator function that equals 1 if $x \geq X_i$ and 0 otherwise. This function steps up by $\frac{1}{n}$ at each data point X_i, providing a piecewise constant function that approximates the true cdf as n increases.

The Glivenko-Cantelli Theorem

The Glivenko-Cantelli Theorem, also known as the Fundamental Theorem of Statistics, is a cornerstone in the study of empirical distributions. Establishes a strong uniform convergence of the empirical cdf to the true population cdf. Formally, the theorem states:

$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|=0\right)=1$$

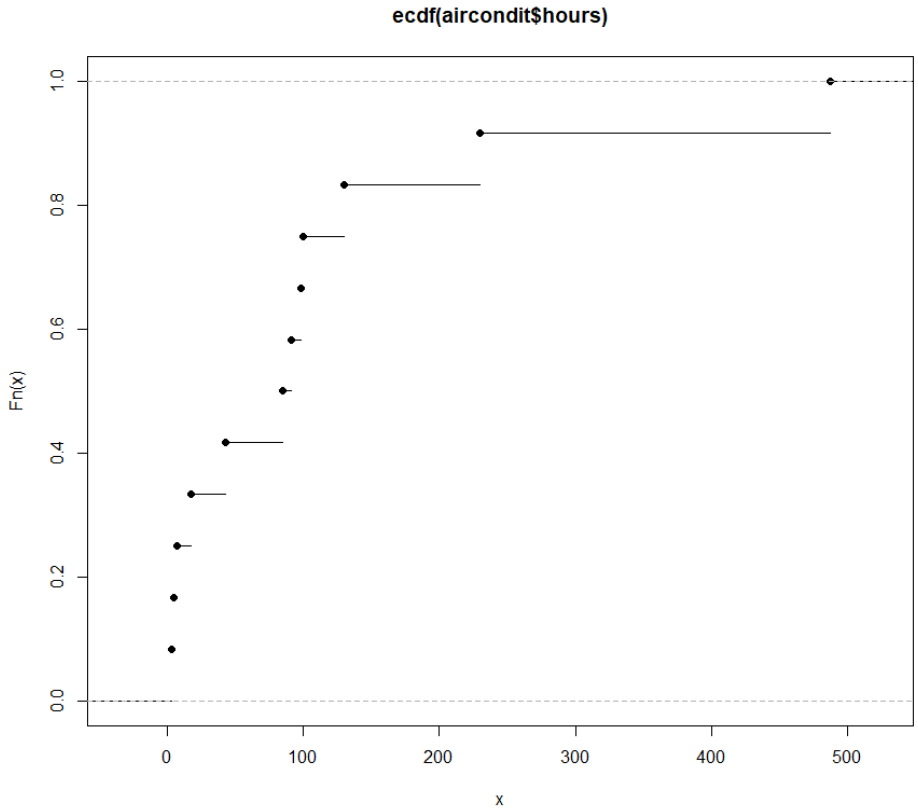
where F denotes the population cdf. This result implies that, with a sufficiently large sample size, the empirical cdf will converge uniformly to the true cdf of the population from which the sample is drawn.

The theorem guarantees that almost surely the maximum difference between the empirical cdf and the population cdf vanishes as the sample size grows to infinity.

The Glivenko-Cantelli Theorem has profound implications for statistical practice. It underpins the validity of using empirical distributions to estimate the population distribution and supports the development of statistical methods that rely on empirical cdfs, such as nonparametric tests and confidence intervals for the population cdf.

Example

```
data(aircondit, package="boot")
plot(ecdf(aircondit$hours))
```



The Dvoretzky-Kiefer-Wolfowitz (DKW) theorem

Let $X_1, \ldots, X_n$ be a random sample from a random variable with distribution function F. Then, for every $\varepsilon > 0$,

$$P\left(\sup_x \left|F(x) - \widehat{F}_n(x)\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

This inequality shows that the probability of the supreme (the maximum difference) between the true distribution function $F(x)$ and the empirical distribution function $\widehat{F}_n(x)$ being greater than ε is bounded by $2e^{-2n\varepsilon^2}$.

This result does not assume anything specific about the distribution F, making it a nonparametric result.

