STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 4

Last Week

- Empirical distribution function
- Plug-in principle
- Estimated standard errors

Sampling from the empirical distribution $\hat{F}$

Suppose we want to draw an iid sample $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ from $\widehat{F}$, where $\widehat{F}$ puts mass $1/n$ on each observations $x_i,\ i = 1, 2, \ldots, n$.

Hence, when sampling from $\hat{F}$, the i^{th} observation x_i in the original sample is selected with probability $1/n$. This leads to the following two-step procedure:

- Draw $i_1, i_2, \ldots, i_n$ independently from the uniform distribution on $\{1, 2, \ldots, n\}$.
- Set $x_j^* = x_{i_j}$ and $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$.
- Thus, we might have $\mathbf{x}^* = (x_3, x_7, x_3, x_{31}, \ldots, x_5)$.

In other words, we sample with replacement from the original sample.

Problem

Aflatoxin residues in peanut butter: In actual testing, 12 lots of peanut butter had aflatoxin residues in parts per billion of

```
aflatoxin <- c(4.94, 5.06, 4.53, 5.07, 4.99, 5.16, 4.38,
               4.43, 4.93, 4.72, 4.92, 4.96)
```

1. Using R and the **sample()** function, or a random number generator, generate five resamples of the integers from 1 to 12.
2. For each of the resamples in a, find the mean of the corresponding elements of the aflatoxin data set. Print out the 5 bootstrap means.
3. Find the mean of the resample means. Compare this with the mean of the original data set.
4. Find the minimum and the maximum of the five resample means. This is a crude bootstrap confidence interval on the mean. (If you had used 1000 resamples, and used the 25th and 975th largest means, this would have given a reasonable 95% confidence interval.)

Chapter 6: The bootstrap estimate of standard error

The Bootstrap Principle

Suppose

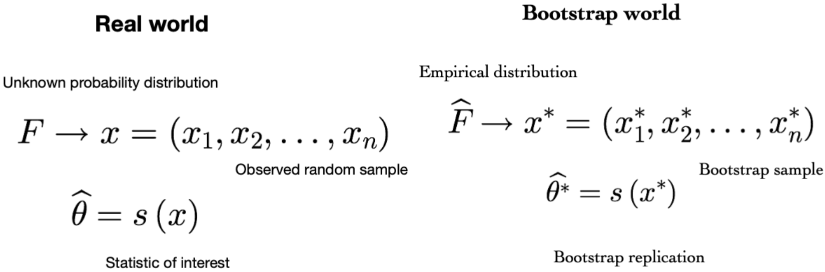
- $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is an observed random sample from an unknown probability distribution F.
- $\theta = t(F)$ is a parameter of interest on the basis of $\mathbf{x}$.
- $\hat{\theta} = s(\mathbf{x})$ is an estimate for θ.

For an evaluation of the statistical properties such as bias and standard error for the estimate $\hat{\theta}$, we wish to estimate the sampling distribution of $\hat{\theta}$.

The bootstrapping method mimics the data-generating process by sampling from an estimate $\hat{F}$ of the unknown distribution F. Thus the role of the above real quantities is taken by their analogous quantities in the “bootstrap world”:

- $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is a bootstrap from $\hat{F}$.
- $\theta^* = t(\hat{F})$ is the parameter in the bootstrap world.
- $\hat{\theta}^* = s(\mathbf{x}^*)$ is the bootstrap replication of θ.

The Bootstrap Principle



Real world

Unknown probability distribution

$$F \longrightarrow x = (x_1, x_2, \ldots, x_n)$$

Observed random sample

$$\widehat{\theta} = s\,(x)$$

Statistic of interest

Bootstrap world

Empirical distribution

$$\widehat{F} \longrightarrow x^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

Bootstrap sample

$$\widehat{\theta}^* = s\,(x^*)$$

Bootstrap replication

Monte Carlo Method

The bootstrap estimate of the sampling distribution of $\hat{\theta}$ is generally computed using Monte Carlo methods:

Example: Assume we know the population distribution, say a standard normal distribution $N(0,1)$, and the sample size $n = 100$. What will the distribution of the sample median be?

Answer: We can find this out using the Monte Carlo simulation approach. First we draw a random sample using **R** and compute the sample median,

```r
x <- rnorm(100)
(x_med <- median(x))
```

```
## [1] 0.1139092
```

This gives us one realization of the median. However, every time we apply the same program, we obtain a different value of the sample median because of a different sets of points we are having.

