4. Resampling techniques: Bootstrap and Jackknife

4.1 The Bootstrap When we want to make statistical inference about a characteristic (random variable) of a population, we have random sample available which is representative of the population, but we **do not have any information about the distribuition of that random variable**, some dificulties arise.

Difficulties in performing inference:

- It could not be possible to evaluate the quality of the estimators: bias and efficiency (variance).
- The distributions of the statistics used to perform tests or estimate confidence intervals are in general also unknown.

The first task in the Bootstrap method is precisely to found an *approximated distribution* **for the interest variable.**

Bootstrap method

First of all: We suppose that we have **a very representative sample** of the population:

$x_1, x_2, ..., x_n$.

There are two possible types of information about the distribution of the interest variable.

1. We know that the distribution belongs to some parametric family indexed by the parameter (or vector of parameters) F_ψ, but the parameters are unknown. For example, we know that the distribution is the exponential distribution but we do not known the value of λ.

   In this case we use the available sample to estimate the unknown variable ψ. (We can use the maximum likelihood method to found an estimator for ψ since we know the distribution function). Then we add in the estimated parameter , $\hat{\psi}$, in the distribution and we obtain our estimated distribution function that we denote by,

$$\boxed{F_{\hat{\psi}}.}$$

2. The distribution function is completely unknown. In this situation we estimate F, through the empirical distribution function for the available sample, $\hat{F}$, that is, the distribution function associated to the probability mass function which assigns the same probability to all the observations of the sample, $x_1, x_2, ..., x_n$, That is,

$$\boxed{\hat{F}(x) = \frac{\#\{i \ : x_i \leq x\}}{N}.}$$

Now since we have already defined the approximated distribution function of the interest random variable, X, $F_{\hat{\psi}}$ or $\hat{F}(x)$, if we want to apply the Monte Carlo method we use this distribution functions to generate other samples of the random variable.

The samples of X, that are generated using the approximated distribution function are called Bootstrap samples and are denoted by,

$$\boxed{x_1^*, x_2^*, ..., x_n^*.}$$

Let $\hat{\theta}$ be an estimator of a given parameter θ of the population. When $\hat{\theta}$ is computed using a Bootstrap sample, we denote it by,

$$\boxed{\hat{\theta}^*.}$$

Bootstrap method - Properties of a statistic

Let $\hat{\theta}$ be a estimator of a parameter from one population. When we want to study the properties of the statistics or of a pivotal statistic $\hat{\theta} - \theta$ to perform tests or estimate a confidence interval, we need to know the distribuition function of the relevant statistics. If we do not know the distribution of the interest random variable, it is expected that the distribution of the statistic is also unknown.

Bootstrap method fundament: if the available sample is very representative of the population, then the distribution of $\hat{\theta} - \theta$ is approximately equal to the distribution of $\hat{\theta}^* - \hat{\theta}$.

$$\boxed{(\hat{\theta} - \theta) \overset{D}{\approx} (\hat{\theta}^* - \hat{\theta}).}$$

- $\hat{\theta}$ - statistic evaluated at the available sample.
- $\hat{\theta}^*$ - statistic evalueted at Bootstrap samples.

If we want to study the properties of $\hat{\theta} - \theta$, namely,

- $E(\hat{\theta} - \theta)$
- $Var(\hat{\theta} - \theta)$

We could use the sample estimators, for it we need to generate samples of $\hat{\theta}$. We could apply a Monte Carlo method by generating samples of X, however we do not know the distribution of X, then we can only use bootstrap samples of X. Then what we are going to evaluate is:

- $E(\hat{\theta}^* - \hat{\theta})$
- $Var(\hat{\theta}^* - \hat{\theta})$

This procedure is validated by $(\hat{\theta} - \theta) \overset{D}{\approx} (\hat{\theta}^* - \hat{\theta})$.

We compute samples of $\hat{\theta}^*$ using boostrap samples: see next slide.

To obtain a sample of $\hat{\theta}^*$, $\hat{\theta}^*_1, \hat{\theta}^*_1, ..., \hat{\theta}^*_R$ we generate Bootstrap samples of the random variable X :

$$\begin{array}{lccccccc}
\text{amostra 1:} & x_1^{*1} & x_2^{*1} & \cdots & x_n^{*1} & \rightarrow & \hat{\theta}^*_1 = T(x^{*1}) \\
\text{amostra 2:} & x_1^{*2} & x_2^{*2} & \cdots & x_n^{*2} & \rightarrow & \hat{\theta}^*_2 = T(x^*2) \\
 & & & \vdots & & & \\
\text{amostra R:} & x_1^{*R} & x_2^{*R} & \cdots & x_n^{*R} & \rightarrow & \hat{\theta}^*_R = T(x^{*R})
\end{array}$$