Confidence Intervals for $\widehat{F}$

Thus, confidence intervals can be constructed for $\widehat{F}$, based on the empirical distribution function.

We define the lower and upper bounds $L(x)$ and $U(x)$ of the confidence interval for any x as follows:

- $L(x) = \max\left\{\widehat{F}_n(x) - \varepsilon_n, 0\right\}$

- $U(x) = \min\left\{\widehat{F}_n(x) + \varepsilon_n, 1\right\}$

  where $\varepsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$.

This formula for ε_n ensures that, for any distribution function F and for any x,

$$P\left(L(x) \leq F(x) \leq U(x)\right) \geq 1 - \alpha$$

This means that the probability that $F(x)$ lies within the interval $[L(x), U(x)]$ is at least $1 - \alpha$, providing a way to construct confidence intervals around the empirical distribution function that hold with a specified confidence level $1 - \alpha$.

The bounds can be easily programmed, but note that the intervals may be wide for small sample sizes.

Alternatives to the empirical cdf

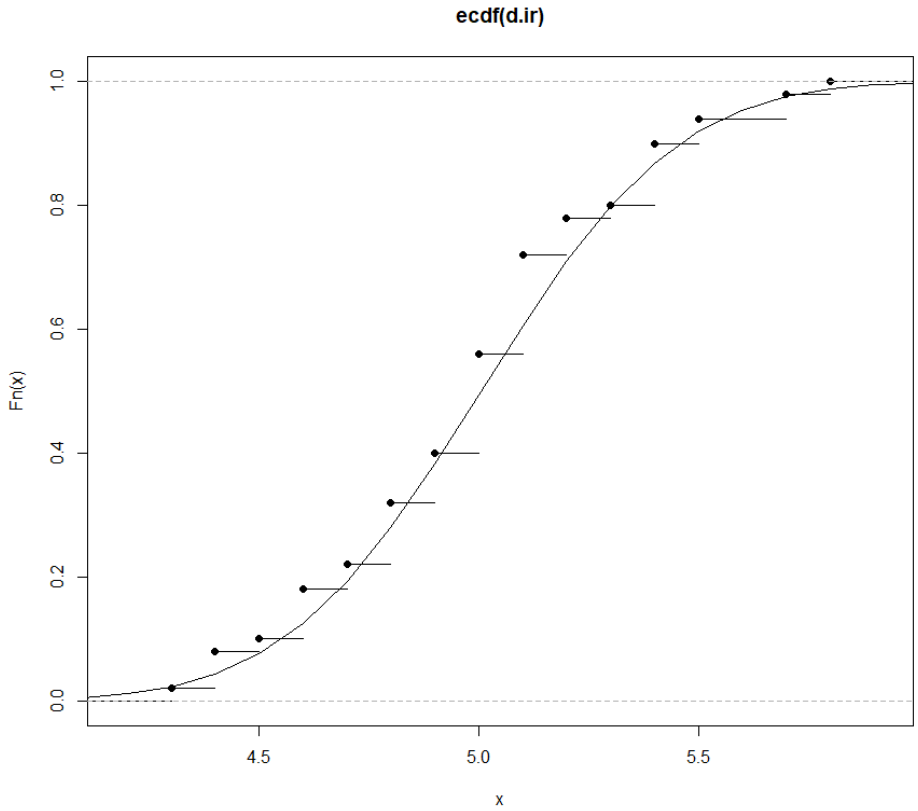
An alternative to the empirical cumulative distribution function is to use a smoothed version of it.

This can be achieved by applying smoothing techniques, such as kernel smoothing, to the empirical data, resulting in a more continuous and visually appealing estimate of the distribution function.

Alternatively, if the data are assumed to come from a known distribution model, the corresponding cdf of that model can be used, with parameters estimated from the data. This approach is particularly useful when the underlying distribution of the data is well understood or can be approximated by a common distribution model, such as the normal distribution.

The figure illustrates how to compare the empirical cdf with a parametric cdf, specifically the normal distribution, using the *iris* dataset.

It overlays a line representing the cdf of the normal distribution, with the mean and standard deviation parameters estimated from the *setosa* sepal lengths.



Plug-in methods

The **plug-in** principle is a fundamental concept in statistical estimation. It posits that if a parameter of interest in a statistical model is unknown, we can replace it with an estimate. This principle is widely applicable in various statistical methodologies, including non-parametric statistics, where the distribution of the data might not be fully specified.

