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.

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

$x_1, x_2, \ldots, x_n$.

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,

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

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,

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

Bootstrap method - Bootstrap Samples

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,

$$x_1^*, x_2^*, \ldots, x_n^*.$$

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,

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

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

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

- $\hat{\theta}$ - 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 :

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

4.1.1 Bootstrap method - Bootstrap Mean and Variance

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

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

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

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

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

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

This interval is obtained by using the pivolal statistic:

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

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

Generate a sample z^*_1 , z^*_2 , ... , z^*_R, add 0 and reorder in ascending order.

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

- $z^*_{((R+1)\alpha)}$ 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.

Exercise

Consider a sample of a bidimensional random variable with unknown distribution F.

Inhabitants ($\times$ 10 000) of 10 cities in 1930 and 1950.

1930	138	93	61	179	48
1950	143	104	69	260	75

1930	37	29	23	30	2
1950	63	50	48	111	50

The quantity of interest is

$$\theta = \frac{\int\int x dF(x,y)}{\int\int y dF(x,y)} = \frac{\mu_X}{\mu_Y}.$$

The nonparametric estimate of this is $\hat{\theta} = t(\hat{\mathrm{F}}) = \frac{\overline{X}}{\overline{Y}}$.

1. Compute the original estimate $\hat{\theta} = t(\hat{F})$.
2. Compute the bootstrap bias and variance of $\hat{\theta} - \theta$ using the Bootstrap technique.
3. Compute the basic Bootstrap interval for $\hat{\theta}$.

4.1.3. The Bootstrap - Hypothesis Testing

Recall:

- To test a null hypothesis H_0 we specify a test statistic T with observed value t_{obs} and we calculate the p-value $p = \Pr(T > t|H_0)$.
- To do this we need to know the distribution of T when the null hypothesis is true.
- Even for simple H_0, the distribution of T may not be easy to find and we may use Monte Carlo tests.

- If we know the distribution F_0 of the random variable X under the null hypothesis then we generate samples of the random variable and for each sample we compute the tests statistic. This way we create a sample of the test statistic and we can use it to estimate the p-value.

- When the distribution F_0 of the random variable X is unknown even under the null hypothesis then to generate a sample of the statistic we generate samples of the random variable X using the approximate distributions of F_0, namely, $\hat{\mathrm{F}}_0$ or $(\mathrm{F}_{\hat{\psi}})_0$ depending on the information that we have about F_0.

$$p_{\text{boot}} = \frac{1 + \#\{t_r^* \geq t_0\}}{R+1}.$$

where $t_r^* = T(x_r^*)$ is the estimated value of the statistic for a sample generated from the distribution $\hat{\mathrm{F}}_0$ or $(\mathrm{F}_{\hat{\psi}})_0$.

Let us consider a specific test that compare two distributions.

Suppose that we have two independent samples,

$$\mathcal{X} = (x_1, ..., x_{n_1}) \text{ from } \mathrm{F}_1$$
$$\mathcal{Y} = (y_1, ..., y_{n_2}) \text{ from } \mathrm{F}_2$$

Test: $H_0 : \mathrm{F}_1 = \mathrm{F}_2 \vee H_1 : \mathrm{F}_1 \neq \mathrm{F}_2$

The test statistic T depends on the caracteristic of the distribution that we want to compare. Suppose it is the mean. Then,

$$T = |\mu_1 - \mu_2|.$$

Under the null hypothesis, X and Y have a common distribution F so the joint sample $\mathcal{U} = (\mathcal{X}, \mathcal{Y})$ is a sample from F with dimension $n_1 + n_2$.

An approximated distribution of the common distribution F_0 is the empirical distribution for the joint sample $\mathcal{U} = (\mathcal{X}, \mathcal{Y})$, $\hat{F}_0 = \hat{\mathrm{F}}_{\mathcal{U}}$.

Then we generate two samples from the distribution $\hat{\mathrm{F}}_{\mathcal{U}}$ with dimension dimension n_1 and n_2 and estimate

$$T^* = |\bar{X}^* - \bar{Y}^*|.$$

We simulate R times and estimate T^* and then,

$$p_{\text{boot}} = \frac{1 + \#\{t_r^* \geq t_0\}}{R + 1}.$$

Consider the following independent samples:

Sample 1: $(0, 4, 3, 1, 2, 3, 4, 3, 3, 0, 1, 4, 4, 0, 4)$

Sample 2: $(0, 2, 1, 6, 4, 2, 1, 1, 7, 2, 1, 2, 6, 3, 5, 3, 0, 5, 10, 4)$

We want to test if both samples comes from the same distribution, more precisely, if they have the same expected value, that is,

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2.$$

Consider the statistic which is the diference between the means.

$$T(x, y) = |\bar{x} - \bar{y}|.$$

4.2. The Jackknife

Suppose that $\theta = t(F)$.

