Chapter 3
The Classical Bootstrap

In Chap. 1, we briefly introduced the idea of bootstrapping. Now, together with the first applications, we will also give some theoretical results of the classical bootstrap approximation as first published simultaneously by Bickel and Freedman (1981) and Singh (1981). The methods of proof in these two papers are different and we follow mainly the work of Singh (1981) here. However, in Sect. 3.5, we will go into more detail about a proof concept applied in Bickel and Freedman (1981).

The first two sections of this chapter contain programming examples and the essential theorems for the classical bootstrap procedure. The last four sections give a deeper insight into the mathematical background. They are rather intended for readers who have a deeper knowledge of probability theory and mathematical statistics.

3.1 An Introductory Example

Recall from Chap. 1 the basic idea of the bootstrap. Starting with an i.i.d. sample

$$X_1, \ldots, X_n \sim F$$

with common unknown df. F we consider a statistic $T_n(F) = T_n(X_1, \ldots, X_n; F)$ whose df. we want to approximate. For the approximation, we use the df. of $T_n(\hat{F})$, where $\hat{F}$ is a known df. which is close to F.

In the situation of the *classical bootstrap* (cb.), the edf. F_n is used for $\hat{F}$. Hence

$$T_n(\hat{F}) = T_n(F_n) = T_n(X_1^*, \ldots, X_n^*; F_n),$$

where

$$X_1^*, \ldots, X_n^* \sim F_n \tag{3.1}$$

is an i.i.d. sample with common df. F_n. We call $X_1^*, \ldots, X_n^*$ the *bootstrap sample*. The underlying probability measure will be denoted here by $\mathbb{P}_n^* \equiv \mathbb{P}_{F_n}$. Note that the probability measure of the bootstrap distribution, $\mathbb{P}_n^*$, depends on the original observations $X_1, \ldots, X_n$, thus it is random! Furthermore, it changes from n to $n+1$. Notice, in (3.1), we notationally suppress the fact that the bootstrap sample changes its distribution with n. Hence, in an asymptotics setting, i.e., $n \to \infty$, it would be more precise to write

$$X_{1,n}^*, \ldots, X_{n,n}^* \sim F_n. \tag{3.2}$$

Nevertheless, for notational convenience we simply write $X_1^*, \ldots, X_n^*$ for the triangular scheme (3.2).

In the following set of examples, we will describe how the classical bootstrap can be used to construct a confidence interval for the expectation of an rv. Note that this is just an introductory example.

Example 3.1 Confidence interval for the expectation μ, part 1. Recall the situation of Sect. 1.1 and assume that we want to construct a confidence interval for the expectation $\mu = \mathbb{E}(X)$ of an rv. $X \sim F$ whose variance $\mathrm{VAR}(X) = \sigma^2 < \infty$ is unknown to us but expected to be finite. We observe an i.i.d. sample $X_1, \ldots, X_n$ and use the CLT to construct a 90% asymptotic confidence interval for μ. Based on Eq. (1.3), we get

$$\mathbb{P}\big(\Phi^{-1}(0.05) \leq \sqrt{n}(\bar{X}_n - \mu)/s_n \leq \Phi^{-1}(0.95)\big) \approx 0.9.$$

Here Φ^{-1} denotes the quantile function of the $\mathcal{N}(0, 1)$ distribution. Since $\Phi^{-1}(0.05)$ is equal to $-\Phi^{-1}(0.95)$, the confidence interval can be obtained from the result above. After some algebraic rearrangements, we get

$$\mathbb{P}\left(\mu \in \big[\bar{X}_n - s_n \times \Phi^{-1}(0.95)\big/\sqrt{n}\,,\ \bar{X}_n + s_n \times \Phi^{-1}(0.95)\big/\sqrt{n}\big]\right) \approx 0.9.$$

In this classical construction, the quantiles of the approximating normal distribution Φ are taken to approximate the corresponding quantiles of $\mathbb{P}_F\big(\sqrt{n}(\bar{X}_n - \mu)/s_n \leq x\big)$. It is Eq. (1.3) which allows this construction.

Now assume that the following approximation is a.s. correct:

$$\sup_{x \in \mathbb{R}} \left|\mathbb{P}\big(\sqrt{n}(\bar{X}_n - \mu)\big/s_n \leq x\big) - \mathbb{P}_n^*\big(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)\big/s_n^* \leq x\big)\right| \longrightarrow 0, \quad \text{as } n \to \infty, \tag{3.3}$$

where

$$\bar{X}_n^* := \frac{1}{n} \sum_{i=1}^n X_i^*, \qquad s_n^{*2} := \frac{1}{n-1} \sum_{i=1}^n \big(X_i^* - \bar{X}_n^*\big)^2.$$

As in the construction above, we can use $q_{0.05}$ and $q_{0.95}$ the 0.05 and 0.95 quantile of the approximating df. of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/s_n^*$ (with respect to the probability measure $\mathbb{P}_n^*$), respectively, to get

$$\mathbb{P}\big(q_{0.05} \leq n^{1/2}(\bar{X}_n - \mu)/s_n \leq q_{0.95}\big) \approx 0.9. \tag{3.4}$$

With some minor algebraic rearrangements, we finally derive

$$\big[\bar{X}_n - s_n \times q_{0.95}/\sqrt{n}\,,\ \bar{X}_n - s_n \times q_{0.05}/\sqrt{n}\big], \tag{3.5}$$

the bootstrap confidence interval for μ.

   But we still have to determine the two quantiles $q_{0.05}$ and $q_{0.95}$. In principle, it is possible to calculate these quantiles since we know the underlying distribution. With respect to the computing time involved, this will be impossible in most cases. However, since we know the underlying df. we can now use a *Monte Carlo* approach (mc.) to get at least an acceptable approximation for these quantiles. To see how this works in practice, we continue with Example 3.1.

Example 3.2   Confidence interval for the expectation μ, part 2. We start with a resampling scheme for the bootstrap data:

Resampling scheme 3.3   Classical bootstrap.

(A) $X_1, \ldots, X_n$ *observed data.*
(B)   *Calculate $q_{0.05}$ and $q_{0.95}$ the 0.05 and 0.95 quantile of $\mathbb{P}_n^*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/s_n^* \leq x)$, where $X_1^*, \ldots, X_n^*$ are i.i.d. according to F_n, the edf. of the observed data.*
(C)   *Take $[\bar{X}_n - s_n \times q_{0.95}/\sqrt{n}, \bar{X}_n - s_n \times q_{0.05}/\sqrt{n}]$ as a confidence interval.*

To apply a Monte Carlo approximation for step (B), one can use the following basic approach:

(B1)   Generate m bootstrap datasets $X_{\ell;1}^*, \ldots, X_{\ell;n}^* \sim F_n$, $1 \leq \ell \leq m$ and calculate $T_{\ell;n} := \sqrt{n}(\bar{X}_{\ell;n}^* - \bar{X}_n)/s_{\ell;n}^*$.
(B2)   Take $T_{[0.05 \times m]:m;n}$ and $T_{[0.95 \times m]:m;n}$ as an approximation for $q_{0.05}$ and $q_{0.95}$ in the interval under (C), where $(T_{\ell:m;n})_{1 \leq \ell \leq m}$ are the ordered $(T_{\ell;n})_{1 \leq \ell \leq m}$, that is, $T_{1;n} \leq T_{2;n} \leq \ldots \leq T_{m;n}$.

