1921 UN SECOLO DI STORIA DAVANTI A NOI 2021

UNIVERSITÀ CATTOLICA del Sacro Cuore

POINT ESTIMATE OF THE STANDARD ERROR: THE JACKNIFE AND THE BOOTSTRAP

INTRODUCTION

Estimation problem.

Assume that we observe a sample of size n of data following a certain (unknown) distribution F:

$$\mathbf{x} = (x_1, \ldots, x_n), \text{ with } x_i \sim F \quad \forall i \in \{1, \ldots, n\}$$

We focus on the problem of estimating one (or more) parameter θ of the distribution F. We denote it as $\theta = t(F)$.

Example. Expected value of the distribution:

$$\theta = \mathbb{E}_F(x)$$

INTRODUCTION

Estimation problem.

Assume that we observe a sample of size n of data following a certain (unknown) distribution F:

$$\mathbf{x} = (x_1, \dots, x_n), \text{ with } x_i \sim F \quad \forall i \in \{1, \dots, n\}$$

We focus on the problem of estimating one (or more) parameter θ of the distribution F. We denote it as $\theta = t(F)$.

Aims:

1. Give a point estimate of *t(F)* for a general case with unknown F (nonparametric point estimation);
2. Assess the standard error of the estimate and its bias (nonparametric point estimation of the standard error);
3. Give an interval estimate of *t(F)* and assess the interval coverage probability.

1. POINT ESTIMATE OF THETA

The plug-in principle.

The plug-in principle is a simple method to estimate the parameter θ from the sample $\mathbf{x}$.

Idea:

- Denote as $\hat{F}$ the empirical distribution function, that is the discrete distribution that puts probability $1/n$ on each value x_i $i = 1, \ldots, n$ and zero otherwise.
- The plug-in estimator $\hat{\theta}$ of the parameter $\theta = t(F)$ is defined as $\hat{\theta} = t(\hat{F})$.

Example. Expected value of the distribution:

$$\theta = \mathbb{E}_F(x) \quad \Rightarrow \quad \hat{\theta} = \mathbb{E}_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1. POINT ESTIMATE OF THETA

- The plug-in estimator is easy to compute in all cases (it is just a functional of a discrete probability distribution that is always well defined).
- No matter how complex is theta, its plugin estimator can often be numerically computed very easily.
- The plug-in estimator does not need to specify a parametric form of the data distribution F, so it is very flexible.
- If information is available of the distribution F, the parametric estimator tend to have better properties (lower variance and bias).

2. STANDARD ERROR AND BIAS OF $\hat{\theta}$

How good is the estimate of a parameter?

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\theta - \hat{\theta})^2] = \text{Var}(\hat{\theta}) + (\theta - \mathbb{E}(\hat{\theta}))^2$$

Variance of the estimator. How variable is $\hat{\theta}$ on different data sets

Bias of the estimator. How different is in average $\hat{\theta}$ from the true value

In order to assess how good is the estimator, we need to compute (or find an estimate for ...) the variance and the bias, without knowing the distribution *F*.

2. STANDARD ERROR AND BIAS OF $\hat{\theta}$

Methods for estimating variance and bias of an estimator:

- Analytic computation (only with F known and simple θ)
- **Monte Carlo simulation** (only with F known)
- **Jacknife** (F unknown)
- **Bootstrap** (F unknown or known)

MONTE CARLO SIMULATION

Imagine that F is fully know, we could answer then the question about the distribution of $\hat{\theta}$ by analytical calculation (very rarely) or by simulations.

Simulation with F known.

- For $b = 1, 2, \ldots, R$:
  - generate a random sample (of size n) $\mathbf{x}_b^* = (x_{1_b}^*, \ldots, x_{n_b}^*) \sim F$
  - compute $\hat{\theta}_b^* = \hat{\theta}(\mathbf{x}_b^*)$ and collect the obtained value
- Output after B iterations: $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$
- Use the output to estimate variance and bias:
  - $\widehat{\text{Var}}(\hat{\theta})_{MC} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2$, with $\hat{\theta}_{(\cdot)}^* = \sum_{b=1}^{B} \hat{\theta}_b^*$
  - $\widehat{\text{Bias}}(\hat{\theta})_{MC} = \theta - \hat{\theta}_{(\cdot)}^*$

Note that the values $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ can be used also to estimate the distribution of $\hat{\theta}$ (histogram, density estimator, ...).