One way to investigate the distribution of sample median is to repeat the above procedure many times and keep track of the sample median of each sample we generate.

Monte Carlo Method

The sampling distribution of $\hat{\theta}$ using Monte Carlo Method

Step 1 Draw B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ from $\hat{F}$.

Step 2 Evaluate bootstrap replications,

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), \quad b = 1, 2, \ldots, B.$$

Step 3 Estimate the sampling distribution of $\hat{\theta}$ by the empirical distribution of the bootstrap replications $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$:

$$\widehat{\text{Prob}}\left(\hat{\theta} \in A\right) = \frac{1}{B} \sum_{b=1}^{B} I_{\left\{\hat{\theta}^*(b) \in A\right\}}$$

for appropriate subsets A.

Often we are only interested in one characteristic of the sampling distribution of $\hat{\theta}$, for example the **standard error** or the bias. Estimates for these quantities can be straightforwardly obtained from the bootstrap replications.

The Non-Parametric Bootstrap

The bootstrap algorithm for estimating the standard error of $\hat{\theta} = s(\mathbf{x})$.

Step 1 Construct an empirical probability distribution, $\hat{F}$, from the sample by placing a probability of $1/n$ at each point, $x_1, x_2, \ldots, x_n$ of the sample. This is the empirical distribution function of the sample, which is the nonparametric maximum likelihood estimate of the population distribution, F.

Step 2 From the empirical distribution function $\hat{F}$, draw a random sample of size n with replacement. This is a *resample*.

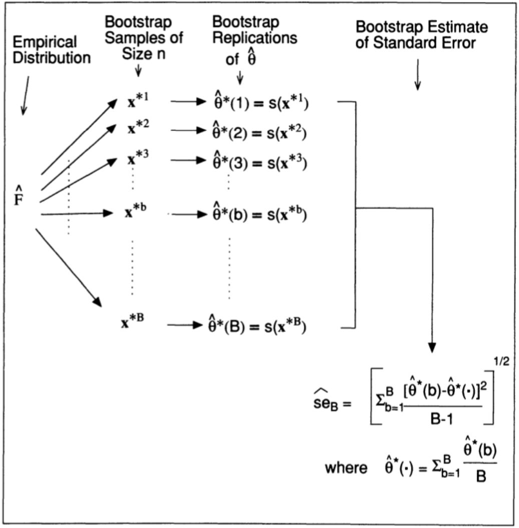
Step 3 Calculate the statistic of interest, $s(\mathbf{x})$, for this resample, yielding $s(\mathbf{x}^*)$.

Step 4 Repeat steps 2 and 3 B times, where B is a large number, in order to create B resamples. The practical size of B depends on the test to be run on the data. Typically, B is at least equal to 1000 when an estimate of confidence interval around $s(\mathbf{x})$ is required.

Step 5 Construct the relative frequency histogram from the B number of $s(\mathbf{x}^*)$'s by placing a probability of $1/B$ at each point, $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots, s(\mathbf{x}^{*B})$. The distribution obtained is the bootstrapped estimate of the sampling distribution of $s(\mathbf{x})$. This distribution can now be used to make inferences about the parameter θ, which is to be estimated by $s(\mathbf{x})$.

The Non-Parametric Bootstrap

The bootstrap algorithm for estimating the standard error of $\hat{\theta} = s(\mathbf{x})$.

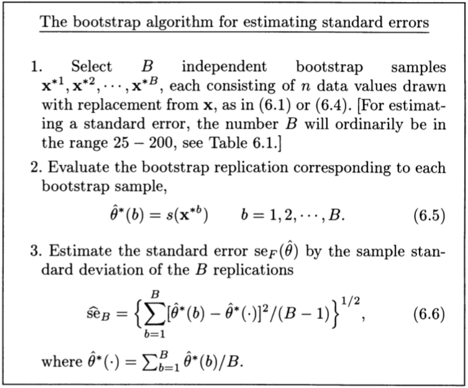

Empirical Distribution
Bootstrap Samples of Size n
Bootstrap Replications of $\hat{\theta}$
Bootstrap Estimate of Standard Error
$\hat{F}$
$\mathbf{x}^{*1}$
$\mathbf{x}^{*2}$
$\mathbf{x}^{*3}$
$\mathbf{x}^{*b}$
$\mathbf{x}^{*B}$
$\hat{\theta}^*(1) = s(\mathbf{x}^{*1})$
$\hat{\theta}^*(2) = s(\mathbf{x}^{*2})$
$\hat{\theta}^*(3) = s(\mathbf{x}^{*3})$
$\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$
$\hat{\theta}^*(B) = s(\mathbf{x}^{*B})$
$\widehat{se}_B = \left[\Sigma_{b=1}^{B} \frac{[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}\right]^{1/2}$
where $\hat{\theta}^*(\cdot) = \Sigma_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$