R-Example 3.4   Confidence interval for the expectation μ, part 3. The following R-code shows how this Monte Carlo approximation is applied under R to find the quantiles. Note that the unbiased estimates

$$s_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2, \quad s_n^{*2} := \frac{1}{n-1} \sum_{i=1}^{n} \big(X_i^* - \bar{X}_n^*\big)^2$$

are used in the R-code. Further, the sample quantiles obtained from the R-function "quantile" differ slightly from the sample quantiles defined in step (B2).

```r
step.B1 <- function(x, m = 1000){
  # x - observed data
  # m - number of MC replications

  # Realize step (B1)
  # generates m classical bootstrap data sets
  # calculates the standardized statistic and returns
  # them as a vector

  n <- length(x)
  mean.observed.x <- mean(x)

  # studentize x.boot according to step (B1)
  studentize <- function(x.boot){
    sqrt(n) * (mean(x.boot) - mean.observed.x) / sd(x.boot)
  }

  # step (B1)
  replicate(m, {
    x.boot <- sample(x, n, replace=TRUE)
    studentize(x.boot)
  })
}

ci <- function(x, conf.level){
  # x - observed data
  # conf.level - confidence level
  # returns left and right bound of the confidence interval
  # as a vector

  alpha  <- 1 - conf.level
  t.boot <- step.B1(x, m=999)

  # quantiles based on the MC simulation, see step (B) and (B2)
  ql.boot <- quantile(t.boot, alpha / 2)
  qr.boot <- quantile(t.boot, 1 - alpha / 2)

  n <- length(x)
  mean.observed.x <- mean(x)
  sd.observed.x   <- sd(x)

  # calculations according to step (C)
  left  <- mean.observed.x - qr.boot * sd.observed.x / sqrt(n)
  right <- mean.observed.x - ql.boot * sd.observed.x / sqrt(n)

  ret <- c(left, right)
  names(ret) <- c("lower", "upper")
  ret
}

set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
#x is a vector with the observed data
x <- rnorm(20, mean = 5, sd = 2)
```

```
ci(x = x, conf.level = 0.90)
```

```
##    lower    upper
## 4.583343 6.052707
```

A more convenient way to obtain the confidence interval is to apply the function
"boot.ci" from the `boot` package.

```r
bootstrap.ci <- function(x, conf.level, R){
  # x - observed data
  # conf.level - confidence level
  # R number of MC simulations

  # calculate mean and variance for the bootstrapped sample
  mean_and_var <- function(d, i){
    # d - observed data
    # i - boot::boot tells us which indices to be used to
    #     obtain the resampling version of the observed data

    d_boot <- d[i]
    c(mean_boot <- mean(d_boot), variance_boot = var(d_boot))
  }

  # resampled mean and variance
  b <- boot::boot(x, mean_and_var, R = R)

  # type = "stud" stands for studentized statistics
  ret <- boot::boot.ci(b, conf = conf.level, type = "stud",
                    # var.t0 - variance of the observed data
                    var.t0 = var(x),
                    # var.t - variances for every
                    #         resampled data sets
                    var.t = b$t[,2])$student[4:5]

  names(ret) <- c("lower", "upper")
  ret
}
```

```r
set.seed(123,kind ="Mersenne-Twister",normal.kind ="Inversion")
bootstrap.ci(rnorm(20, mean=5, sd=2), conf.level=0.9, R=999)
```

```
##    lower    upper
## 4.510934 6.097640
```

The slight differences between the two calculated confidence intervals originate
from two facts. First, the resampled data are generated differently; cf. the parameter
"simple" in the help page of the function "boot". Second, the quantiles are calculated
differently, cf. Davison and Hinkley (1997, p. 195).

Table 3.1 Observed coverage and mean interval length of 80% and 90% confidence intervals, based on resampling scheme 3.3 and normal approximation. The underlying distribution functions of the random samples (*n*=10) are *Exp*(0.1) and *UNI*(0,6)

fun	conf.level	proc	obsCoverage	meanIntervalWidth
rexp	0.8	bootstrap.ci	0.80	10.51
rexp	0.8	normal.ci	0.71	7.35
rexp	0.9	bootstrap.ci	0.90	14.36
rexp	0.9	normal.ci	0.84	9.76
runif	0.8	bootstrap.ci	0.85	1.51
runif	0.8	normal.ci	0.75	1.38
runif	0.9	bootstrap.ci	0.94	2.08
runif	0.9	normal.ci	0.86	1.77

Example 3.5 Confidence interval for the expectation μ, part 4. According to this classical bootstrap approach, we constructed confidence intervals and compared them with the corresponding ones constructed by approximation with the normal distribution. The result of this simulation study is given below in Table 3.1. For the distribution function, we choose the uniform df. on the interval [0, 6] and the exponential df. with parameter 0.1.

```r
normal.ci <- function(x, conf.level){
  # x - observed data
  # conf.level - confidence level

  # calculate confidence interval based on the
  # central limit theorem

  mean_observed_x <- mean(x)
  sd_observed_x   <- sd(x)
  n <- length(x)
  q <- qnorm((1-conf.level)/2)

  c(lower = mean_observed_x + q * sd_observed_x / sqrt(n),
    upper = mean_observed_x - q * sd_observed_x / sqrt(n))
}

# data will be generated using the uniform distribution
# and the exponential distribution
dg <- bind_rows(
  simTool::expand_tibble(fun = "runif", n = 10, max = 6),
  simTool::expand_tibble(fun = "rexp",  n = 10, rate = 0.1)) %>%
  as.data.frame

# 80% and 90% confidence intervals will be calculated by using
# the function boostrap.ci() and normal.ci()
pg <- bind_rows(
  simTool::expand_tibble(proc = "bootstrap.ci",
                    conf.level = c(0.8, 0.9), R = 999),
  simTool::expand_tibble(proc = "normal.ci",
```

```
                          conf.level = c(0.8, 0.9))) %>%
  as.data.frame
```

```
# create data sets according to data.frame dg
# calculate confidence intervals according to data.frame pg
# do this 1000 times on 4 cpus on the local machine
eg <- simTool::eval_tibbles(dg, pg, replications = 1000,
                            ncpus = 4,
                            cluster_seed = rep(123456, 6),
                            discard_generated_data = TRUE,
                            simplify = FALSE,
                            # convert the resulting vectors
                            # to a tibble
                            post_analyze = tibble::enframe)

results <- eg$simulation %>%
  # put the lower and upper limit
  # into two columns
  tidyr::unnest(results) %>%
  tidyr::spread(key = name, value = value) %>%
  # calculate the width of every constructed
  # confidence interval and determine if it
  # covers the true mean
  dplyr::mutate(width = upper - lower,
         mu = ifelse(fun == "rexp", 10, 3),
         muCovered = lower <= mu & mu <= upper) %>%
  dplyr::group_by(fun, conf.level, proc) %>%
  # calculate the covarage and mean length of the
  # confidence intervals conditioned on the different
  # distribution function, the confidence level of
  # the interval and the function used to calculate
  # the confidence interval
  dplyr::summarize(obsCoverage = mean(muCovered),
                   meanIntervalWidth = mean(width))
```

3.2 Basic Mathematical Background of the Classical Bootstrap

In this section, we give some mathematical justifications for the validity of the classical bootstrap approximations. We start with an example to show that the bootstrap approximation is not always correct!

Example 3.6 Let $X_1, \ldots, X_n \sim UNI$ be an i.i.d. sample with $UNI \equiv F$ as common df. The right-hand point of the support of F is obviously $1 (= T(F))$. To "estimate" $T(F)$, we take the largest observation $T_n(F) \equiv T_n(X_1, \ldots, X_n; F) = X_{n:n}$, where

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$$

denotes the order statistic corresponding to the observations. As we will see in Exercise 3.26,

$$\mathbb{P}_F(n(T(F) - T_n(F)) \le x) = \mathbb{P}_F(n(1 - X_{n:n}) \le x)$$
$$\longrightarrow 1 - \exp(-x), \quad \text{as } n \to \infty,$$

for all $x \ge 0$. In particular, we get for $x = 0$ that

$$\mathbb{P}_F(n(T(F) - T_n(F)) \le 0) = \mathbb{P}_F(n(1 - X_{n:n}) \le 0) \longrightarrow 0.$$

Now we mimic this situation for the bootstrap approximation. Having observed the sample $X_1, \ldots, X_n$, the right-hand point of the support of F_n is obviously the largest observation, thus $T(F_n) \equiv X_{n:n}$. To "estimate" $T(F_n)$ from the bootstrap sample $X_1^*, \ldots, X_n^*$ we have to take the largest bootstrap observation, thus $T_n(F_n) = X_{n:n}^*$. But now we get (see Exercise 3.26)

$$\mathbb{P}_n^*(n(T(F_n) - T_n(F_n)) \le 0) = \mathbb{P}_n^*(n(X_{n:n} - X_{n:n}^*) \le 0)$$
$$\longrightarrow 1 - \exp(-1), \quad \text{as } n \to \infty.$$

This shows that the bootstrap approximation is not correct here.   □

This disillusioning example points out clearly that we cannot expect a bootstrap approximation to be always possible. Furthermore, it tells us that we have to prove its correctness before we are allowed to use it.