MONTE CARLO SIMULATION

- The methods gives us an estimate of bias and variance of the parameter, as well as an estimate of its distribution, that can be used to perform **inference** (interval estimation, tests).
- Thanks to the **law of large numbers** we know that as B tends to infinity, we get a perfect match to theoretical calculation.
- However in reality we can't simulate an infinite number of replicates, and hence we introduce the Monte Carlo Error.

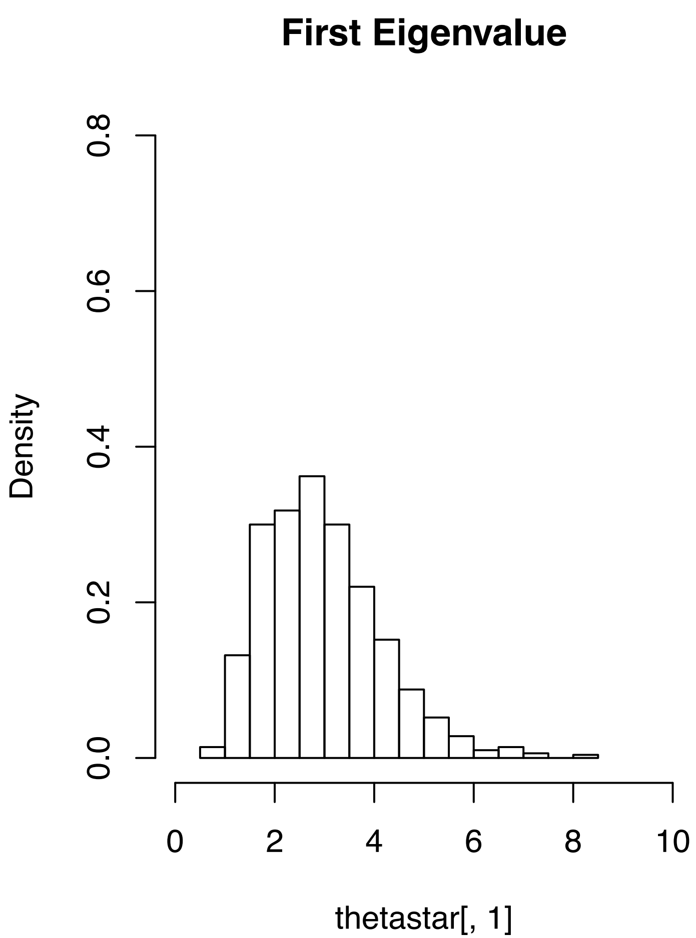
Example:

We observe a sample of size 10 drawn from a bivariate Normal distribution with Mean (0,0) and covariance matrix rbind(c(2,1) ; c(1,2)).