The Bootstrap Estimate of Standard Error

The bootstrap algorithm for estimating the standard error of $\hat{\theta} = s(\mathbf{x})$.

Let $\hat{\theta} = s(\mathbf{x})$ be an estimator for θ and suppose we want to know the standard error of $\hat{\theta}$. A bootstrap estimate of standard error can be obtained by the following algorithm:

Algorithm 6.1



The bootstrap algorithm for estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$, as in (6.1) or (6.4). [For estimating a standard error, the number B will ordinarily be in the range $25 - 200$, see Table 6.1.]
2. Evaluate the bootstrap replication corresponding to each bootstrap sample,

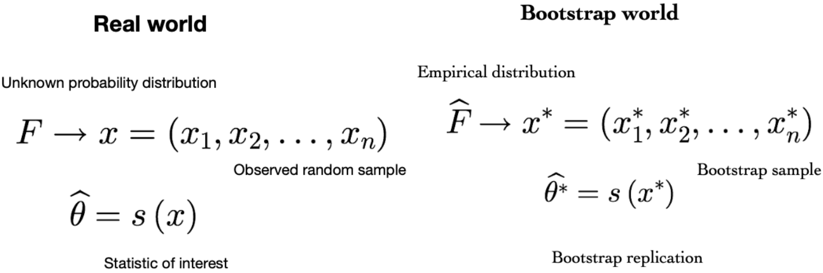
$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \qquad b = 1, 2, \cdots, B. \tag{6.5}$$

3. Estimate the standard error $\text{se}_F(\hat{\theta})$ by the sample standard deviation of the B replications

$$\widehat{\text{se}}_B = \left\{ \sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2/(B-1) \right\}^{1/2}, \tag{6.6}$$

where $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

The Bootstrap Principle



Real world

Unknown probability distribution

$$F \longrightarrow x = (x_1, x_2, \ldots, x_n)$$

Observed random sample

$$\widehat{\theta} = s(x)$$

Statistic of interest

Bootstrap world

Empirical distribution

$$\widehat{F} \longrightarrow x^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

Bootstrap sample

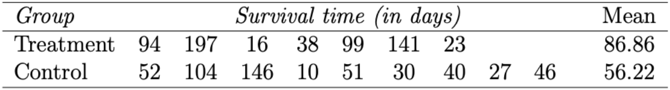
$$\widehat{\theta}^* = s(x^*)$$

Bootstrap replication

Example

Mouse Data

- A small randomized experiment were done with 16 mouse, 7 to treatment group and 9 to control group. Treatment was intended to prolong survival after a test surgery.



Group	*Survival time (in days)*									Mean
Treatment	94	197	16	38	99	141	23			86.86
Control	52	104	146	10	51	30	40	27	46	56.22

```
library(bootstrap)
(trt <- mouse.t)
```

```
## [1]  94 197  16  38  99 141  23
```

```
(ctl <- mouse.c)
```

```
## [1]  52 104 146  10  50  31  40  27  46
```

Example

Mouse Data

Suppose that we want to assess in the accuracy of the sample mean of the treatment group.

```
result <- bootstrap(mouse.t, 1000, theta = mean)
sd(result$thetastar)
```

```
## [1] 23.22694
```

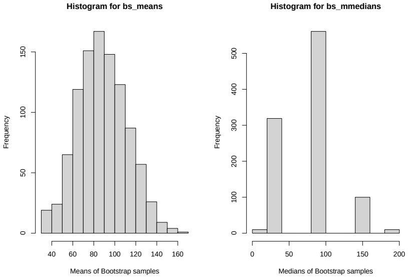
Note that this is a Monte Carlo approximation to the ideal bootstrap estimate, which in the special case of the sample mean is given by

$$\widehat{\mathsf{se}}(\bar{x}) = \left\{ \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{1/2} = 23.36.$$

The two estimates agree quite well.