In the following considerations, the sample size and the resampling size will be n, and the bootstrap sample will be taken from F_n the edf. corresponding to the i.i.d. sample $X_1, \ldots, X_n \sim F$. The bootstrap sample will be denoted as usual by $X_1^*, \ldots, X_n^* \sim F_n$ and we skip the second index n here. Furthermore, we use $\mathbb{E}(X) = \mu$, $\text{VAR}(X) = \sigma^2$,

$$\mathbb{E}_n^*(X^*) = \int x\, F_n(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} = \bar{X}_n, \quad \text{VAR}_n^*(X^*) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = s_n^2,$$

and finally

$$\bar{X}_n^* = \frac{1}{n} \sum_{i=1}^{n} X_i^*, \quad s_n^{*2} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i^* - \bar{X}_n^*\right)^2.$$

Note that now $1/n$ is used for s_n^2 and s_n^{*2} instead of $1/(n-1)$. In asymptotic considerations, this is irrelevant. With this definition s_n^2 becomes the variance of the bootstrap variable. This has, as will be seen later, advantages in theoretical considerations.

The Weak Law of Large Number (WLLN) guarantees that

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}h(X_i) - \int h(x)\,F(dx)\Big| > \varepsilon\Big) \longrightarrow 0, \quad \text{for every } \varepsilon > 0,$$

whenever the integral is defined. As to the bootstrap sample, we show

Theorem 3.7   Weak Law of Large Numbers for the classical bootstrap. *Assume that $\int |h(x)|\,F(dx) < \infty$. Then with probability one (w.p.1):*

$$\mathbb{P}_n^*\Big(\Big|n^{-1}\sum_{i=1}^{n}h(X_i^*) - \int h(x)\,F(dx)\Big| > \varepsilon\Big) \longrightarrow 0, \quad as\ n \to \infty,$$

for every $\varepsilon > 0$.

Proof Assume w.l.o.g. that the $F-$integral of h vanishes, otherwise we consider $h - \int h\,dF$. The bootstrap variables form a triangular array of independent rvs. within each row with common df. F_n. Define for $n \in \mathbb{N}$

$$h_n(x) = h(x)\,\mathrm{I}_{\{|h(x)|<n\}}$$

to get

$$\mathbb{P}_n^*\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}h(X_i^*)\Big| > \varepsilon\Big) \leq \mathbb{P}_n^*\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}h_n(X_i^*)\Big| > \varepsilon\Big) + \mathbb{P}_n^*\Big(\bigcup_{i=1}^{n}\{|h(X_i^*)| \geq n\}\Big).$$

The second probability on the right-hand side is bounded by

$$n\,\mathbb{P}_n^*\big(|h(X_1^*)| \geq n\big)$$

and the first by

$$\mathbb{P}_n^*\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}h_n(X_i^*) - \mathbb{E}_n^*(h_n(X_1^*))\Big| > \varepsilon/2\Big) + \mathrm{I}_{\{|\mathbb{E}_n^*(h_n(X_1^*))| > \varepsilon/2\}}.$$

Apply Chebyshev's inequality to get

$$\mathbb{P}_n^*\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}h_n(X_i^*) - \mathbb{E}_n^*(h_n(X_1^*))\Big| > \varepsilon/2\Big) \leq \frac{4}{n\varepsilon^2}\mathrm{VAR}_n^*(h_n(X_1^*)).$$

Thus, the proof is complete if we can show that

$$\lim_{n \to \infty} n\mathbb{P}_n^*\big(|h(X_1^*)| \geq n\big) = 0 \quad \mathbb{P} - a.s. \tag{3.6}$$

$$\lim_{n \to \infty} \mathbb{E}_n^*(h_n(X_1^*)) = 0 \quad \mathbb{P} - a.s. \tag{3.7}$$

$$\lim_{n \to \infty} \frac{1}{n} \mathrm{VAR}_n^*(h_n(X_1^*)) = 0 \quad \mathbb{P} - a.s. \tag{3.8}$$

hold. Now, apply Markov's inequality to get

$$n\mathbb{P}_n^*\big(|h(X_1^*)| \geq n\big) = n\mathbb{P}_{F_n}\{x : |h(x)| \geq n\} \leq \int\limits_{\{x : |h(x)| \geq n\}} |h(x)| \, F_n(\mathrm{d}x).$$

Fix for a moment $K \geq 0$ as a constant integer and apply the SLLN and the last inequality to get w.p.1

$$\limsup_{n \to \infty} n\mathbb{P}_n^*\big(|h(X_1^*)| \geq n\big) \leq \limsup_{n \to \infty} \int\limits_{\{x : |h(x)| \geq K\}} |h(x)| \, F_n(\mathrm{d}x) = \int\limits_{\{x : |h(x)| \geq K\}} |h(x)| \, F(\mathrm{d}x).$$

But the right-hand side can be made arbitrarily small by letting $K \uparrow \infty$ since $\int |h(x)| \, F(\mathrm{d}x) < \infty$ by assumption. Thus (3.6) holds.
To verify (3.7), we first observe that w.p.1

$$\lim_{n \to \infty} \int h(x) \, F_n(\mathrm{d}x) = \int h(x) \, F(\mathrm{d}x) = 0$$

according to SLLN. Combine this result with

$$\limsup_{n \to \infty} \int\limits_{\{x : |h(x)| \geq n\}} |h(x)| \, F_n(\mathrm{d}x) \leq \int\limits_{\{x : |h(x)| \geq K\}} |h(x)| \, F(\mathrm{d}x)$$

and use the same argument as in the proof of (3.6) to show that (3.7) holds. According to (3.7) it remains to show that w.p.1.

$$\frac{1}{n} \mathbb{E}_n^*\big(h_n^2(X_1^*)\big) \longrightarrow 0.$$

Note that

$$\frac{1}{n} \mathbb{E}_n^*\big(h_n^2(X_1^*)\big) \leq \frac{1}{n} \sum_{k=1}^{n} k^2 \, \mathbb{P}_n^*\big(k - 1 \leq |h(X_1^*)| < k\big)$$

$$\leq \frac{2}{n} \sum_{k=1}^{n} \sum_{j=1}^{k} j \, \mathbb{P}_n^*\big(k - 1 \leq |h(X_1^*)| < k\big)$$

$$\leq \frac{2}{n} \sum_{j=1}^{n} j \, \mathbb{P}_n^* \big(j - 1 \leq \big|h(X_1^*)\big| < n\big)$$

$$\leq 2 \sup_{j \in \mathbb{N}} \big|x_{nj} - x_j\big| + \frac{2}{n} \sum_{j=1}^{n} j \, \mathbb{P}\big(j - 1 \leq \big|h(X_1)\big|\big),$$

where $x_{nj} = j \, \mathbb{P}_n^* \big(j - 1 \leq \big|h(X_1^*)\big|\big)$ and $x_j = j \, \mathbb{P}\big(j - 1 \leq \big|h(X_1)\big|\big)$ for $n, j \in \mathbb{N}$. But the last sum is a Cesaro average, cf. Billingsley (1995, A30), of a sequence which tends to 0 w.p.1 by virtue of $\int |h(x)| \, F(dx) < \infty$. Hence, it remains to show that $\sup_{j \in \mathbb{N}} \big|x_{nj} - x_j\big| = o(1)$ almost surely, as $n \to \infty$. Note that for every fixed $j_0 \in \mathbb{N}$, $\big|x_{nj_0} - x_{j_0}\big| = o(1)$ almost surely, as $n \to \infty$, according to the SLLN. If the uniform convergence would not hold, a subsequence $(j_n)_{n \in \mathbb{N}}$ must exist such that $\big|x_{nj_n} - x_{j_n}\big| \geq c$ for all $n \in \mathbb{N}$ and some $c > 0$. But this is impossible due to $\int |h(x)| \, F(dx) < \infty$ and (3.6). This finally completes the proof. □

The next theorem shows that the approximation given under (3.3) is correct. Our proof here is based on Singh (1981).