Our aim is estimate the bias and variability of the estimator.

It is often of interest to know how θ behaves due to minor perturbations of F.

This is essentially the derivative of $t(F)$ with respect to the function F.

This derivative is often called the influence function of t.

Definition: Suppose that F is a cumulative distribution function and $t(\cdot)$ is some functional. Then **the influence function** of t at F is the function

$$L_t(y; F) = \lim_{\varepsilon \to 0} \frac{t[(1-\varepsilon)F + \varepsilon H_y] - t(F)}{\varepsilon}$$

and

$$H_y(u) = \begin{cases} 0 & , \quad u < y \\ 1 & , \quad u \geq y \end{cases}$$

Suppose $x_1, x_2, ..., x_n$ is a dataset and $\hat{F}$ is the empirical function of the data.

The **empirical influence function** is defined to be

$$I(y) = L_t(y; \hat{F}).$$

The values of the empirical influence function at the data points are called the **empirical influence values**

$$I_j = I(x_j) = L_t(x_j; \hat{F}), \quad j = 1, 2, ..., n.$$

An extension of Taylors Theorem says that for functionals t and measures G and F

$$t(G) \simeq t(F) + \int L_t(y; F) dG(y)$$

This is an exact result if t is a linear statistic.

Applying this with F a cdf and $G = \hat{F}$ we get

$t(\hat{F}) \simeq t(F) + \int L_t(y; F) d\hat{F}(y)$, that is,

$$t(\hat{F}) \simeq t(F) + \frac{1}{n} \sum_{j=1}^{n} L_t(x_j; \hat{F}) = t(F) + \frac{1}{n} \sum_{j=1}^{n} l_j$$

$$\Leftrightarrow \theta - \hat{\theta} = -\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{l_j}$$

.

The jackknife provides a way of approximating the empirical influence values through resampling the data.

$$l_j = \lim_{\varepsilon \to 0} \frac{t[(1-\varepsilon)\hat{F} + \varepsilon H_{x_j}] - t(\hat{F})}{\varepsilon}$$

$$\simeq \frac{t[(1-\varepsilon)\hat{F} + \varepsilon H_{x_j}] - t(\hat{F})}{\varepsilon},$$

If we take $\varepsilon = -\frac{1}{n-1}$ then

$$(1-\varepsilon)\hat{F} + \varepsilon H_{x_j} = \frac{n}{n-1}\hat{F} - \frac{1}{n-1}H_{x_j} = \hat{F}_{-j}$$

is a distribution with no weight on the point x_j and weight $\frac{1}{n-1}$ on the rest of the sample.

This is equivalent to just having the sample of size $n-1$ found by omitting x_j from the original sample.

Then the jackknife approximation to the empirical influence value l_j is

$$l_{\text{jack};j} = (n-1)[t(\hat{F}) - t(\hat{F}_{-j})] = (n-1)(\hat{\theta} - \hat{\theta}_{-j}).$$

The jackknife unbiased estimates of the bias and variance that use the Jackknife empirical influence values that we call Jackknife bias and Jackknife variance are:

- $\text{b}_{\text{jack}} = -\frac{1}{n}\sum_{j=1}^{n} l_{\text{jack};j}$.
- $\text{Var}_{\text{jack}} = \frac{1}{n(n-1)}\left(\sum_{j=1}^{n} l_{\text{jack};j}^2 - n\text{b}_{\text{jack}}^2\right)$.

It is common to use the Jackknife technique to evaluate the bias and variance of the Monte Carlo estimators.

For example, to estimate the variance in the studentized confidence intervals t-bootstrap, $z_r^* = \dfrac{\hat{\theta}_r^* - \hat{\theta}}{\sqrt{v_r^*}}$, we can estimate v_r^* using the Jackknife estimator avoiding this way a new bootstrap iteration.

Exercise:
Let $\mu = t(F) = \int x dF(x)$.
Estimate the Jackknife bias and the variance for $\hat{\theta} = \hat{\mu} = t(\hat{F})$.