KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIWERSYTET GDAŃSKI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Bootstrap Methods in Data Processing.

Sample problems.

PWP Interdysciplinary Doctoral Studies in Mathematical Modeling
UDA-POKL.04.01.01-00-026/13-00
Projekt jest współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Problem 1. Let us consider the two-point distribution so the random variables X_i have only two values belonging to the set $\{0,1\}$ and $P(X_i = 1) = p$, $P(X = 0) = 1 - p$ for certain parameter $p \in (0,1)$. Find the distribution of the estimator T of the parameter p, given by

$$t = t(X_1, ..., X_N) = \frac{1}{N}\sum_{k=1}^{N} X_k.$$

Find the distribution of the bootstrap statistics t^*.

Solution: Let us summarize basic facts:

- for X_1, X_2, ..., X_N the space of all samples consists of the sequences $\{0,1\}^N$, so all possible values of the statistics t are $\{\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \dots, \frac{N}{N}\}$,
- for each $k \in \{0,1,2,\dots,N\}$ we have $t = \frac{k}{N}$ is and only if our sample (N-sequence) contains exactly k "ones" and $N - k$ "zeroes", so

  $$P\left(t = \frac{k}{N}\right) = \binom{N}{k} p^k (1-p)^{N-k},$$

  meaning that statistics Nt is distributed as the binomial distribution $B(N, p)$,
- this implies that $E(t) = p$ and $Var(t) = \frac{1}{N}p(1-p)$.

Now for the given sample $x_1, ..., x_N$ we may estimate p as the sample mean i.e. $t = \bar{x}$. Moreover each observation $x_i \in \{0,1\}$, so the bootstrap sample distribution is also two-point distribution, but with parameter $\bar{x}$. This means that t^* satisfies

$$Nt^* \sim B(N; \bar{x}).$$

Problem 2. Perform the simulation of the bootstrap method for two-point distribution and sample consisting of $N = 10$ values and $R = 10\ 000$ bootstrap samples. Use `boot` package in `R`.

Solution: In order to randomly choose $N = 10$ elements from the two-point distribution we call

```
data <- sample(0:1, 10, replace=T)
```

The result is

```
data
[1] 1 0 0 1 0 0 0 1 1 0
```

In order to find $R = 10\ 000$ bootstrap samples we call

```
library(boot)
samplemean <- function(x, d) mean(x[d])
b=boot(data,samplemean,R=10000)
```

In order to show results of the bootstrap procedure we should call

```
print(b)
```

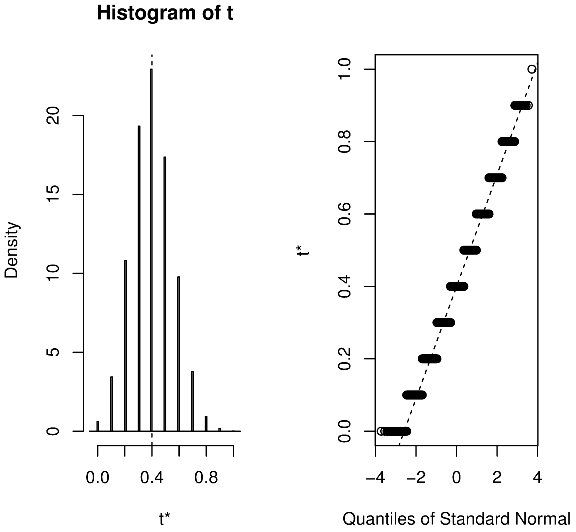
The result is:

```
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:  boot(data = data, statistic = samplemean, R = 10000)
Bootstrap Statistics :
     original    bias      std.  error
t1* 0.4        -0.00096   0.1546398
```

To see the histogram of the bootstrap distribution one can call

```
plot(b)
```

Then we can see the result as



Problem 3. Consider the data sample with three values $(1, -1, 0)$. Find the bootstrap distribution of the mean.

Solution: All possible bootstrap samples may be found as all 3-element sequences containing values from the set $\{-1, 0, 1\}$, so there are $3^3 = 27$ possible bootstrap samples. Actually we may list all of them, but what is important is to find all possible mean value estimations based on the bootstrap sample, and understand how often each of them appears. The mean value of each sample is one of $-1, -\frac{2}{3}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{2}{3}, 1$.

The table below shows the number of sequences corresponding to each of the possible bootstrap values

-1	$-\frac{2}{3}$	$-\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
1	3	6	7	6	3	1

This gives the bootstrap distribution

-1	$-\frac{2}{3}$	$-\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
1/27	3/27	6/27	7/27	6/27	3/27	1/27

Problem 4. The set of 5 values is collected given in the table below. Write R code generating the full bootstrap distribution of this sample.