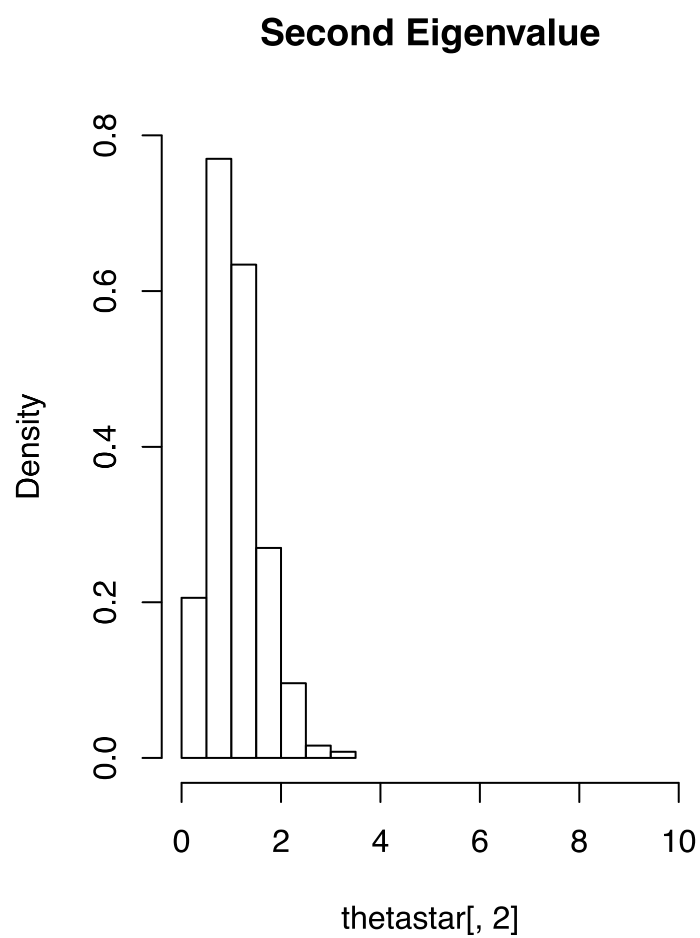
We are interested in estimating the eigenvalues of the covariance matrix, using the plugin estimator.

How variable is the estimator?

MONTE CARLO SIMULATION



Standard deviation: 1.17



Standard deviation: 0.51

JACKNIFE

- Monte Carlo simulation is very simple, but in reality we do not have access to the true distribution *F*.
- In such cases, we need to find a method to (approximately) simulate samples from *F*.

Remember that we start from a sample $\mathbf{x}$ of size n, with $x_i \sim F$, $\forall i = 1, \dots, n$. Assume that we want to estimate the parameter $\theta = t(F)$ with the estimator $\hat{\theta} = s(\mathbf{x})$ (not necessarily the plug-in estimator).
We focus on samples that leave out one observation at a time:

$$\mathbf{x}_{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

Jacknife samples. For a sample of size n, we can construct *n* Jacknife samples.

$$\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$$

Jacknife replications. *n* estimates of theta, each based on *n-1* units.

JACKNIFE

Estimate of variance:

$$\widehat{\mathrm{Var}}_J = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2$$

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$$

- Looks at the sample variance of the Jacknife replications.
- The sum of squares is inflated with the factor *(n-1)/n*, which is bigger than the usual terms *1/(n-1)* or *1/n*.
- The inflation factor is needed since the Jacknife samples are very similar between each other, so $\hat{\theta}_{(i)}$ tends to vary less than $\hat{\theta}$.
- In the special case of estimating the sample mean, the inflation factor makes the estimate of the variance correct.

JACKNIFE

Special case: $\hat{\theta} = \overline{X}$.
By theory we know:

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

$$\widehat{\mathrm{Var}}(\overline{X}) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

And for the Jacknife estimate $\hat{\theta}_{(i)} = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} X_j$ we have:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{(i)}$$

$$= \frac{1}{n(n-1)}\sum_{j=1,j\neq i}^{n} X_j = \overline{X}.$$

The Jacknife estimate of the variance is then:

$$\widehat{\mathrm{Var}}_J = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \overline{X})^2$$

JACKNIFE

Now:

$$\hat{\theta}_{(i)} - \overline{X} = \frac{1}{n-1}\sum_{j\neq i} X_j - \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$= \frac{1}{n(n-1)}\left[n\sum_{j\neq i} X_j - (n-1)\sum_{j=1}^{n} X_j\right] = \frac{\overline{X} - X_i}{n-1}$$

And finally:

$$\overline{\mathrm{Var}}_J = \frac{n-1}{n}\sum_{i=1}^{n}\frac{(\overline{X} - X_i)^2}{(n-1)^2} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\overline{X} - X_i)^2$$

- In the special case of estimating the sample mean, the inflation factor makes the estimate of the variance coinciding with the plug-in estimator of the variance of the sample mean (that is in this case unbiased).
- Note that this holds, however, just in this case!

