Resampling Methods.

Anatoli Iambartsev
IME-USP

Resampling Methods.

The resampling method is tied to the *Monte-Carlo simulation*, in which researchers "make up" data and draw conclusions based on many possible scenario. The fundamental difference between Monte Carlo simulation and resampling is that in the former data could be totally hypothetical, while in the latter the simulation must be based upon some real data.

1. **Bootstrap.**

2. **Jackknife.**

3. **Cross-validation.**

4. **Randomization exact test**. Also known as the **permutation test**.

1. **Bootstrap.** 2. **Jackknife.** 3. **Cross-validation.** 4. **Permutation test**.

Bootstrap means that one available sample given rise to many others by resampling. While the original objective of cross-validation is to verify replicability of resultas and that of Jackknife is to detect outliers.

The principles of cross-validation, Jackknife, and bootstrap are very similar, but bootstrap overshadows the others for it is a more thorough procedure in the sense that it draws many more sub-samples then the others. Through simulations Fan and Wang (1996) found that the bootstrap technique provides less biased and more consistent results than the Jackknife method does. Nevertheless, Jackknife is still useful in EDA for assessing how each sub-sample affects the model.

Resampling Methods.

In probability:

cdf $F(x) \Longrightarrow$ parameter of interest $\theta = s(F)$

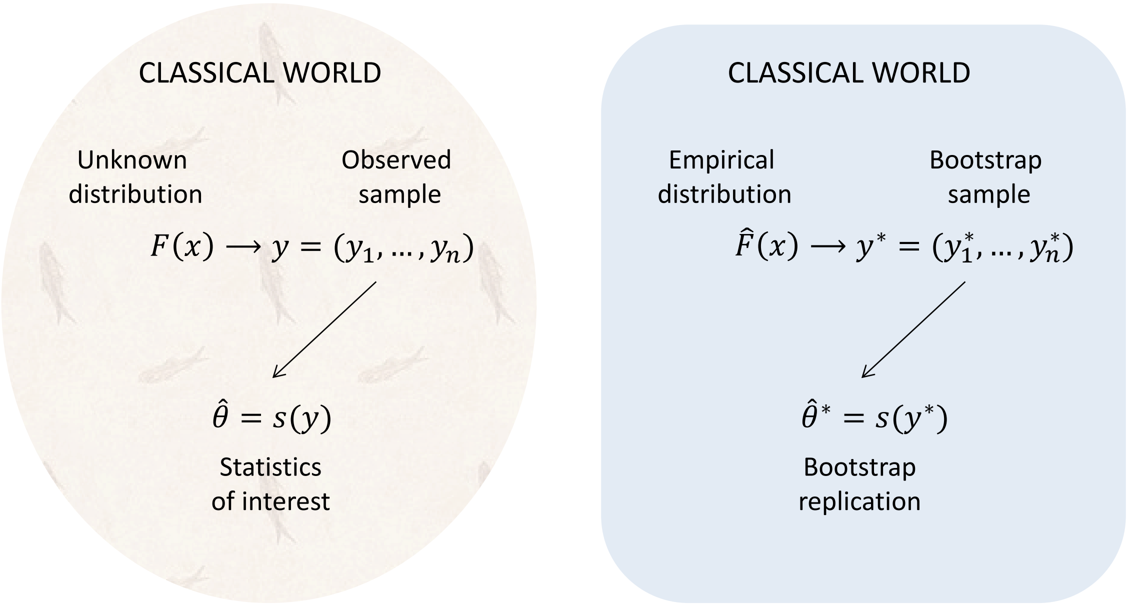
In statistics: given a sample $y_1, \ldots, y_n$

cdf $\widehat{F}(x) \Longrightarrow$ parameter of interest $\widehat{\theta} = s(\widehat{F}), s.e.(\widehat{\theta})$

How to estimate standard error $s.e.(\widehat{\theta})$ when we have no the exact analytic formula?

We can use the bootstrap method. The bootstrap method "consider" an observed sample $x_1, \ldots, x_n$, as an total "population", i.e. the population cdf for bootstrap is empirical cdf $\widehat{F}$.

Aula 5. Monte Carlo Method III. Resampling.