Theorem 3.8 Central limit theorem for the classical bootstrap. *Let* $\mathbb{E}(X^2) < \infty$ *and set* $\mu = \mathbb{E}(X)$. *Then w.p.1*

$$\sup_{x \in \mathbb{R}} \Big| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu) \leq x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n) \leq x\big) \Big| \longrightarrow 0, \quad as \; n \to \infty.$$

Proof By the CLT, the continuity of Φ, the standard normal df., we get from a classical argument, cf. Loève (1977, p. 21), that it suffices to prove

$$\mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)/s_n \leq x\big) \longrightarrow \Phi(x), \quad \text{as } n \to \infty, \text{ for each } x \in \mathbb{R},$$

w.p.1. For this, we have to check the validity of Lindeberg's condition, cf. Serfling (1980, 1.9.3):

$$s_n^{-2} \int\limits_{\{|X_1^* - \bar{X}_n| \geq \varepsilon n^{1/2} s_n\}} (X_1^* - \bar{X}_n)^2 \, d\mathbb{P}_n^* \longrightarrow 0, \quad \text{as } n \to \infty, \text{ for each } \varepsilon > 0,$$

where the left-hand term equals

$$s_n^{-2} n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \mathbb{I}_{\{|X_i - \bar{X}_n| \geq \varepsilon n^{1/2} s_n\}}. \tag{3.9}$$

Note that for all $\tilde{\varepsilon} > 0$

$$\sum_{i \geq 1} \mathbb{P}\left(\frac{|X_i|}{\sqrt{i}} > \tilde{\varepsilon}\right) = \sum_{i \geq 1} \int_{[i-1,i[} \mathbb{P}\left(\frac{X^2}{\tilde{\varepsilon}} > i\right) dx \leq \int_0^{\infty} \mathbb{P}\left(\frac{X^2}{\tilde{\varepsilon}} > x\right) dx = \frac{\mathbb{E}(X^2)}{\tilde{\varepsilon}} < \infty.$$

Therefore, according to the Borel-Cantelli Lemma:

$$\limsup_{i \to \infty} \frac{|X_i|}{\sqrt{i}} = 0, \qquad \text{w.p.1.}$$

Since $s_n \to \sigma$ and $\bar{X}_n \to \mu$ w.p.1 according to SLLN, the last result ensures w.p.1 that, for $n \equiv n(\omega)$ sufficiently large,

$$\left| X_i - \bar{X}_n \right| \geq \varepsilon n^{1/2} s_n$$

can only hold for finitely many i. Hence, the indicator function under (3.9) can only be 1 in finitely many cases. This completes the proof.   □

This result, however, is not exactly what we stated under (3.3).

Corollary 3.9 *Under the assumptions of Theorem 3.8, we get w.p.1*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu)\big/s_n \leq x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)\big/s_n^* \leq x\big) \right| \longrightarrow 0, \quad \text{as } n \to \infty.$$

Proof As we have discussed in the proof of Theorem 3.8, we have to show

$$\mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)/s_n^* \leq x\big) \longrightarrow \Phi(x), \quad \text{as } n \to \infty, \tag{3.10}$$

w.p.1, for each $x \in \mathbb{R}$, while we already know that

$$\mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)/s_n \leq x\big) \longrightarrow \Phi(x), \quad \text{as } n \to \infty,$$

w.p.1, for each $x \in \mathbb{R}$. Since

$$\mathbb{P}_n^*\big(\left| s_n\big/s_n^* - 1 \right| > \varepsilon \big) \longrightarrow 0, \quad \text{as } n \to \infty,$$

w.p.1, for every $\varepsilon > 0$, according to an application of Theorem 3.7, (3.10) follows from Slutsky's theorem, cf. Serfling (1980, Theorem 1.5.4).   □

3.3 Discussion of the Asymptotic Accuracy of the Classical Bootstrap

In this section, we review some of Singh (1981) the results on the classical bootstrap without any proof. We have already seen in Theorem 3.8 that the CLT holds for the standardized mean when the classical bootstrap is used. But this result does not tell us anything about the quality of the approximation.

Again, going through the proof of Theorem 3.8, we find it to be in line with the classical argumentation. The same is true for the next theorem which gives us the

rate of convergence. Note that in the classical situation an appropriate bound is given
by the Berry-Esséen theorem, cf. Loève (1977, p. 300):

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu) \le x\big) - \Phi(x/\sigma) \right| \le K \rho \, \sigma^{-3} n^{-1/2},$$

where K is a universal constant and $\rho = \mathbb{E}(|X - \mu|^3)$. Based on the Berry-Esséen
theorem and the Law of Iterated Logarithm (LIL), i.e.,

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^{n}(X_i - \mu)}{(2\sigma^2 n \log(\log(n)))^{1/2}} = 1, \qquad \text{w.p.1},$$

cf. Serfling (1980, 1.10 Theorem A), Singh proved the following result:

Theorem 3.10 *Let* $\mathbb{E}(X^4) < \infty$. *Then w.p.1*

$$\limsup_{n \to \infty} n^{1/2} (\log(\log(n)))^{-1/2} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu) \le x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n) \le x\big) \right|$$
$$= (2\sigma^2 \sqrt{2\pi e})^{-1} \big(2\text{VAR}((X - \mu)^2)\big)^{1/2}.$$

As we already mentioned in the introduction, the bootstrap is a vehicle to approx-
imate the df. of a given statistic. Theorem 3.8 shows that the classical bootstrap
approximation holds in the case of arithmetic mean. But the normal approximation
also holds due to CLT. In a particular situation, we have to decide which approxima-
tion is preferable. Therefore, we have to compare the order of convergence of these
two approximations. Theorem 3.10 says that w.p.1

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu) \le x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n) \le x\big) \right| = O\left(\left(\frac{\ln(\ln(n))}{n} \right)^{1/2} \right). \tag{3.11}$$

The Berry-Esséen theorem shows that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu)/\sigma \le x\big) - \Phi(x) \right| = O(n^{-1/2}). \tag{3.12}$$

But (3.11) and (3.12) are not comparable since for (3.12) we have to know the
variance σ^2 which is unknown in most situations and which is not used in (3.11).
If $\mathbb{E}(|X|^3) < \infty$, Singh (1981) showed by applying an Edgeworth expansion that
w.p.1

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu)/\sigma \le x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)/s_n \le x\big) \right| = o(n^{-1/2}), \tag{3.13}$$

which is better than the approximation under (3.12). Furthermore, Abramovitch and
Singh (1985) proved under the assumption $\mathbb{E}(X^6) < \infty$ that w.p.1

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu)/s_n \le x\big) - \mathbb{P}_n^*\big(n^{1/2}(\bar{X}_n^* - \bar{X}_n)/s_n^* \le x\big) \right| = o(n^{-1/2}), \quad (3.14)$$

where $s_n^{*2} = n^{-1} \sum_{i=1}^{n} (X_i^* - \bar{X}_n^*)^2$.