JACKNIFE

Estimate of bias:

$$\widehat{\text{Bias}}_J = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

- Looks at the deviations between the Jacknife replications and the estimate.
- This deviation is also inflated with a factor, that is in this case *(n-1)*, which is bigger than 1.
- The inflation factor is needed since the Jacknife samples are very similar between each other, and they also tend to be similar to the estimate based on the whole sample.
- In the special case of estimating the sample mean, the bias is zero.
- In the case of the sample variance(divided by *n*), the inflation factor makes the estimate of the bias correct.

JACKNIFE

Special case: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

By theory we know:

$$\text{Bias}(\hat{\theta}) = -\frac{1}{n}\sigma^2.$$

And, for the jackknife estimate of bias, it is possible to show that:

$$\widehat{\text{Bias}}_J = -\frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right).$$

- In the special case of estimating the sample variance with its biased estimator, the inflation factor makes the estimate of the bias coinciding with the plug-in estimator of the bias.
- Note that this also holds, however, just in this case!

JACKNIFE

Example: test score data.

We consider the test score data from Mardia et al. (1979). A group of n=88 students took five tests in Mechanics, Vectors, Algebre, Analysis, and Statistics.

The quantity that we would like to estimate is the ratio between the largest eigenvalue of the covariance matrix of the scores, and the sum of all eigenvalues of the covariance matrix.

$$\theta = \frac{\lambda_{(1)}}{\sum_{j=1}^{5} \lambda_j}$$

The closest is this ratio to one, the most the model can be reduced to a 1D model where each student has a QI that explains the scores to all five tests (a single measure can capture all information).