CLASSICAL WORLD

Unknown distribution

Observed sample

$F(x) \longrightarrow y = (y_1, \dots, y_n)$

$\hat{\theta} = s(y)$

Statistics of interest

CLASSICAL WORLD

Empirical distribution

Bootstrap sample

$\hat{F}(x) \longrightarrow y^* = (y_1^*, \dots, y_n^*)$

$\hat{\theta}^* = s(y^*)$

Bootstrap replication

Bootstrap: parametric non-parametric.

A bootstrap resampling can be executed in two forms: parametric and non-parametric.

- **Non-parametric bootstrap**: The sampling is based on the empirical cdf $\widehat{F}$. Sampling from a data (with reposition).
- **Parametric bootstrap**: The sampling is based on cdf $F(\widehat{\theta})$.

Bootstrap simulation.

Observed data:

2.4, 1.5, 3.7, 1.9, 2.5 $\Longrightarrow \quad \widehat{\theta} = s(y) = \bar{y} = 2.40$

Bootstrap samples:

1.5, 1.9, 1.5, 2.4, 3.7 $\Longrightarrow \quad \widehat{\theta}^*(1) = 2.20$

1.9, 3.7, 2.4, 2.4, 1.5 $\Longrightarrow \quad \widehat{\theta}^*(2) = 2.38$

2.4, 1.9, 2.5, 2.4, 3.7 $\Longrightarrow \quad \widehat{\theta}^*(3) = 2.58$

$\vdots$

3.7, 1.9, 3.7, 2.5, 1.5 $\Longrightarrow \quad \widehat{\theta}^*(1) = 2.66$

Bootstrap standard error.

From bootstrap sampling we can estimate any aspect of the distribution of $\widehat{\theta} = s(y)$ (which is any quantity computed from the data $y = (y_1, \ldots, y_n)$, for example its standard error is

$$s.e.b.(\widehat{\theta}) = \Big(\frac{1}{B-1} \sum_{b=1}^{B} \big(\widehat{\theta}^*(b) - \widehat{\theta}^*(\cdot) \big)^2 \Big)^{1/2}$$

where $\widehat{\theta}^*(b)$ is the bootstrap replication of $s(y)$ and

$$\widehat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}^*(b).$$

Jackknife.

In some sense the bootstrap method is a generalization of the method jackknife, in the sense that the resampling is made randomly and not deterministically as in jackknife “leave-one-out”.

Jackknife.

1. We have a sample $y = (y_1, \ldots, y_n)$ and estimator $\widehat{\theta} = s(y)$.

2. Target: estimate the bias and standard error of the estimator.

3. The *leave-one-out* observation samples

$$y_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n),$$

   for $i = 1, \ldots, n$ are called *jackknife samples*.

4. Jackknife estimators are $\widehat{\theta}_{(i)} = s(y_{(i)})$.

Jackknife bias-reduction. Quenouille bias.

The bias of $\widehat{\theta} = s(y)$ is defined as

$$bias_J(\widehat{\theta}) = (n-1)\big(\widehat{\theta}_{(\cdot)} - \widehat{\theta}\big),$$

where $\widehat{\theta}_{(\cdot)}$ is the average of Jackknife estimators $\widehat{\theta}_{(i)}$

$$\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}.$$

This leads to a bias-reduced *jackknife estimator* of parameter θ

$$\widehat{\theta}_J = \widehat{\theta} - bias_J(\widehat{\theta}) = n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}$$

Jackknife bias-reduction. Quenouille bias.

Why it works? In general the estimators are the maximum-likelihood estimators, and the expectation $E_n := \mathbb{E}_F \widehat{\theta}(X_1, \ldots, X_n)$ can be represented as

$$E_n = \mathbb{E}_F \widehat{\theta}(F) = \theta(F) + \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + o\Big(\frac{1}{n^2}\Big)$$

where a_1, a_2 do not depend on n, but usually are unknown in practice. Note that for any i

$$E_{n-1} = \theta(F) + \frac{a_1(F)}{n-1} + \frac{a_2(F)}{(n-1)^2} + o\Big(\frac{1}{(n-1)^2}\Big) = \mathbb{E}_F \widehat{\theta}_{(\cdot)}.$$

Remember

$$\widehat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{(i)}.$$

Jackknife bias-reduction. Quenouille bias.