In summary, one might think that the classical bootstrap approximation is always better than the normal approximation since it incorporates the Edgeworth terms automatically. However, if, for instance, F, the underlying df. is symmetric around μ, we get

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\big(n^{1/2}(\bar{X}_n - \mu)/\sigma \le x\big) - \Phi(x) \right| = o(n^{-1/2}), \quad (3.15)$$

which shows the same order of convergence that we find under (3.14). Furthermore, (3.15) still holds if we replace σ by s_n, cf. Abramovitch and Singh (1985).

Remark 3.11 A detailed discussion of the bootstrap and its relation to Edgeworth expansions can be found in Hall (1992).

3.4 Empirical Process and the Classical Bootstrap

Assume throughout this section that $X_1, \ldots, X_n \sim F$ is an i.i.d. sample with common continuous df. F, and let

$$\alpha_n(x) := n^{1/2}(F_n(x) - F(x)) \quad (3.16)$$

be the *empirical process*. The classical invariance principle of this process says, cf. Billingsley (1968, Theorem 16.4), that

$$\alpha_n \underset{n \to \infty}{\longrightarrow} B^o(F), \qquad \text{in distribution}$$

in the Skorokhod topology, where $B^o(F)$ is a transformed Brownian bridge, i.e., a centered Gaussian process with covariance structure given by

$$\mathbb{E}\big(B^o(F)(s) \cdot B^o(F)(t)\big) = F(s)(1 - F(t)), \quad s \le t.$$

To analyze the distribution of this process, one often takes a special version of α_n given by

$$\bar{\alpha}_n(F(x)), \qquad x \in \mathbb{R},$$

where $\bar{\alpha}_n(u) = n^{1/2}(\bar{F}_n(u) - u)$ is the *uniform empirical process* based on an uniform sample $U_1, \ldots, U_n \sim UNI$. Note that

$$\alpha_n(x) = n^{1/2}\Big(n^{-1}\sum_{i=1}^{n} I_{\{]-\infty,x]\}}(X_i) - F(x)\Big)$$

$$= n^{1/2}\Big(n^{-1}\sum_{i=1}^{n} I_{\{]-\infty,x]\}}(F^{-1}(U_i)) - F(x)\Big)$$

$$= n^{1/2}\Big(n^{-1}\sum_{i=1}^{n} I_{\{]0,F(x)]\}}(U_i) - F(x)\Big)$$

$$= \bar{\alpha}_n(F(x)).$$

In the following, we will consider the empirical process built according to the classical bootstrap resampling scheme. Denote this process by

$$\alpha_n^*(x) := n^{1/2}(F_n^*(x) - F_n(x)).$$

Theorem 3.12 *Assume F to be continuous. Then w.p.1*

$$\alpha_n^* \xrightarrow[n\to\infty]{} B^o(F), \qquad \text{in distribution}$$

in the Skorokhod topology.

Proof Since we know from the classical invariance principle that $\bar{\alpha}_n(F)$ converges to this limit process, it is enough to find a version of the empirical bootstrap process which is close to $\bar{\alpha}_n(F)$. To be precise, take for α_n^* the version given by

$$\alpha_n^* = \bar{\alpha}_n(F_n),$$

where now the same sample $U_1, \ldots, U_n$ is used as for the process $\bar{\alpha}_n(F)$. For notational reason, we use $\bar{\mathbb{P}}$ for the probability measure corresponding to this uniform sample. Let $\varepsilon > 0$ be arbitrarily chosen. Then

$$\bar{\mathbb{P}}\Big(\sup_{x\in\mathbb{R}} |\bar{\alpha}_n(F_n(x)) - \bar{\alpha}_n(F(x))| \ge \varepsilon\Big) \le \bar{\mathbb{P}}(\bar{w}_n(\|F_n - F\|) > \varepsilon),$$

where

$$\bar{w}_n(\delta) := \sup_{|t-s|\le\delta} |\bar{\alpha}_n(t) - \bar{\alpha}_n(s)|$$

denotes the modulus of continuity. Since $\|F_n - F\| \to 0$ $\mathbb{P}$−a.s. we can apply a general result on the oscillation of the uniform empirical process given by Stute (1982, (0.3)), to obtain w.p.1

$$\bar{\mathbb{P}}(\bar{w}_n(\|F_n - F\|) > \varepsilon) \xrightarrow[n\to\infty]{} 0,$$

which finally proves the theorem.   □

Remark 3.13 The proof of this theorem can be found in Swanepoel (1986) and in Dikta (1987, Appendix). Further bootstrap versions of important processes are also discussed there.

3.5 Mathematical Framework of Mallow's Metric

To analyze the classical bootstrap of the mean, Bickel and Freedman (1981) used a different concept than Singh (1981). Parts of their analysis are based on the relation of the classical bootstrap approximation to Mallow's metric. In this section, we will outline their approach and start with an important minimization result given in Major (1978, Theorem 8.1).

Theorem 3.14 *Let F and G be distribution functions such that $\int |x| \, F(\mathrm{d}x)$ and $\int |x| \, G(\mathrm{d}x)$ are finite. Assume that H is a two-dimensional df. on $(\mathbb{R}^2, \mathscr{B}_2^*)$, the product space equipped with the Borel $\sigma-$algebra, with marginal df. F and G, respectively, and define $\mathscr{M}$ to be the collection of all those H's. Then for every convex function $f : \mathbb{R} \longrightarrow \mathbb{R}$*

$$\inf_{H \in \mathscr{M}} \int f(x - y) \, H(\mathrm{d}x, \mathrm{d}y) = \int_0^1 f(F^{-1}(u) - G^{-1}(u)) \, \mathrm{d}u. \qquad (3.17)$$

Proof By an application of the separation theorem for convex functions, compare Rockafellar (1997, Corollary 11.5.1), we can find appropriate constants c and d such that $f(x - y) \geq c(x - y) + d$. This shows that $\int f(x - y) \, H(\mathrm{d}x, \mathrm{d}y)$ might be infinite but is always defined for every $H \in \mathscr{M}$.

In the first step of the proof, we assume that F and G define discrete distributions with common support on $x_1 < x_2 < \ldots < x_n$. Since $\{\int f(x - y) \, H(\mathrm{d}x, \mathrm{d}y) : H \in \mathscr{M}\}$ is closed and bounded in $\mathbb{R}$, the infimum under (3.17) is attained for some H. Fix such a minimizer, denote it by $H \in \mathscr{M}$, and let $X \sim F$ and $Y \sim G$ be defined over some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with joint df. H. Denote $\mathbb{P}(X = x_i, Y = x_k)$ by $p_{i,k}$, for $1 \leq i, k \leq n$. Then we can assume that the following property (3.18) holds:

$$\min(p_{i,j}, p_{k,\ell}) = 0, \quad \text{for all } k < i \text{ and } j < \ell. \qquad (3.18)$$

To prove this property assume that it is not correct for this H. Then we can find some $k < i$ and $j < \ell$ such that

$$p = \min(p_{i,j}, p_{k,\ell}) > 0.$$

We now define a new distribution $\tilde{H}$ by

$$\tilde{p}_{i,\ell} = p_{i,\ell} + p, \quad \tilde{p}_{k,j} = p_{k,j} + p, \quad \tilde{p}_{i,j} = p_{i,j} - p, \quad \tilde{p}_{k,\ell} = p_{k,\ell} - p,$$
$$\tilde{p}_{s,t} = p_{s,t} \quad \text{otherwise.}$$

Note that the marginal distributions of $\tilde{H}$ are identical to those of H and that

$$x_k - x_\ell < x_k - x_j < x_i - x_j \quad \text{and} \quad x_k - x_\ell < x_i - x_\ell < x_i - x_j.$$