$$\mathbf{x}_i = Q_i \mathbf{v}_{(1)}$$

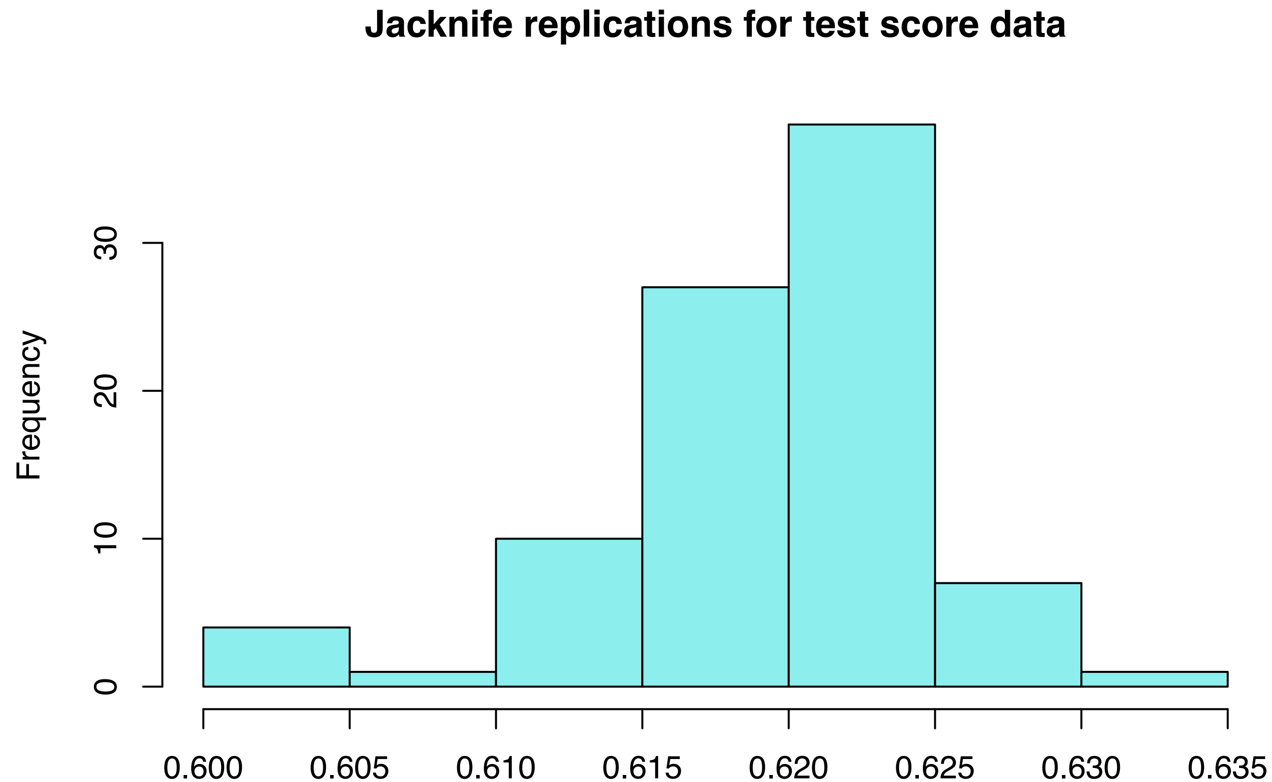
JACKNIFE

Example: test score data.

$$\hat{\theta} = 0.6191, \quad \widehat{\text{se}}_J = 0.0496, \quad \widehat{\text{Bias}}_J = 0.0011$$

Example: test score data.

$$\hat{\theta} = 0.6191, \quad \widehat{\text{se}}_J = 0.0496, \quad \widehat{\text{Bias}}_J = 0.0011$$



JACKNIFE

- The jackknife estimates of the variance and bias are very easy to define.
- The computational time is very low: we only need to create n Jacknife data sets, and n replications of theta.
- The formulas for Jacknife variance and bias do not have any theoretical justification in general. In very special cases, we can prove good properties (for instance, that they are unbiased estimates of the variance and bias).
- It is not trivial to create confidence intervals with good properties based on the Jacknife estimates.
- Jacknife can fail when the estimator of theta is not smooth.

JACKNIFE

- The jackknife estimates of the variance and bias are very easy to define.
- The computational time is very low: we only need to create n Jacknife data sets, and n replications of theta.
- The formulas for Jacknife variance and bias do not have any theoretical justification in general. In very special cases, we can prove good properties (for instance, that they are unbiased estimates of the variance and bias).
- It is not trivial to create confidence intervals with good properties based on the Jacknife estimates.
- **Jacknife can fail when the estimator of theta is not smooth.**

JACKNIFE

An example of Jacknife failure: $\hat{\theta} = \text{Median}(F)$

- We simulate a data set of sample size n=20 from a binomial distribution of size 10 and probability 0.5 .
- The true median is of course 0.5 * 10=5 (the distribution is symmetric).
- When estimating the median with the Jacknife, we notice that the Jacknife replications are all really similar (they all coincide in most cases).
- In many cases (when repeating the experiment), all Jacknife replications are equal to 5 (the true median).
- When computing the Jacknife estimate of the variance, we obtain zero!
- Jacknife estimates are too similar to the original data set.

JACKNIFE

An example of Jacknife failure: $\hat{\theta} = \mathrm{Median}(F)$

- We simulate a data set of sample size n=10 from a binomial distribution of size 10 and probability 0.5 .
- The true median is of course 0.5 * 10=5 (the distribution is symmetric).
- When estimating the median with the Jacknife, we notice that the Jacknife replications are all really similar (they all coincide in most cases).
- In many cases (when repeating the experiment), all Jacknife replications are equal to 5 (the true median).

When computing the Jacknife estimate of the variance, we obtain zero!

- Jacknife estimates are too similar to the original data set.

Delete-*d* Jacknife: leave out *d* observations at a time.

DELETE-*d* JACKNIFE

We can try to leave out d observations at a time. In this case the estimate of the variance is:

$$\widehat{\text{Var}}_{J-d} = \frac{n/d}{\binom{n}{d}} \sum_{s=1}^{\binom{n}{d}} (\hat{\theta}_{(s)} - \hat{\theta}_{(\cdot)})^2.$$

Number of Jacknife samples

It can be shown that the delete-d Jacknife is consistent for the median when $\sqrt{n}d \to \infty$ and $n - d \to \infty$.
In practice, we have to leave out more than $\sqrt{n}$ and less than n observations at a time.