In the context of Probability and Statistics, many parameters of a distribution can be considered as *functionals* of the cumulative distribution function (cdf).

In this setting, a functional is a function T that acts on cdfs, mapping them to real numbers. This means that the parameter can be represented as a certain function of the entire distribution.

Two common examples of such parameters are the **mean** and the **variance** of a distribution. These can be expressed as:

$$\begin{aligned} \mu(F) &= \int x\, dF(x), \\ \sigma^2(F) &= \int (x-\mu(F))^2\, dF(x), \end{aligned}$$

where F denotes the cdf of the distribution, and the integral is calculated over the entire space where F is defined.

The mean $\mu(F)$ is the expected value of the random variable X with distribution F, and the variance $\sigma^2(F)$ measures the dispersion of X around its mean.

The Plug-in Principle for Estimation

According to the plug-in principle, when a parameter such as $\mu(F)$ or $\sigma^2(F)$ is unknown, we estimate it using sample data. This involves replacing the theoretical cdf F with the empirical cdf F_n derived from the sample. Consequently, the plug-in estimates for the mean and variance are given by:

$$\mu(F_n) = \overline{X}_n,$$

and

$$\sigma^2(F_n) = \frac{n-1}{n} S_n^2,$$

where $\overline{X}_n$ is the sample mean, S_n^2 is the sample variance, and n is the sample size.

The variance of the sample is often multiplied by $\frac{n-1}{n}$ to make it an unbiased estimator of the variance of the population in the case of a normally distributed population.

An important property of plug-in estimators is their consistency, which means that as the sample size n increases, the estimator converges in probability to the true value of the parameter.

If the functional T is smooth (continuous and differentiable), the Glivenko-Cantelli theorem provides a strong foundation for this consistency. Specifically, for a smooth T, the statistic $T(F_n)$ is a consistent estimator of $T(F)$, ensuring that our plug-in estimates become increasingly accurate as we collect more data.

The bootstrap principle

This method involves resampling with replacement from the sample and recalculating the statistic many times to create a distribution known as the bootstrap distribution. The bootstrap distribution provides useful information on the variability and bias of the estimate, allowing for more accurate inference about the population parameter.

Let F denote the true population distribution from which we are sampling. Our goal in statistical inference is often to estimate a parameter θ that describes some characteristic of this population. The parameter θ is defined by a functional T applied to the population distribution, such that $\theta = T(F)$.

Empirical Distribution and Estimation

Given a random sample $X_1, \ldots, X_n$ drawn from F, we can construct the empirical distribution F_n, which assigns equal probability to each observed sample point. To estimate the parameter θ, we often use a statistic S calculated from the sample, where $\theta \approx S(X_1, \ldots, X_n)$. This statistic S serves as our estimate of θ and can be thought of as an application of a function to the sample data, potentially analogous to $T(F_n)$, the application of the functional to the empirical distribution. However, it is important to note that S does not necessarily equal $T(F_n)$. The choice of S depends on the parameter of interest and the characteristics of the population distribution.

Examples of Estimators

For example, if we are interested in estimating the mean of the population, we might use the sample mean as our statistic S. However, in cases where the population distribution is known to be symmetric but not necessarily normal, we might choose the sample median as a robust estimator of the population mean. Alternatively, a trimmed mean, which removes outliers before calculating the mean, can be used to estimate the population mean, especially in distributions with heavy tails.

The bootstrap procedure involves repeatedly sampling, with replacement, from the empirical distribution F_n to generate many bootstrap samples. For each bootstrap sample, we calculate our statistic of interest. The distribution of these bootstrap statistics approximates the sampling distribution of the statistic under the true population distribution F. This bootstrap distribution allows us to estimate the standard error, confidence intervals, and bias of our statistic, providing a deeper understanding of its behavior and how it relates to the true parameter θ.

How the nonparametric bootstrap works

Suppose that we have a sample from a population, but we know nothing about the population's distribution. The fundamental idea of nonparametric bootstrap is to use the sample itself as a population model.