4.1.1 Bootstrap method - Bootstrap Mean and Variance

$$\boxed{E(\hat{\theta}-\theta) \approx \mathrm{b}_{\mathrm{boot}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r^* - \hat{\theta}).}$$

$$\boxed{Var(\hat{\theta}-\theta) \approx \mathrm{Var}_{\mathrm{boot}}(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_r^* - \bar{\hat{\theta}}^*\right)^2.}$$

4.1.2 Bootstrap method - Confidence intervals

Let us fix the pivotal statistic: $\boxed{\hat{\theta} - \theta.}$

When we know the distribution of $\hat{\theta} - \theta$ we estimate a confidence interval by computing the quantiles a_α and $a_{\alpha-1}$ which satisfy,

$$P(a_\alpha < \hat{\theta} - \theta < a_{1-\alpha}) = 1 - 2\alpha \;\Leftrightarrow\; \begin{cases} P(\hat{\theta} - \theta \leq a_\alpha) = \alpha \\ P(\hat{\theta} - \theta \geq a_{1-\alpha}) = \alpha \end{cases}$$

In this case the confidence interval is,

$$]\hat{\theta} - a_{1-\alpha}, \hat{\theta} - a_\alpha[.$$

Since we do not know the distribution of $\hat{\theta} - \theta$ we use the empirical distribution function of $\hat{\theta}^* - \hat{\theta}$ to estimate a_α e $a_{1-\alpha}$.

The estimators of a_α and $a_{1-\alpha}$ are the quantiles that satisfies,

$$P(\hat{a}_\alpha < \hat{\theta}^* - \hat{\theta} < \hat{a}_{1-\alpha}) = 1 - 2\alpha.$$

Remark: A **pivotal statistic** is a function that depends on a sample and an unknown parameter.

Bootstrap basic confidence interval

$$\boxed{\left]\hat{\theta}-\hat{a}_{\alpha-1},\hat{\theta}-\hat{a}_{\alpha}\right[=\left]2\hat{\theta}-\hat{\theta}^*_{((R+1)(1-\alpha))}\,,\,2\hat{\theta}-\hat{\theta}^*_{((R+1)\alpha)}\right[.}$$

How to compute $\hat{a}_{\alpha}$ e $\hat{a}_{\alpha-1}$:

Generate a sample $\hat{\theta}^*_1-\hat{\theta},\hat{\theta}^*_2-\hat{\theta},...,\hat{\theta}^*_R-\hat{\theta}$, add $0=\hat{\theta}-\hat{\theta}$ and reorder the sample in ascending order,

$$\hat{\theta}^*_{(1)}-\hat{\theta}\,,\,\hat{\theta}^*_{(2)}-\hat{\theta}\,,...,\,\hat{\theta}^*_{(R+1)}-\hat{\theta}$$

and define

$$\begin{cases}\hat{a}_{\alpha}=\hat{\theta}^*_{((R+1)\alpha)}-\hat{\theta}\\ \hat{a}_{1-\alpha}=\hat{\theta}^*_{((R+1)(1-\alpha))}-\hat{\theta}.\end{cases}$$

- $\theta^*_{((R+1)\alpha)}$ e $\theta^*_{((R+1)(1-\alpha))}$ are the elements in the ordered sample which are in the positions $(R+1)\alpha$ and $(R+1)(1-\alpha)$, respectively.
- Choose R in such a way that $(R+1)\alpha$ and $(R+1)(1-\alpha)$, are integer numbers.

Studentized-Bootstrap confidence interval

$$\boxed{\left]\hat{\theta} - z^*_{((R+1)(1-\alpha))}\sqrt{v}\,,\ \hat{\theta} - z^*_{((R+1)\alpha)}\sqrt{v}\right[.}$$

This interval is obtained by using the pivolal statistic:

$$\boxed{\hat{z} = \frac{\hat{\theta}-\theta}{\sqrt{\mathrm{var}(\hat{\theta}-\theta)}}.}$$

$v = \mathrm{Var}_{\mathrm{boot}}(\hat{\theta})$.

Generate a sample z_1^* , z_2^* , ... , z_R^*, add 0 and reorder in ascending order.

$$z_r^* = \frac{\hat{\theta}_r^* - \hat{\theta}}{\sqrt{v_r^*}} \qquad \text{and} \qquad v_r^* = \mathrm{Var}_{\mathrm{boot}}(\hat{\theta}_r^*).$$

- $z^*_{((R+1)\alpha)}$ and $z^*_{((R+1)(1-\alpha))}$ are the elements in the ordered sample which are in the positions $(R+1)\alpha$ and $(R+1)(1-\alpha)$, respectively.
- Choose R in such a way that $(R+1)\alpha$ and $(R+1)(1-\alpha)$, are integer numbers.