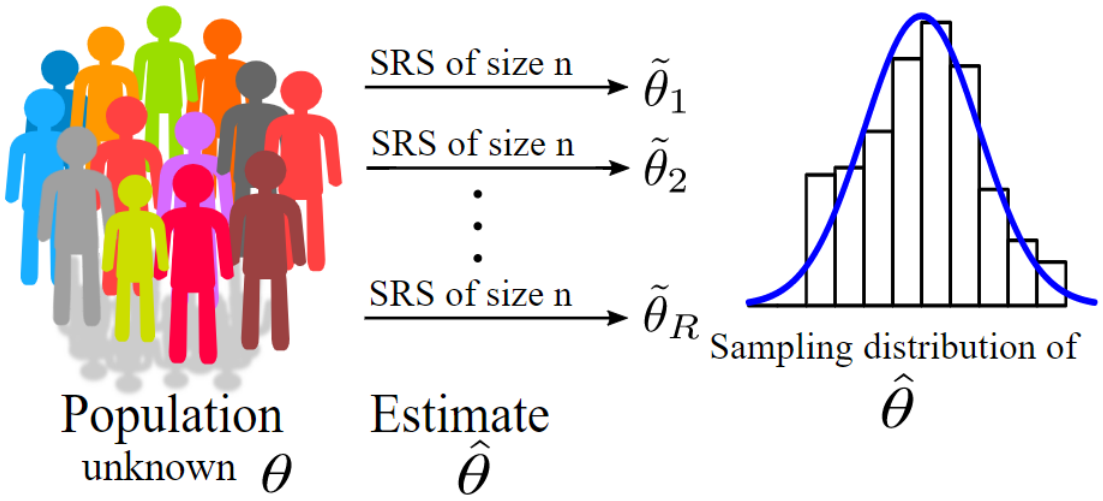
CM 2062 - Statistical Computing with R
Lab Sheet 12

Bootstrap Confidence Interval in R

Bootstrapping is a statistical method for inference about a population using sample data. It can be used to estimate the confidence interval(CI) by drawing samples with replacement from sample data. Bootstrapping can be used to assign CI to various statistics that have no closed-form or complicated solutions. Suppose we want to obtain a 95% confidence interval using bootstrap re-sampling the steps are as follows:

1. Sample n elements with replacement from original sample data.
2. For every sample calculate the desired statistic eg. mean, median etc.
3. Repeat steps 1 and 2 m times and save the calculated stats.
4. Plot the calculated stats which forms the bootstrap distribution
5. Using the bootstrap distribution of desired stat we can calculate the 95% CI

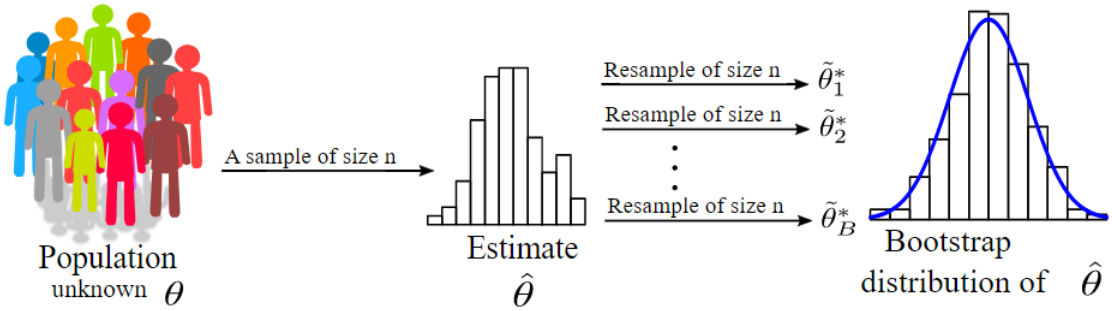
Figure Illustration of the sampling distribution



Bootstrap is analogous to the procedure to get the sampling distribution. First, from a population, we can get a random sample with size n. Using the sample, we can obtain an estimate of

θ denoted as $\hat{\theta}$. Second, using the random sample as a "population", we can randomly sample the subjects to form new samples with the same size. In doing so, we allow the subjects in the original sample to appear more than once in the new, bootstrapping samples. Each time we get a bootstrapping sample, we can calculate the statistic we are interested in, denoted by $\tilde{\theta}$. By repeating the procedure for B times, we can get a set of $\tilde{\theta}_j, j = 1...B$. With the set of values, we can get an empirical distribution, e.g., as a histogram, for $\hat{\theta}$. Inference about the parameter can be conducted by viewing the bootstrap distribution as the "sampling" distribution. This procedure is illustrated in the figure below.

Figure Illustration of the bootstrap distribution



Bootstrap standard error and confidence intervals

The bootstrap method is often used to obtain bootstrap standard error or confidence intervals for statistics of interest. The bootstrap standard error of the parameter estimate $\hat{\theta}$ can be calculated as

$$s.e(\hat{\theta}) = \sum_{j=1}^{B} (\tilde{\theta}_j - \bar{\tilde{\theta}}_j)^2/(B-1)$$

$$\bar{\tilde{\theta}} = \sum_{j=1}^{B} \tilde{\theta}_j/B$$

Then $(1-\alpha)100\%$ bootstrap confidence interval for θ is $(\hat{\theta} \pm z_{\alpha/2} s.e(\hat{\theta}))$.

Bootstrap confidence interval Methods

There are five main bootstrap confidence interval methods.

1. Normal bootstrap or Standard confidence limits method

Use the standard deviation for calculation of CI.
Use when statistic is unbiased.
Is normally distributed.