With $0 < \alpha < 1$ defined by

$$\alpha = \frac{(x_k - x_j) - (x_i - x_j)}{(x_k - x_\ell) - (x_i - x_j)}$$

we get

$$x_k - x_j = \alpha(x_k - x_\ell) + (1 - \alpha)(x_i - x_j),$$

and also, since

$$(1 - \alpha) = \frac{(x_i - x_\ell) - (x_i - x_j)}{(x_k - x_\ell) - (x_i - x_j)},$$

$$x_i - x_\ell = (1 - \alpha)(x_k - x_\ell) + \alpha(x_i - x_j).$$

Convexity of f now yields

$$f(x_k - x_j) + f(x_i - x_\ell) \leq \alpha f(x_k - x_\ell) + (1 - \alpha)f(x_i - x_j)$$
$$+ (1 - \alpha)f(x_k - x_\ell) + \alpha f(x_i - x_j)$$
$$= f(x_k - x_\ell) + f(x_i - x_j),$$

and, therefore,

$$\int f(x - y)\,\tilde{H}(dx, dy) - \int f(x - y)\,H(dx, dy)$$
$$= p\left(f(x_k - x_j) + f(x_i - x_\ell) - f(x_k - x_\ell) - f(x_i - x_j)\right) \leq 0.$$

Overall, this shows that switching from H to $\tilde{H}$ does not increase the integral but fulfills (3.18) for this particular choice of i, j, k, ℓ. Furthermore, if $\tilde{H}$ should not have the property (3.18), we can apply the same procedure as above and, after finitely many steps, we end up with a df. such that fulfills the required property (3.18), and that minimizes $\int f(x - y)\,H(dx, dy)$ over $\mathcal{M}$.

Define the matrix $P_n = (p_{i,j})_{1 \leq i, j \leq n}$. Then property (3.18) says that for every coefficient $p_{i,j} > 0$ of P_n all the other northeast and southwest coefficients have to be zero. One can easily check (by induction on n) that this property together with the given marginal distributions uniquely determines the matrix P_n and, therefore, the joint distribution of (X, Y). Now check that the joint distribution of

$(F^{-1}(U), G^{-1}(U))$ has the property (3.18) when U is uniformly distributed on the unit interval. Therefore, (3.17) is correct in the discrete case.

In the next step, we assume that F and G are concentrated on the interval $[-T, T]$. Since f as a convex function on $\mathbb{R}$ is continuous and F and G are concentrated on a bounded interval, we can assume without loss of generality that f is bounded and that the infimum on the left-hand side of (3.17) is finite. Thus, we can find for an arbitrary $\varepsilon > 0$ a df. $H_0 \in \mathscr{M}$ such that

$$\inf_{H \in \mathscr{M}} \int f(x - y)\, H(\mathrm{d}x, \mathrm{d}y) > \int f(x - y)\, H_0(\mathrm{d}x, \mathrm{d}y) - \varepsilon.$$

The Rosenblatt transformation 2.4 guarantees that two rvs. (X, Y) on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ exist with joint df. H_0. Define, for every $n \in \mathbb{N}$,

$$(X_n, Y_n) = \Big(\frac{[nX]}{n}, \frac{[nY]}{n} \Big),$$

where $[t]$ denotes the integer part of t, and use F_n and G_n to denote the df. of X_n and Y_n, respectively. Obviously, $(X_n, Y_n) \longrightarrow (X, Y)$ w.p.1 and F_n and G_n define discrete distributions. Since the w.p.1 convergence also implies the convergence in distribution, the proof of the elementary Skorokhod theorem, compare Billingsley (1995, Theorem 25.6), shows that $F_n^{-1}(u) \longrightarrow F^{-1}(u)$ and $G_n^{-1}(u) \longrightarrow G^{-1}(u)$. This convergence holds for all $0 < u < 1$ out of a set with Lebesgue measure 1. Now, apply Lebesgue's dominated convergence theorem to get with the first part of our proof

$$\int f(x - y)\, H_0(\mathrm{d}x, \mathrm{d}y) = \lim_{n \to \infty} \int f(X_n - Y_n)\, \mathrm{d}\mathbb{P}$$

$$\geq \liminf_{n \to \infty} \int_0^1 f\big(F_n^{-1}(u) - G_n^{-1}(u) \big)\, \mathrm{d}u$$

$$= \int_0^1 f\big(F^{-1}(u) - G^{-1}(u) \big)\, \mathrm{d}u.$$

Overall, this shows that

$$\inf_{H \in \mathscr{M}} \int f(x - y)\, H(\mathrm{d}x, \mathrm{d}y) > \int_0^1 f\big(F^{-1}(u) - G^{-1}(u) \big)\, \mathrm{d}u - \varepsilon,$$

which proves (3.17) for F and G concentrated on intervals of the type $[-T, T]$.

In the third step of the proof, we now can take arbitrary F and G. Without loss of generality, we assume that the infimum in (3.17) is finite. As in the second step, we can find, for an arbitrary $\varepsilon > 0$, a df. $H_0 \in \mathscr{M}$ and random variables (X, Y) on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with joint df. H_0 such that

$$\inf_{H \in \mathcal{M}} \int f(x - y) \, H(\mathrm{d}x, \mathrm{d}y) > \int f(x - y) \, H_0(\mathrm{d}x, \mathrm{d}y) - \varepsilon.$$

Now set $A_n = \{|X| \leq n\} \cap \{|Y| \leq n\}$ and define $X_n = X \cdot \mathrm{I}_{\{A_n\}}$ and $Y_n = Y \cdot \mathrm{I}_{\{A_n\}}$, for $n \in \mathbb{N}$, and note that w.p.1 $X_n \longrightarrow X$ and $Y_n \longrightarrow Y$, respectively. Furthermore,

$$\left| f(X_n - Y_n) \right| \leq \left| f(X - Y) \right| + \left| f(0) \right|$$

for every $n \in \mathbb{N}$, and the bound on the right-hand side is integrable with respect to the chosen probability. Thus, Lebesgue's dominated convergence theorem is applicable. With the same argumentation used in the second step, we finally can complete the proof of (3.17) for these arbitrary F and G.   □

If the convex function f is defined by $f(x) = |x|^p$, for $p \geq 1$, we get an important application of the last theorem which leads to the following definition.

Definition 3.15 For $p \geq 1$, denote with $\mathscr{F}_p$ the class of all df. F with $\int |x|^p \, F(dx) < \infty$. Let $F, G \in \mathscr{F}_p$. Then

$$d_p(F, G) := \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p \, \mathrm{d}u \right)^{1/p} \tag{3.19}$$

defines *Mallow's p-metric*. For notational convenience, we will also use $d_p(X, Y)$, where $X \sim F$ and $Y \sim G$ and the joint distribution of (X, Y) minimizes the L^p distance over $\mathscr{M}$.

Corollary 3.16 *If we put $X = F^{-1}(U)$ and $Y = G^{-1}(U)$, where $U \sim UNI$, then*

$$d_p(F, G) = \| X - Y \|_p,$$

where the basic probability space is the unit interval with the Lebesgue measure and $\| \cdot \|_p$ denotes the L^p−norm.

Remark 3.17 For any scalars a, b let $F_{a,b}$ be the df. of $aX + b$, where $X \sim F$. For $F, G \in \mathscr{F}_p$, we then get

$$d_p(F_{a,b}, G_{a,b}) = |a| d_p(F, G).$$

Proof Apply Theorem 3.14 to get

$$
\begin{aligned}
d_p(F_{a,b}, G_{a,b}) &= \inf_{X \sim F, \, Y \sim G} \| (aX + b) - (aY + b) \|_p \\
&= |a| \inf_{X \sim F, \, Y \sim G} \| X - Y \|_p \\
&= |a| d_p(F, G).
\end{aligned}
$$

□

As the next lemma shows, Mallow's metric is closely related to weak convergence, where we now use the term "weak convergence" instead of "convergence in distribution", since we are dealing here with dfs. and not with rvs.