This approach involves resampling with replacement from the original sample to generate new bootstrap samples. Each bootstrap sample is the same size as the original sample and is denoted by $X_1^*, \ldots, X_n^*$.

The distribution of these bootstrap samples, represented by F_n^*, serves as an empirical approximation of the population distribution, F_n. This bootstrap distribution allows us to estimate the

variability and distribution of a statistic without relying on traditional parametric assumptions.

The bootstrap estimate of a parameter, $\widehat{\theta}^*$, is calculated using the statistic $S(X_1^*, \ldots, X_n^*)$ applied to a bootstrap sample. This contrasts with the plug-in estimator $T(F_n^*)$, which directly uses the empirical bootstrap distribution.

The Bootstrap Algorithm

The bootstrap method can be summarized in the following steps:

1. Generate a bootstrap sample by resampling with replacement from the empirical distribution of the original sample. This sample is denoted $X_1^*, \ldots, X_n^*$.

2. Compute the bootstrap estimate of the parameter $\widehat{\theta}^* = S(X_1^*, \ldots, X_n^*)$. This involves applying the statistic of interest to the bootstrap sample.

3. Repeat steps 1 and 2 a large number of times, denoted by B.

This process generates a distribution of the bootstrap estimates, which approximates the sampling distribution of the estimator $\widehat{\theta}$.

This distribution can then be used to estimate various characteristics of the estimator, such as its bias, variance, confidence intervals, or to perform hypothesis testing.

Estimating standard errors with bootstrap

Standard errors measure the variability or spread of a sampling distribution and are crucial in hypothesis testing and constructing confidence intervals.

Consider a dataset consisting of observations $x_1, x_2, \ldots, x_n$ of size n. Let F_n denote the empirical distribution function of the observed sample. The goal of the bootstrap method is to estimate the standard error of a statistic $\widehat{\theta}$, which is a function of the sample, without making strong assumptions about the underlying population distribution.

Algorithm

To estimate the standard error of $\widehat{\theta}$ using the bootstrap method, we can follow the following steps:

1. **Generate Bootstrap Samples:** Draw B independent bootstrap samples from the original data set. Denote these samples as $X_1^*, X_2^*, \ldots, X_B^*$. Each bootstrap sample is of the same size

n as the original data set and is drawn with replacement. This means that each sample can include duplicates of some observations and potentially exclude others.

2. **Compute the Statistic:** For each bootstrap sample $b \in \{1, 2, \ldots, B\}$, compute the statistic of interest $\widehat{\theta}^{*b} = S(x^{*b})$. The function S represents the statistical operation or formula applied to the bootstrap sample to calculate the statistic.