Thus

$$\mathbb{E}_F(\widehat{\theta}_J) = \mathbb{E}_F(n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}) = nE_n - (n-1)E_{n-1}$$
$$= \theta(F) - \frac{a_2(F)}{n(n-1)} + a_3(F)\Big(\frac{1}{n^2} - \frac{1}{(n-1)^2}\Big) + \dots$$

The bias of $\widehat{\theta}_J$ is of the order $O(1/n^2)$ comparing with original which is of the order $O(1/n)$.

Jackknife bias-reduction. Example.

Consider estimator of mean $\widehat{\theta} = \bar{y}$ and $\widehat{\theta}_{(\cdot)} = \bar{y}$, thus

$$bias_J(\widehat{\theta}) = (n-1)\big(\widehat{\theta}_{(\cdot)} - \widehat{\theta}\big) = 0.$$

Jackknife bias-reduction. Example.

Let $\theta(F)$ be variance $\theta(F) = \int (x - \mathbb{E}_F(X))^2 dF$ and $\widehat{\theta} = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$. The simple calculations provide

$$bias_J(\widehat{\theta}) = (n-1)\big(\widehat{\theta}_{(\cdot)} - \widehat{\theta}\big) = -\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

providing

$$\widehat{\theta}_J = \widehat{\theta} - bias_J(\widehat{\theta}) = n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

In this case $E_n = \theta - \theta/n$, and $a_1(F) = -\theta, a_i(F) = 0$ for all $i > 1$.

Jackknife standard error.

[Ef] Tukey (1958) suggested how the recomputed statistics $\widehat{\theta}_{(i)}$ could also provide a nonparametric estimate of variance. Let

$$\mathbb{V}ar_F(\widehat{\theta}) = \mathbb{E}_F\Big(\widehat{\theta}(X_1, \ldots, X_n) - \mathbb{E}_F\widehat{\theta}\Big)^2.$$

The Tukey's jackknife standard deviation estimation is

$$s.d.j.(\widehat{\theta}) = \Big(\frac{n-1}{n}\sum_{i=1}^{n}\big(\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\big)^2\Big)^{1/2}$$

Bootstrap bias-reduction.

Let $\widehat{\theta}$ be a consistent estimator, but biased. Target: to reduce the bias of the estimator.

The bias of $\widehat{\theta}$ is the systematic error $bias = \mathbb{E}_F\widehat{\theta} - \theta$. Em general the bias depends on the unknown parameter θ, because why we cannot to have $\widehat{\theta} - bias$.

Consider the following bootstrap bias correction

$$\widehat{\theta}_B = \widehat{\theta} - bi\widehat{a}s.$$

where

$$bi\widehat{a}s = \mathbb{E}_{\widehat{F}}\widehat{\theta} - \widehat{\theta} = \widehat{\theta}^*_{(\cdot)} - \widehat{\theta},$$

where $\widehat{\theta}^*_{(\cdot)}$ is the average of bootstrap estimators, i.e.

$$\widehat{\theta}^*_{(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*_b.$$

Bootstrap bias-reduction.

Thus

$$\widehat{\theta}_B = \widehat{\theta} - \widehat{bias} = 2\widehat{\theta} - \widehat{\theta}^*_{(\cdot)}$$

In terms of asymptotic behaviors, the jackknife and traditional (linearization) estimators are usually first order asymptotically equivalent. Some limited empirical results show that the jackknife variance estimator is less biased but more variable than the linearization estimator.

Bootstrap hypotheses testing.

- Set the two hypotheses.

- Choose a test statistic T that can discriminate between the two hypotheses. We do not care that our statistic has a known distribution under the null hypothesis.

- Calculate the observed value t_{obs} of the statistic for the sample.

- Generate B samples from the distribution implied by the null hypothesis.

- For each sample calculate the value $t_{(i)}$ of the statistic, $i = 1, \ldots, B$.

- Find the proportion of times the sampled values are more extreme than the observed.