Lemma 3.18 *Assume that $F_n, F \in \mathscr{F}_p$, where F_n denotes not necessarily an edf. Then the following criteria are equivalent:*

(i) $d_p(F_n, F) \longrightarrow 0$, as $n \to \infty$.
(ii) As $n \to \infty$, $F_n \longrightarrow F$ weakly and $\int |x|^p F_n(\mathrm{d}x) \longrightarrow \int |x|^p F(\mathrm{d}x)$.
(iii) $F_n \longrightarrow F$ weakly, as $n \to \infty$ and $\{|X_n|^p : n \geq 1\}$ is uniformly integrable, where $X_n \sim F_n$.
(iv) $\int \varphi \, \mathrm{d}F_n \longrightarrow \int \varphi \, \mathrm{d}F$, as $n \to \infty$ for all continuous φ such that $\varphi(x) = O(|x|^p)$ as $x \to \pm\infty$.

Proof (i)$\Rightarrow$(ii): According to the last corollary we can use the rv.

$$X_n = F_n^{-1}(U), \quad X = F^{-1}(U).$$

Then,

$$\left| \left(\int |x|^p F_n(\mathrm{d}x) \right)^{1/p} - \left(\int |x|^p F(\mathrm{d}x) \right)^{1/p} \right| = \left| \|X_n\|_p - \|X\|_p \right| \leq \|X_n - X\|_p$$
$$= d_p(F_n, F) \longrightarrow 0$$

which shows the convergence of the integrals under (ii). It also guarantees the L^p-convergence of X_n to X which implies the weak convergence. This completes the proof of (ii).

(ii)$\Rightarrow$(iii): We only have to show uniform integrability. For this, we fix a point $a > 0$ such that a and $-a$ are continuity points of F. Then, we get

$$\int\limits_{\{|x|>a\}} |x|^p F_n(\mathrm{d}x) = \int |x|^p F_n(\mathrm{d}x) - \int\limits_{\{|x|\leq a\}} |x|^p F_n(\mathrm{d}x) \equiv I_n.$$

Since $\pm a$ are continuity points of F, the assumed weak convergence of $F_n \to F$ together with a slight modification of Billingsley (1995, Theorem 29.1) guarantees that

$$\int\limits_{\{|x|\leq a\}} |x|^p F_n(\mathrm{d}x) \longrightarrow \int\limits_{\{|x|\leq a\}} |x|^p F(\mathrm{d}x).$$

This combined with the assumed convergence of the p-th moments yields

$$I_n \longrightarrow \int |x|^p F(\mathrm{d}x) - \int\limits_{\{|x|\leq a\}} |x|^p F(\mathrm{d}x) = \int\limits_{\{|x|>a\}} |x|^p F(\mathrm{d}x).$$

The integral on the right-hand side can be made arbitrarily small by increasing a, which proves the uniform integrability.

(iii)$\Rightarrow$(iv): Let φ be as under (iv). Again, we take a fixed such that $\pm a$ are continuity points of F to get from the weak convergence of $F_n \to F$ that

$$\int_{\{|x| \leq a\}} \varphi(x)\, F_n(\mathrm{d}x) \longrightarrow \int_{\{|x| \leq a\}} \varphi(x)\, F(\mathrm{d}x).$$

Since $\varphi = O(|x|^p)$ there exists a constant c such that

$$|\varphi(x)| \leq c|x|^p,$$

for all x such that $|x| \geq a$. Thus,

$$\int_{\{|x| > a\}} |\varphi(x)|\, F_n(\mathrm{d}x) \leq c \int_{\{|x| > a\}} |x|^p\, F_n(\mathrm{d}x).$$

Furthermore, the assumed uniform integrability implies that we can choose for every given $\varepsilon > 0$ a continuity point $a = a(\varepsilon)$ of F such that

$$\sup_{n \geq 1} c \int_{\{|x| > a\}} |x|^p\, F_n(\mathrm{d}x) \leq \varepsilon$$

which completes the proof of (iv).

(iv)$\Rightarrow$(i): Obviously, (iv) implies (ii). Therefore, it suffices to show (ii)$\Rightarrow$(i). But the weak convergence of F_n to F implies the almost sure convergence of $X_n = F_n^{-1}(U)$ to $X = F^{-1}(U)$ w.r.t. the Lebesgue measure on the unit interval ($U \sim UNI$), cf. the proof of Billingsley (1995, Theorem 25.6). Since $\mathbb{E}(|X_n|^p) \longrightarrow \mathbb{E}(|X|^p) < \infty$, as $n \to \infty$, according to (iv), we finally get from Loève (1977, L^p-Convergence Theorem) that $\|X_n - X\|_p \to 0$, as $n \to \infty$. This completes the proof of (i). $\quad\square$

In the special case that F_n is the edf. of an i.i.d. sample, we get the following corollary.

Corollary 3.19 *Assume $F \in \mathscr{F}_p$ and let F_n be the edf. Then, $d_p(F_n, F) \longrightarrow 0$ w.p.1.*

Proof The Glivenko-Cantelli theorem says that w.p.1

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0.$$

Thus, w.p.1, $F_n \to F$ weakly. Furthermore, from the SLLN, we conclude

$$\int |x|^p\, F_n(\mathrm{d}x) \longrightarrow \int |x|^p\, F(\mathrm{d}x)$$

w.p.1. Now, apply the last lemma to complete the proof.   □

In the following lemma, we bound Mallow's distance of two sums of independent rv. by Mallow's distance of the summands.

Lemma 3.20 *Assume that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are two sequences of independent rv. in $\mathscr{F}_p$. Then,*

$$d_p\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\Big) \le \sum_{i=1}^n d_p(X_i, Y_i).$$

Proof Take $U_1, \ldots, U_n \sim UNI$ as an i.i.d. sample and set

$$\tilde{X}_i := F_i^{-1}(U_i), \quad \tilde{Y}_i := G_i^{-1}(U_i)$$

for $i = 1, \ldots, n$, where $X_i \sim F_i$ and $Y_i \sim G_i$, for $1 \le i \le n$. According to the definition of $d_p(X, Y)$, we get for each $i = 1, \ldots, n$

$$d_p(X_i, Y_i) = d_p(\tilde{X}_i, \tilde{Y}_i) = \|\tilde{X}_i - \tilde{Y}_i\|_p.$$

Now, apply Corollary 3.16 and Minkowski's inequality to obtain

$$d_p\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\Big) \le \Big\|\sum_{i=1}^n \tilde{X}_i - \sum_{i=1}^n \tilde{Y}_i\Big\|_p \le \sum_{i=1}^n \|\tilde{X}_i - \tilde{Y}_i\|_p$$

$$= \sum_{i=1}^n d_p(\tilde{X}_i, \tilde{Y}_i) = \sum_{i=1}^n d_p(X_i, Y_i),$$

which finally proves the lemma.   □

If $p = 2$, the last lemma improves in the presence of equal means.

Lemma 3.21 *Assume in addition to the assumptions of Lemma 3.20 that $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$, for $1 \le i \le n$ and $p \ge 2$. Then,*

$$d_2^2\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\Big) \le \sum_{i=1}^n d_2^2(X_i, Y_i).$$

Proof Take $(\tilde{X}_i, \tilde{Y}_i)$ as in the proof of Lemma 3.20. From Corollary 3.16 and Bienaymé's equality, we get

$$d_2^2\Big(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\Big) \le \|\sum_{i=1}^n \tilde{X}_i - \tilde{Y}_i\|_2^2 = \sum_{i=1}^n \|\tilde{X}_i - \tilde{Y}_i\|_2^2 = \sum_{i=1}^n d_2^2(X_i, Y_i).$$

□

Corollary 3.22 *Let $F, G \in \mathscr{F}_p$ with $p \geq 2$. Assume that $X_1, \ldots, X_n$ are i.i.d. with common df. F and $Y_1, \ldots, Y_n$ are i.i.d. with common df. G, respectively. Furthermore, we assume that $\mathbb{E}(X_1) = \mathbb{E}(Y_1)$. Then,*

$$d_2^2\left(n^{-1/2} \sum_{i=1}^{n} X_i, n^{-1/2} \sum_{i=1}^{n} Y_i\right) \leq n^{-1} \sum_{i=1}^{n} d_2^2(X_i, Y_i) = d_2^2(F, G).$$

Remark 3.23 According to the CLT and Lemma 3.18 (ii), for standardized F and $G = \Phi$, the left-hand side of the inequality under Corollary 3.22 tends to zero as $n \to \infty$. Since the right-hand side of this inequality is fixed for each $n \geq 1$ and positive for $G \neq F$, this inequality cannot be used to prove the CLT. However, if G depends on n in such a way that $d_2^2(F, G_n) \to 0$, as $n \to \infty$, we can use this inequality to prove weak convergence. In the particular case of the classical bootstrap, $G_n = F_n$ is the edf. of the i.i.d. sample $X_1, \ldots, X_n$, while $Y_1, \ldots, Y_n$ forms a bootstrap sample, then $d_2^2(F, F_n) \to 0$ w.p.1. Together with the CLT, this proves the CLT for the standardized bootstrap sample under the classical resampling scheme.