Example: Mouse Data

Mean vs Median



Th main aspect of the bootstrap distribution of the median is that it can take on very few values, in the case of the treatment group for instance, 7. The simple bootstrap will always present this discrete characteristic even if we know the underlying distribution is continuous, there are ways to fix this and in many cases it won't matter but it is an important feature

Trimmed Mean

A remedy: Trimmed Mean

- The 25% trimmed mean (when the lowest 25% and the highest 25% are discarded) is known as the interquartile mean.
- The median is the mean of the 1 or 2 middle observations.
- The trimmed mean often does a better job of representing the average of typical observations than does the median. *Bootstrapping trimmed means* also works better than bootstrapping medians, because the bootstrap doesn't work well for statistics that depend on only 1 or 2 observations.

The combinatorics of bootstrap samples

Bootstrap samples

- There is a total of $\binom{2n-1}{n}$ distinct bootstrap samples.
- The probability of obtaining one of these samples under sampling with replacement can be obtained from the multinomial distribution, i.e. we are drawing bootstrap samples from the multinomial distribution, a vector $(k_1, k_2, \ldots, k_n)$ with each of $n(= k_1 + k_2 + \cdots + k_n)$ categories being equally likely, $p_i = 1/n$, so that the probability of a possible vector is

$$\text{Prob}_{boot}(k_1, k_2, ..., k_n) = \binom{n}{k_1, k_2, \ldots, k_n} \left(\frac{1}{n}\right)^n .$$

- For small n, we can find the ideal bootstrap estimate of standard error(SE) using all possible samples, but as n gets larger, the computation of SE gets impractical.

The number of bootstrap replications B

Two rules of Thumb in the textbook

- Even a small number of bootstrap replications, say $B = 25$, is usually informative. $B = 50$ is often enough to give a good estimate of $\text{se}_F(\hat{\theta})$.
- Very seldom are more than $B = 200$ replications needed for estimating a standard error. (Much bigger values of B are required for bootstrap confidence intervals).

The Parametric Bootstrap

When we assume the data is from a parametric model, we can use the parametric bootstrap to access the uncertainty (variance, standard deviation, confidence intervals) of the estimated parameter.

Example: Let $X_1, X_2, \ldots, X_n \sim N(0, \sigma^2)$, where σ^2 is unknown number. A natural way to estimate σ^2 is via the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Because the sample variance is an estimator, it is a random quantity. How do we estimate the variance of the sample variance?

The Parametric Bootstrap

Example (Continued..) Since we know that the sample variance is a good estimator of σ^2, we can use it to replace σ^2, leading to a new distribution $N(0, S_n^2)$ We know to sample from this new distribution, so we just generate bootstrap samples from this distribution. Assume we generate B sets of samples:

$$\begin{gathered} X_1^{*(1)}, \cdots, X_n^{*(1)} \sim N(0, S_n^2) \\ X_1^{*(2)}, \cdots, X_n^{*(2)} \sim N(0, S_n^2) \\ \vdots \\ X_1^{*(B)}, \cdots, X_n^{*(B)} \sim N(0, S_n^2). \end{gathered}$$

Then, to estimate the variability of S_n^2,

$$\widehat{\mathsf{Var}}_B(S_n^2) = \frac{1}{B-1} \sum_{b=1}^{B} \left(S_n^{2*}(b) - \bar{S}_n^{2*} \right)^2,$$

where $S_n^{2*}(b)$, $b = 1, 2, \ldots, B$ are the sample variance of each bootstrap sample, and $\bar{S}_n^{2*} = \frac{1}{B} \sum\limits_{b=1}^{B} S_n^{2*}(b)$.

The Parametric Bootstrap

The sampling distribution of $\hat{\theta}$

Our knowledge about F is incorporated into the bootstrap algorithm by substituting the parametric distribution $F_{\hat{\theta}}$ for the empirical distribution.

Step 1 Draw B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ from $\hat{F}_{\text{par}}$, where $\hat{F}_{\text{par}}$ is the parametric estimate of the population F.