- Accept or reject according to the significance level.

Bootstrap hypotheses testing.

Suppose two samples $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$. We wish to test the hypothesis that the mean of two populations are equal, i.e.

$$H: \ \mu_x = \mu_y \quad \text{vs} \quad A: \ \mu_x \neq \mu_y$$

Use as a test statistic $T = \bar{x} - \bar{y}$.

Under the null hypothesis a good estimate of the population distribution is the combined sample $z = (x_1, \dots, x_n, y_1, \dots, y_m)$

For each of the bootstrap sample calculate $T^*_{(i)}$, $i = 1, \dots, B$.

Estimate the p-value of the test as

$$\widehat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}(T^*_{(i)} \geq t_{obs})$$

Other test statistics are applicable, as for example t-statistics.

Failure of bootstrap.

- Small data sets (because empirical distribution is not good approximation of F).
- Infinite moments.
- Dependences structure (time series, spatial problems). Bootstrap is based on the assumption of independence.
- Dirty data: if outliers exist in our sample, clearly we we do not sample from a good estimate of F and we add variability in our estimates.
- Unsmooth quantities: there are plenty of theoretical results that relate the success of bootstrap with the smoothness of the functional under consideration.
- Multivariate data: when the dimensions of the problem are large, then empirical distribution become less good as an estimate of F. This may cause problems.

Choice of B.

The choice of B depends on

- Computer availability (efficiency).
- Type of the problem: while $B = 1000$ suffices for estimating standard errors, perhaps t is not enough for confidence interval.
- Complexity of the problem.

Randomization test.

- Also called *permutation test*, *randomization exact test*, *exact test*.
- Introduced by Fisher and Pitman in the 1930s.
- Usually require only a few weak assumptions.
  - underlying distributions are symmetric;
  - the alternatives are shifts in value.

Steps of randomization test.

- Choose a test statistics which will distinguish the hypothesis from the alternative.

- Compute the test statistic for the original set (labeling) of the observations.

- Compute the test statistic for all possible rearrangements (permutations) of the observations.

- Obtain the permutation distribution of test statistic.

- Make a decision: reject the hull hypothesis if the value of the test statistic for the original labeling (original data) is an extreme value in the permutation distribution of the statistic. Otherwise, accept the null hypothesis.

The number of permutations.

A	B	C	D	E	F
121	118	110	34	12	22

How many permutations exists?

$$\binom{6}{3} = 20$$

$$\binom{52}{18} = 4.27 \times 10^{13}$$

Bootstrap. Theoretical questions [CL, p.5].

Let $T(\cdot)$ be a functional of interest, for example estimator of a parameter. We are interested in estimation of $T(F)$, where F is population distribution. Let F_n be an empirical distribution based on sample $x = (x_1, \ldots, x_n)$. Bootstrap:

1. generate a sample $x^* = (x_1^*, \ldots, x_n^*)$ with replacement from the empirical distribution F_n for the data (boostrap sample);

2. compute $T(F_n^*)$ the bootstrap estimate of $T(F)$. This is a replacement of the original sample x with a bootstrap sample x^* and the bootstrap estimate of $T(F)$ in place of the sample estimate of $T(F)$;

3. M times repeat steps 1 and 2 where M is large, say 100000.

Bootstrap. Theoretical questions [CL, p.5].

Now a very important thing to remember is that with the Monte Carlo approximation to the bootstrap, there are two sources of error:

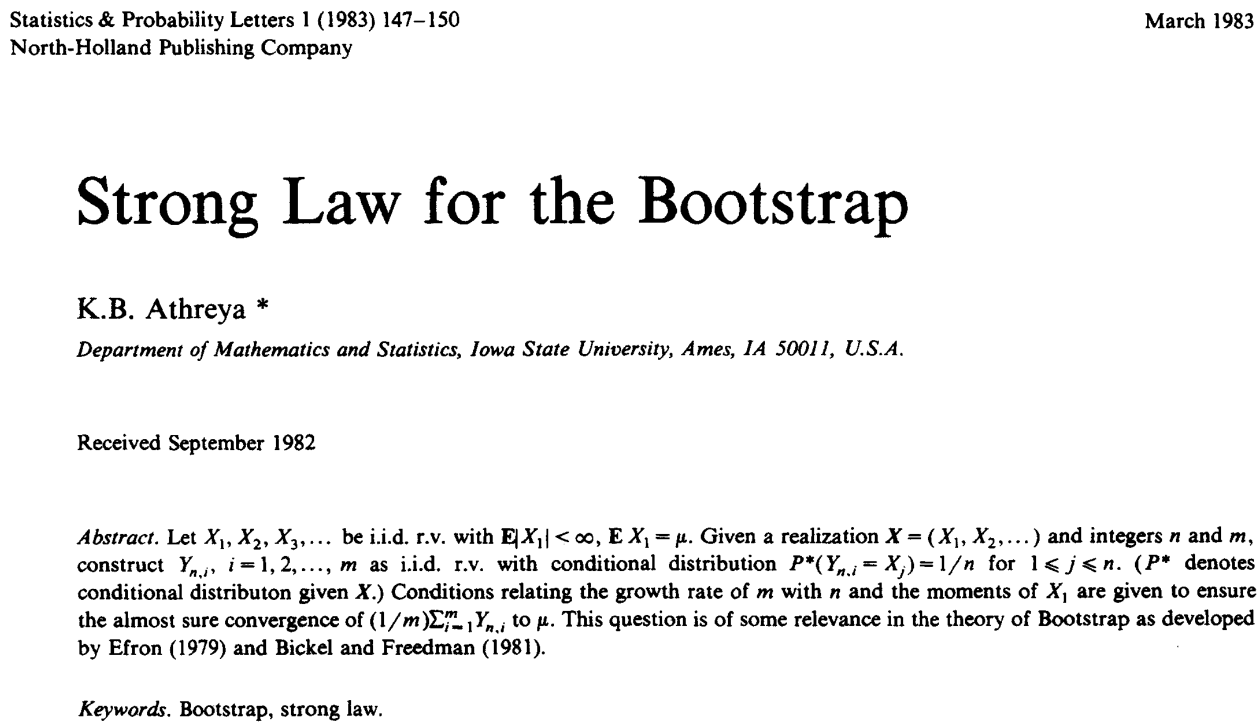
1. the Monte Carlo approximation to the bootstrap distribution, which can be made as small as you like by making M large;

2. the approximation of the bootstrap distribution F_n^* to the population distribution F.

If $T(F_n^*)$ converges to $T(F)$ as $n \to \infty$, then bootstrapping works.

Bootstrap. Theoretical questions [CL, p.5].

“If $T(F_n^*)$ converges to $T(F)$ as $n \to \infty$, then bootstrapping works. It is nice that this works out often, but it is not guaranteed. We know by a theorem called the Glivenko-Cantelli theorem that F_n converges to F uniformly. Often, we know that the sample estimate is consistent (as is the case for the sample mean). So, (1) $T(F_n)$ converges to $T(F)$ as $n \to \infty$. But this is dependent on smoothness conditions on the functional T. So we also need (2) $T(F_n^*) - T(F_n)$ to tend to 0 as $n \to \infty$. In proving that bootstrapping works (i.e., the bootstrap estimate is consistent for the population parameter), probability theorists needed to verify (1) and (2). One approach that is commonly used is by verifying that smoothness conditions are satisfied for expansions like the Edgeworth and Cornish-Fisher expansions. Then, these expansions are used to prove the limit theorems.”



Statistics & Probability Letters 1 (1983) 147–150
North-Holland Publishing Company