2. Basic bootstrap or Hall's (second percentile) method

Use percentile to calculate upper and lower limit of test statistic.
When statistic is unbiased and homoscedastic.
The bootstrap statistic can be transformed to a standard normal distribution.

3. Percentile bootstrap or Quantile-based, or Approximate intervals

Use quantiles eg 2.5%, 5% etc. to calculate the CI.
Use when statistic is unbiased and homoscedastic.
The standard error of your bootstrap statistic and sample statistics are the same.

4. BCa bootstrap or Bias Corrected Accelerated Method

Use percentile limits with bias correction and estimate acceleration coefficient corrects the limit and find the CI.
The bootstrap statistic can be transformed to a normal distribution.
The normal-transformed statistic has a constant bias.

5. Studentized bootstrap Method

Resamples the bootstrap sample to find a second-stage bootstrap statistic and use it to calculate the CI.
Use when statistic is homoscedastic.
The standard error of bootstrap statistic can be estimated by second-stage resampling.

In R the package **boot** allows a user to easily generate bootstrap samples of virtually any statistic that we can calculate. We can generate estimates of bias, bootstrap confidence intervals, or plots of bootstrap distribution from the calculated from the boot package.

Example

For demonstration purposes, we are going to use the iris dataset due to its simplicity and availability as one of the built-in datasets in R. The data set consists of 50 samples from each of the three species of Iris (Iris setosa, Iris Virginia, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. We can view the iris dataset using head command and note the features of interests.

Let's say we want to estimate the correlation between Petal Length and Petal Width and then we need to get the bootstrap confidence interval for this correlation estimate.

```
# View the first row
# of the iris dataset
head(iris, 1)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
  Species
1  setosa
```

```
# Custom function to find correlation
# between the Petal Length and Width

corr.fun <- function(data, idx)
{
  df <- data[idx, ]

  # Find the spearman correlation between
  # the 3rd and 4th columns of dataset
  c(cor(df[, 3], df[, 4], method = 'spearman'))
}

corr.fun(iris)
```

```
[1] 0.9376668
```

Using the boot function to find the R bootstrap of the statistic.

```
library(boot)

# Calling the boot function with the dataset
# our function and no. of rounds
bootstrap <- boot(iris, corr.fun, R = 1000)
bootstrap
```

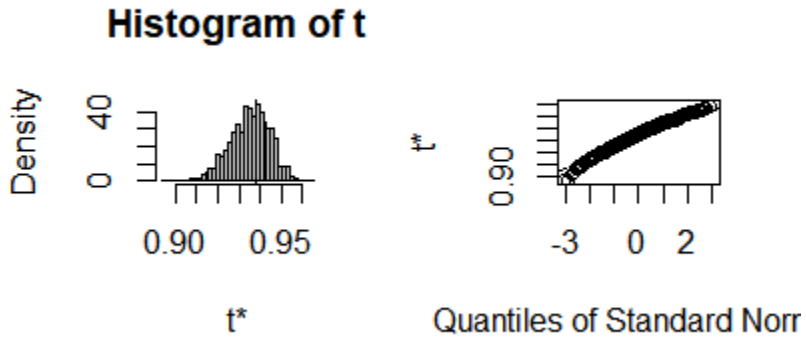
```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = iris, statistic = corr.fun, R = 1000)


Bootstrap Statistics :
     original        bias     std. error
t1* 0.9376668 -0.002434814 0.009643576
```

We can plot the generated bootstrap distribution using the plot command with calculated bootstrap.

```
plot(bootstrap)
```



Using the **boot.ci()** function to get the confidence intervals.

```
# Function to find the
# bootstrap Confidence Intervals
boot.ci(boot.out = bootstrap, type = c("norm", "basic","perc", "bca"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = bootstrap, type = c("norm", "basic", "perc",
    "bca"))

Intervals :
Level      Normal              Basic
95%   ( 0.9212,  0.9590 )   ( 0.9224,  0.9613 )
```

```
Level      Percentile            BCa
95%   ( 0.9140,  0.9530 )   ( 0.9176,  0.9546 )
Calculations and Intervals on Original Scale
```

Looking at the Normal method interval of (0.9219, 0.9589) we can be 95% certain that the actual correlation between petal length and width lies in this interval 95% of the time. As we have seen the output consists of multiple CI using different methods according to the type parameter in function boot.ci. The computed intervals correspond to the ("norm", "basic", "perc", "bca") or Normal, Basic, Percentile, and BCa which give different intervals for the same level of 95%. The specific method to use for any variable depends on various factors such as its distribution, homoscedastic, bias, etc.

Exercise 1
Consider the Sepal.Length variable in iris data set.

1. Obtain the bootstrap estimate for median of the Sepal.Length variable.
2. Plot the bootstrap estimate of the median of the Sepal.Length variable.
3. Construct the all five bootstrap confidence intervals for median of the Sepal.Length variable.

Exercise 2
Consider the Petal.Width variable in iris data set.

1. Obtain the bootstrap estimate for mean of the Petal.Width variable.
2. Plot the bootstrap estimate of the mean of the Petal.Width variable.
3. Construct the all five bootstrap confidence intervals for mean of the Petal.Width variable.

Exercise 3
Consider the data set which contain 15 paired observations of student LSAT scores and GPAs. Let's say We want to estimate the correlation between LSAT and GPA scores.

1. Obtain the bootstrap estimate for correlation between LSAT and GPA scores.
2. Plot the bootstrap estimate of the correlation between LSAT and GPA scores.
3. Construct the all five bootstrap confidence intervals for correlation between LSAT and GPA scores.

student	lsat	gpa
1	576	3.39
2	635	3.30
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3.00
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96