1	2	3	4	5
4.21	4.60	1.82	3.61	4.26

Solution: We need to generate all bootstrap samples – i.e. all 5-element sequences consisting of values listed in the table. There are $5^5 = 3125$ such sequences. To do this we should perform the series of operations

```
x <- c(4.21,4.60,1.82,3.61,4.26)
library(gtools)
pp <- permutations(5,5,x,repeats=TRUE)
```

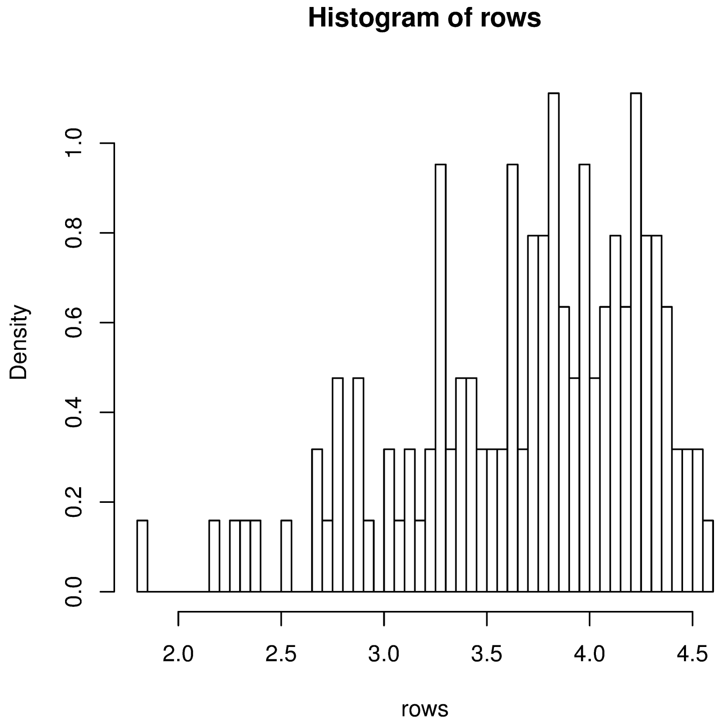
Now pp keeps the list of all possible bootstrap samples. Now we have to calculate the mean value for each sample.

```
rows <- rowMeans(pp)
```

Now we should present the data in the histogram:

```
hist(rows, freq=FALSE, breaks=50)
```

The result is as follows:



Problem 5. Prepare `R` code for the following experiment:

1. Generate 10 element random sample from the $N(0,1)$ distribution.
2. Generate 10 000 bootstrap samples (from the 10 element random sample) and based on this present the histogram of the bootstrap distribution of the mean.
3. Find the interval (centered at t^*) containing 95% of the bootstrap distribution.

Solution: Let us follow the plan sketched above:

1. Generate 10 element random sample from the $N(0,1)$ distribution.

   ```
   x <- rnorm(10,mean = 0, sd = 1)
   ```

   The sample is:

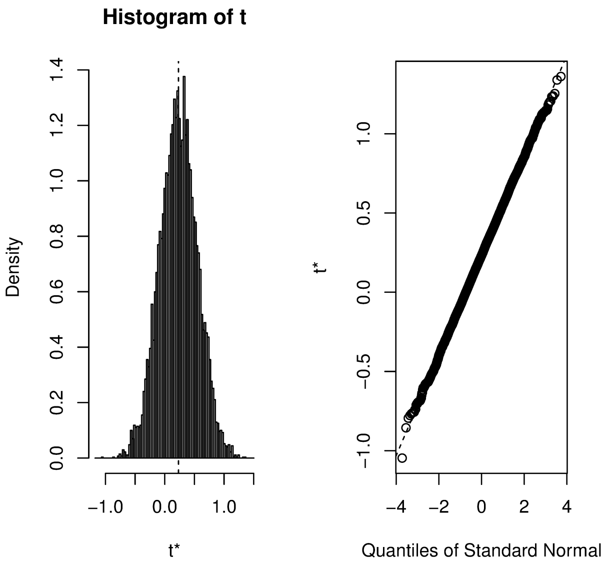
   ```
   0.45502956 1.30490684 -0.46707184 -0.07918046 -0.50447636 1.09248498 -1.60450834
   0.57982589 1.95010243 -0.41063331
   ```

2. Generate 10 000 bootstrap samples (from the 10 element random sample) and based on this present the histogram of the bootstrap distribution.

   This may be performed in the sequence of steps:

   ```
   library(boot)
   samplemean <- function(y, d) mean(y[d])
   b=boot(x,samplemean,R=10000)
   plot(b)
   ```

   The final histogram plot is:



3. Find the interval (centered at t^*) containing 95% of the bootstrap distribution.

   In order to find the confidence intervals we can use the function

   ```
   boot.ci(b,conf=0.95)
   ```

   Which returns (among other results)

   ```
   Percentile
   (-0.3871, 0.8436 )
   ```