March 1983

Strong Law for the Bootstrap

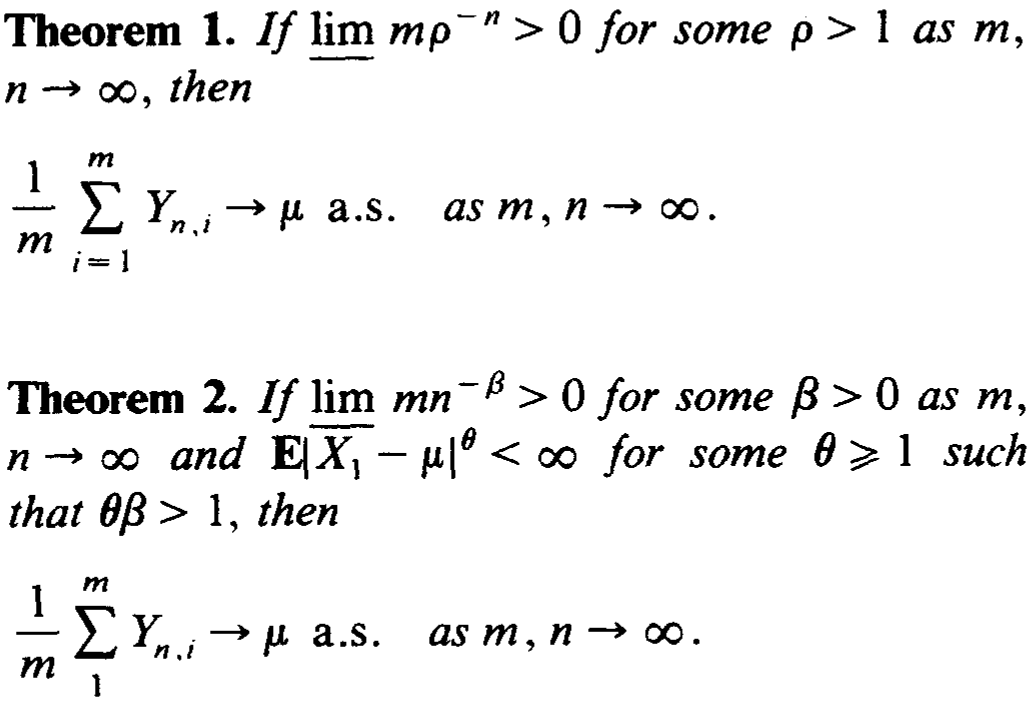
K.B. Athreya *

Department of Mathematics and Statistics, Iowa State University, Ames, IA 50011, U.S.A.

Received September 1982

Abstract. Let $X_1, X_2, X_3, \dots$ be i.i.d. r.v. with $\mathbf{E}|X_1| < \infty$, $\mathbf{E}\,X_1 = \mu$. Given a realization $\boldsymbol{X} = (X_1, X_2, \dots)$ and integers n and m, construct $Y_{n,i}$, $i = 1, 2, \dots, m$ as i.i.d. r.v. with conditional distribution $P^*(Y_{n,i} = X_j) = 1/n$ for $1 \leqslant j \leqslant n$. ($P^*$ denotes conditional distributon given $\boldsymbol{X}$.) Conditions relating the growth rate of m with n and the moments of X_1 are given to ensure the almost sure convergence of $(1/m)\sum_{i=1}^{m} Y_{n,i}$ to μ. This question is of some relevance in the theory of Bootstrap as developed by Efron (1979) and Bickel and Freedman (1981).

Keywords. Bootstrap, strong law.



Theorem 1. *If* $\underline{\lim}\, m\rho^{-n} > 0$ *for some* $\rho > 1$ *as* m, $n \to \infty$, *then*

$$\frac{1}{m}\sum_{i=1}^{m} Y_{n,i} \to \mu \text{ a.s.} \quad \textit{as } m, n \to \infty .$$

Theorem 2. *If* $\underline{\lim}\, mn^{-\beta} > 0$ *for some* $\beta > 0$ *as* m, $n \to \infty$ *and* $\mathbf{E}|X_1 - \mu|^\theta < \infty$ *for some* $\theta \geqslant 1$ *such that* $\theta\beta > 1$, *then*

$$\frac{1}{m}\sum_{1}^{m} Y_{n,i} \to \mu \text{ a.s.} \quad \textit{as } m, n \to \infty .$$

References.

[Ef] Efron Bradley. *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS-NSF Regional conference series in applied mathematics, 1982.

[RC] Cristian P. Robert and George Casella. Introducing Monte Carlo Methods with R. Series “Use R!”. Springer.

[CL] Chernick, M. R., anf LaBudde, R. A. (2014). An introduction to bootstrap methods with applications to R. John Wiley & Sons.