Step 2 Evaluate bootstrap replications,

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), \quad b = 1, 2, \ldots, B.$$

Step 3 Estimate the sampling distribution of $\hat{\theta}$ by the empirical distribution of the bootstrap replications $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$:

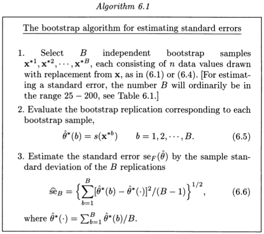
$$\widehat{\text{Prob}}\left(\hat{\theta} \in A\right) = \frac{1}{B} \sum_{b=1}^{B} I_{\left\{\hat{\theta}^*(b) \in A\right\}}$$

for appropriate subsets A.

Non-parametric vs Parametric Bootstrap

Non-parametric Bootstrap

$$X_1^{*1}, X_2^{*1}, \ldots, X_n^{*1} \sim \hat{F}$$
$$X_1^{*2}, X_2^{*2}, \ldots, X_n^{*2} \sim \hat{F}$$
$$\vdots$$
$$X_1^{*B}, X_2^{*B}, \ldots, X_n^{*B} \sim \hat{F}$$



Algorithm 6.1

The bootstrap algorithm for estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$, as in (6.1) or (6.4). [For estimating a standard error, the number B will ordinarily be in the range 25 – 200, see Table 6.1.]
2. Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \qquad b = 1, 2, \cdots, B. \tag{6.5}$$

3. Estimate the standard error $\mathrm{se}_F(\hat{\theta})$ by the sample standard deviation of the B replications

$$\widehat{\mathrm{se}}_B = \left\{ \sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2 / (B-1) \right\}^{1/2}, \tag{6.6}$$

where $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

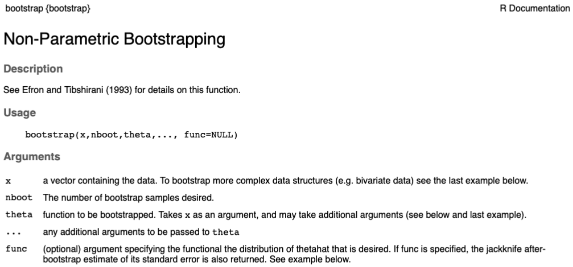
Parametric Bootstrap

$$X_1^{*1}, X_2^{*1}, \ldots, X_n^{*1} \sim \hat{F}_{\mathsf{par}}$$
$$X_1^{*2}, X_2^{*2}, \ldots, X_n^{*2} \sim \hat{F}_{\mathsf{par}}$$
$$\vdots$$
$$X_1^{*B}, X_2^{*B}, \ldots, X_n^{*B} \sim \hat{F}_{\mathsf{par}}$$

Packages in R

Two well known R packages concerned with the bootstrap.

- The “bootstrap” package, which is documents by our textbook.



bootstrap {bootstrap} R Documentation

Non-Parametric Bootstrapping

Description

See Efron and Tibshirani (1993) for details on this function.

Usage

```
    bootstrap(x,nboot,theta,..., func=NULL)
```

Arguments

`x` a vector containing the data. To bootstrap more complex data structures (e.g. bivariate data) see the last example below.

`nboot` The number of bootstrap samples desired.

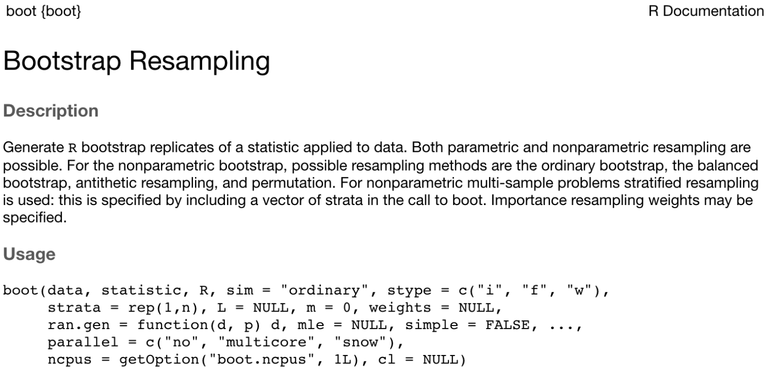
`theta` function to be bootstrapped. Takes `x` as an argument, and may take additional arguments (see below and last example).

`...` any additional arguments to be passed to `theta`

`func` (optional) argument specifying the functional the distribution of thetahat that is desired. If func is specified, the jackknife after-bootstrap estimate of its standard error is also returned. See example below.

- The “boot” package with the textbook Davison and Hinkley (1997): Bootstrap Methods and Their Applications.