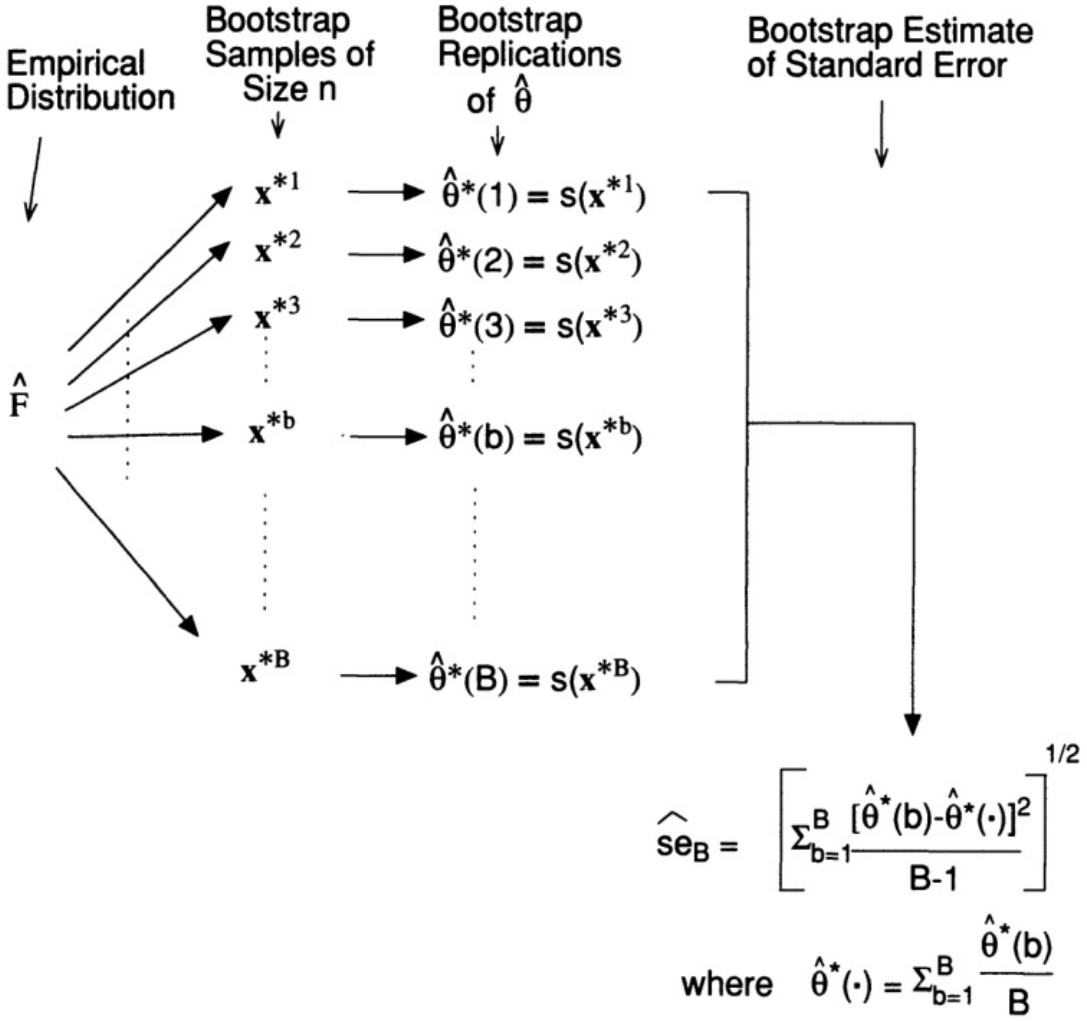
3. **Estimate the Standard Error:** The bootstrap estimate of the standard error of $\widehat{\theta}$, denoted as $\widehat{\sigma}_B$, is calculated using the formula:

$$\widehat{\sigma}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^{*b} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{*b}\right)^2}.$$

This formula computes the standard deviation of bootstrap estimates, providing an estimate of the standard error of $\widehat{\theta}$.

It is important to note that as B, the number of bootstrap samples becomes very large, $\widehat{\sigma}_B$ converges to the ideal bootstrap estimate of the standard error, $se_{F_n}(\widehat{\theta}^*)$. This convergence ensures the reliability of the bootstrap method in estimating the standard error without requiring strict assumptions about the form of the population distribution.

From Efron and Tibshirani, 1993



Bias estimation

The bias of an estimator is a measure of how far the estimator's expected value deviates from the true value of the parameter being estimated. The bias of estimator S with respect to F is defined as:

$$\text{Bias}_F = E_F\left[S\left(X_1, \ldots, X_n\right)\right] - T(F),$$

where $E_F[\cdot]$ denotes the expected value under the distribution F.

In practice, since F is unknown, we approximate it with F_n, the empirical distribution, to estimate the bias. Thus, the estimated bias can be expressed as:

$$\text{Bias}_{F_n} = E_{F_n}\left[S\left(X_1^*, \ldots, X_n^*\right)\right] - S(x_1, x_2, \ldots, x_n),$$

where $\{X_1^*, \ldots, X_n^*\}$ denotes a bootstrap sample.

Algorithm

1. Draw B independent bootstrap samples $x^{*1}, x^{*2}, \ldots, x^{*B}$, each of size n, drawn with replacement from the original sample $x_1, x_2, \ldots, x_n$.

2. For each bootstrap sample $b \in \{1, 2, \ldots, B\}$, compute $\widehat{\theta}^{*b} = S(x^{*b})$, the estimate of θ based on the b-th bootstrap sample.

3. Approximate the bias of the estimator based on the bootstrap samples as follows:

$$\widehat{\text{Bias}}_B = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{*b} - S(x_1, x_2, \ldots, x_n),$$

   where $\widehat{\text{Bias}}_B$ is the bootstrap estimate of the bias of S.