Sometimes, however, the bootstrap sample comes from $\tilde{F}_n$, the edf. of a not necessarily independent sample $X_1^*, \ldots, X_n^*$. For example, in linear regression, we have the situation that the residuals $X_i^* = \tilde{\varepsilon}_i, 1 \leq i \leq n$, are not independent. In such a case, the approach outlined above for F_n cannot be applied for $\tilde{F}_n$ unless it is guaranteed that in some sense $\tilde{F}_n$ is close to F_n. A condition which will work in this setup is given in the next lemma; compare also Freedman (1981).

Lemma 3.24 *Assume that $X_1, \ldots, X_n$ are i.i.d. with df. F and edf. F_n. Let $X_1^*, \ldots, X_n^*$ be a second i.i.d. sample with edf. $\tilde{F}_n$ such that w.p.1*

$$\frac{1}{n} \sum_{i=1}^{n} |X_i^* - X_i|^p \longrightarrow 0, \quad as \ n \to \infty \qquad (3.20)$$

holds. If $F \in \mathscr{F}_p$ for some $p \geq 1$, then, w.p.1, $d_p(\tilde{F}_n, F) \longrightarrow 0$, as $n \to \infty$.

Proof Let $U \sim UNI$ be the uniform distribution on the unit interval. Since

$$
\begin{aligned}
d_p(\tilde{F}_n, F) &= \|\tilde{F}_n^{-1}(U) - F^{-1}(U)\|_p \\
&\leq \|\tilde{F}_n^{-1}(U) - F_n^{-1}(U)\|_p + \|F_n^{-1}(U) - F^{-1}(U)\|_p \\
&= d_p(\tilde{F}_n, F_n) + d_p(F_n, F)
\end{aligned}
$$

and, w.p.1, $d_p(F_n, F) \longrightarrow 0$, as $n \to \infty$, according to Corollary 3.19, we get from assumption (3.20) and Theorem 3.14

$$d_p(\tilde{F}_n, F_n) \leq \left(\frac{1}{n} \sum_{i=1}^{n} |X_i^* - X_i|^p\right)^{1/p} \longrightarrow 0, \quad as \ n \to \infty,$$

w.p.1. This completes the proof of the lemma. □

3.6 Exercises

Exercise 3.25 Repeat the simulation study from Example 3.5 without using the `simTool` package. Note in that simulation the functions "bootstrap.ci" and "normal.ci" are applied to any dataset that is generated.

Exercise 3.26 Recall the assumptions of Example 3.6 and show that for $x \geq 0$

$$\mathbb{P}(n(1 - X_{n:n}) \leq x) \longrightarrow 1 - \exp(-x), \quad \text{as } n \to \infty.$$

Furthermore, show for the bootstrap sample

$$\mathbb{P}_n^*(n(X_{n:n} - X_{n:n}^*) \leq 0) \longrightarrow 1 - \exp(-1), \quad \text{as } n \to \infty.$$

Exercise 3.27 Conduct a simulation that indicates

$$\mathbb{P}_n^*(n(X_{n:n} - X_{n:n}^*) \leq 0) \longrightarrow 1 - \exp(-1), \quad \text{as } n \to \infty,$$

see Exercise 3.26.

Exercise 3.28 Recall the scenario of Theorem 3.7 and assume in addition that

$$\int h^2(x) \, F(\mathrm{d}x) < \infty.$$

Use Chebyshev's inequality to verify the assertion of Theorem 3.7.

Exercise 3.29 Implement in R the simulation study of Example 3.5, without using the `simTool` package.

Exercise 3.30 Use R to generate $U_1, \ldots, U_{100}$ i.i.d. rvs. according to the uniform distribution. Based on this data,

(i) plot the path of the corresponding empirical process;
(ii) generate a classical bootstrap sample to the data and plot the path of the corresponding empirical process.

Exercise 3.31 In the R-library `boot`, many bootstrap applications are already implemented. Read the corresponding help to this library and try to redo the simulation under Exercise 3.29 by using the functions of this library.

Exercise 3.32 Verify that for fixed discrete marginal distributions on $x_1 < x_2 < \ldots < x_n$ the property (3.18) used in the proof of Theorem 3.14 defines exactly one joint distribution.

Exercise 3.33 Verify that for fixed discrete marginal distributions on $x_1 < x_2 < \ldots < x_n$ with df. F and G, respectively, the joint distribution of $(F^{-1}(U), G^{-1}(U))$ has the property (3.18) used in the proof of Theorem 3.14. Here U is uniformly distributed on the unit interval.

Exercise 3.34 Let $F \in \mathscr{F}_p$, that is, $\int |x|^p \, F(dx) < \infty$. Verify that

(i) for a location family, $F_\theta(x) = F(x - \theta)$

$$d_p(F_{\theta_1}, F_{\theta_2}) = |\theta_1 - \theta_2|;$$

(ii) for a scale family, $F_\sigma(x) = F(\sigma x)$ with $\sigma > 0$

$$d_p(F_{\sigma_1}, F_{\sigma_2}) = |\sigma_1 - \sigma_2| \left(\int |x|^p \, F(dx) \right)^{1/p}.$$

Exercise 3.35 Verify that for two (centered) binomial distributions F_1 and F_2 with parameters (n, p_1) and (n, p_2), respectively,

$$d_2^2(F_1, F_2) \le n |p_1 - p_2| (1 - |p_1 - p_2|).$$

Exercise 3.36 Verify that for two (centered) Poisson distributions F_1 and F_2 with parameters λ_1 and λ_2, respectively,

$$d_2^2(F_1, F_2) \le |\lambda_1 - \lambda_2|.$$

References

Abramovitch L, Singh K (1985) Edgeworth corrected pivotal statistics and the bootstrap. Ann Stat 13(1):116–132

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9(6):1196–1217

Billingsley P (1968) Convergence of probability measures. Wiley, New York

Billingsley P (1995) Probability and measure, 3rd edn. Wiley Series in Probability and Mathematical Statistics. Wiley, New York

Davison AC, Hinkley DV (1997) Bootstrap methods and their application, vol 1. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK

Dikta G (1987) Beiträge zur nichtparametrischen Schätzung der Regressionsfunktion. PhD thesis, Justus-Liebig-Universität Gießen

Freedman DA (1981) Bootstrapping regression models. Ann Stat 9(6):1218–1228

Hall P (1992) The bootstrap and Edgeworth expansion. Springer Series in Statistics. Springer, New York

Loève M (1977) Probability theory. I, 4th edn. Springer, New York

Major P (1978) On the invariance principle for sums of independent identically distributed random variables. J Multivar Anal 8(4):487–517

Rockafellar RT (1997) Convex analysis. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ

Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, New York

Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Stat 9(6):1187–1195

Stute W (1982) The oscillation behavior of empirical processes. Ann Probab 10(1):86–107

Swanepoel JWH (1986) A note on proving that the (modified) bootstrap works. Commun Stat Theory Methods 15(11):3193–3203