The boot package



boot {boot} R Documentation

Bootstrap Resampling

Description

Generate `R` bootstrap replicates of a statistic applied to data. Both parametric and nonparametric resampling are possible. For the nonparametric bootstrap, possible resampling methods are the ordinary bootstrap, the balanced bootstrap, antithetic resampling, and permutation. For nonparametric multi-sample problems stratified resampling is used: this is specified by including a vector of strata in the call to boot. Importance resampling weights may be specified.

Usage

```
boot(data, statistic, R, sim = "ordinary", stype = c("i", "f", "w"),
     strata = rep(1,n), L = NULL, m = 0, weights = NULL,
     ran.gen = function(d, p) d, mle = NULL, simple = FALSE, ...,
     parallel = c("no", "multicore", "snow"),
     ncpus = getOption("boot.ncpus", 1L), cl = NULL)
```

The Boot package

Arguments:

- data: The data as a vector, matrix or data frame.
- statistic: A function which when applied to data returns a vector containing the statistic(s) of interest. When *sim = "parametric"*, the first argument to statistic must be the data. For each replicate a simulated dataset returned by *ran.gen* will be passed.
- R: The number of bootstrap replicates.
- sim: A character string indicating the type of simulation required. Possible values are "ordinary" (the default), "parametric", "balanced", etc.
- ran.gen: This function is used only when *sim = "parametric"* when it describes how random values are to be generated. It should be a function of two arguments. The first argument should be the observed data and the second argument consists of any other information needed (e.g. parameter estimates). The second argument may be a list, allowing any number of items to be passed to ran.gen
- mle: The second argument to be passed to ran.gen. Typically these will be maximum likelihood estimates of the parameters.

Non-Parametric bootstrap with boot package

Example:

```
library(boot)
x <- rexp(50)
boot_percentile <- function(x, i, p) quantile(x[i], p)
boot(x, boot_percentile, R = 999, p = 0.95)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = boot_percentile, R = 999, p = 0.95)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 2.658892 -0.01973989   0.5483944
```

```
b <- bootstrap(x, 999, theta = boot_percentile, p = 0.95)
sd(b$thetastar)
```

```
## [1] 0.5502894
```

Parametric bootstrap with boot package

Example:

Let $X_1, X_2, \ldots, X_n \sim \mathsf{Exp}(\lambda)$, where λ is an unknown quantity. We estimate λ by an estimator such as the MLE $\hat{\lambda}_n = 1/\overline{X}_n$. Let's find the sampling distribution of 95^{th} percentile. So we first generate the bootstrap samples:

$$\begin{gathered}
X_1^{*1}, X_2^{*1}, \ldots, X_n^{*1} \sim \mathsf{Exp}(\hat{\lambda}_n) \\
X_1^{*2}, X_2^{*2}, \ldots, X_n^{*2} \sim \mathsf{Exp}(\hat{\lambda}_n) \\
\vdots \\
X_1^{*b}, X_2^{*b}, \ldots, X_n^{*b} \sim \mathsf{Exp}(\hat{\lambda}_n)
\end{gathered}$$

Example: R code

```
library(boot)

x <- rexp(50)
bootpercentile <- function(x, p) quantile(x, p)
exp_boot <- function(x, mle)  rexp(length(x), mle)
b <- boot(x, bootpercentile, R = 999, sim = "parametric",
          ran.gen = exp_boot, mle = 1/mean(x), p = 0.95)
b
```

```
##
## PARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = bootpercentile, R = 999, sim = "parametric",
##     ran.gen = exp_boot, mle = 1/mean(x), p = 0.95)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 2.658892 -0.1986283   0.4604715
```

```
mean(b$t) - b$t0
```

```
##        95%
## -0.1986283
```

```
sd(b$t)
```

```
## [1] 0.4604715
```

Your Turn

Do this example in R

Example1: Let us consider a sample containing two hundred values generated randomly from a standard normal population $N(0, 1)$. This is the original sample. Then the sampling distribution of the sample mean is approximately normal with a mean 0 and a standard deviation $1/\sqrt(200)$. Apply the nonparametric bootstrap method to infer the result.

Example2.: Let $X_1, X_2, \ldots, X_n \sim N(0, \sigma^2)$, where σ^2 is unknown number. How do we estimate the variance of the sample variance?

- Note: Use the sample variance instead of MLE.
- Start with

```
x <- rnorm(50)
```