The parametric bootstrap

The parametric bootstrap assumes that the population F_θ comes from a specified parametric distribution model. However, the parameters θ of this model are not known and must be estimated from the data. We denote these estimated parameters as $\widehat{\theta}$, and the estimated distribution as $F_{\hat{\theta}}$.

The parametric bootstrap then involves resampling from this estimated distribution to generate new samples, which can be used to approximate the sampling distribution of the estimator or to perform hypothesis testing.

Algorithm

1. **Collect data:** Begin by considering the sample $x_1, x_2, \ldots, x_n$ consisting of the values observed n. This sample is assumed to be drawn from the population distribution F_θ, where θ represents the unknown parameters of this distribution.

2. **Estimate parameters:** Use the collected sample to estimate the parameters θ of the population distribution. This is achieved by applying a suitable parameter estimation method (such as the maximum likelihood method) to obtain $\widehat{\theta}$, the estimate of θ.

   The distribution F_θ, with θ replaced by the estimated values $\widehat{\theta}$, is denoted as $F_{\hat{\theta}}$. This estimated distribution serves as a model for generating resampled data.

3. **Resample and apply bootstrap:** With the estimated distribution $F_{\hat{\theta}}$ at hand, we can now apply the bootstrap method. This involves generating a large number of resamples from $F_{\hat{\theta}}$, each of which is the same size as the original sample.

   For each resample, calculate the statistic of interest (e.g., the mean, variance, or another parameter). This step can be applied to various bootstrap algorithms, including those used to estimate standard errors, construct confidence intervals, or perform hypothesis tests.

Note on the Algorithm The parametric bootstrap is particularly useful when the underlying distribution of the data is known or can be reasonably assumed. It leverages the parametric model to generate resamples more efficiently and can provide more accurate inference than the non-parametric bootstrap in situations where the parametric model is appropriate.

However, its performance is highly dependent on the correctness of the assumed model. If the model is poorly chosen, the results of the parametric bootstrap may be misleading.

Jackknife

The Jackknife is a resampling technique that is used to estimate the bias and standard error of statistical estimators. Developed by Quenouille in 1949, it serves as a simpler alternative for bias correction and variance estimation. The Jackknife method is particularly useful because of its simplicity and general applicability to a wide range of problems.

The essence of the Jackknife method lies in systematically resampling the original data set to form new *subsamples* by omitting **one** observation at a time. This process is known as the *leave-one-out* strategy. Through this approach, it assesses the variability of the statistical estimate and provides a way to adjust for bias.

Consider a sample of n observations denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The sample with the i^{th} observation removed, denoted by $\mathbf{x}_{(i)}$, is thus given by:

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

For a estimator S, applied to the sample $\mathbf{x}$, the estimator based on $\mathbf{x}_{(i)}$ is represented as:

$$\widehat{\theta}_{(i)} = S(\mathbf{x}_{(i)}).$$

This process is repeated for each observation in the sample, leading to a series of Jackknife replications.

Jackknife estimate of bias

The bias of an estimator $\widehat{\theta}$ can be estimated using the Jackknife method as follows:

$$\widehat{\text{Bias}}_{jack} = (n-1)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)} - \widehat{\theta}\right),$$

where $\widehat{\theta}$ is the estimate of θ using the full sample, and $\widehat{\theta}_{(i)}$ is the estimate obtained by omitting the i^{th} observation.

This formula adjusts for bias by comparing the average of the Jackknife estimates with the original estimate.

Jackknife estimate of standard error

The standard error of the estimator $\widehat{\theta}$ can be estimated using the Jackknife method as:

$$\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(i)} - \frac{1}{n}\sum_{j=1}^{n}\widehat{\theta}_{(j)}\right)^2} = \frac{n-1}{\sqrt{n}} se\left(\widehat{\theta}_{(\cdot)}\right),$$

where $se(\widehat{\theta}_{(\cdot)})$ denotes the standard error of the Jackknife estimates.

This expression quantifies the variability of the estimator $\widehat{\theta}$ by considering the spread of the Jackknife estimates around their mean.

The Jackknife is a powerful tool for bias correction and estimating the standard error of statistical estimators. Its simplicity and general applicability make it valuable in statistical analysis, especially when dealing with small sample sizes.