SD-TSIA204
Confidence intervals

Ekhine Irurozki
Telecom Paris

December 2021

Confidence interval

- Context: regard an estimator $\widehat{g}(y_1, \ldots, y_n)$ for the value g.
  We would like to have an interval $\widehat{I}$ around $\widehat{g}$ which contains g with high probability.

- We construct $\widehat{I} = [\underline{C}, \overline{C}]$ based on the observations $(y_1, \ldots, y_n)$:
  confidence interval is a random variable

$$\mathbb{P}(\widehat{I} \text{ contains } g) = \mathbb{P}(\underline{C} \leq g \text{ et } \overline{C} \geq g) = 95\%$$

Confidence interval of level α

A confidence interval of **level** α for a value g is a function of the sample

$$\widehat{I} : (y_1, \ldots, y_n) \mapsto \widehat{I} = \left[\underline{C}(y_1, \ldots, y_n), \overline{C}(y_1, \ldots, y_n)\right]$$

such that

$$\mathbb{P}\left[g \in \widehat{I}(y_1, \ldots, y_n)\right] \geq 1 - \alpha$$

Rem:usual choices $\alpha = 5\%, 1\%, 0.1\%$, etc. Defined often by the consideration data complexity / number of observations.

Rem:In the following we will denote confidence interval by CI.

Example: survey

- Election survey with two candidates: A and B.
- Aim: estimate p.
- Sample of size n: a reasonable estimator is then

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}_n \, .$$

- The goal is to establish an oracle: is thare a clear winner in this sruvey? is this estimator relevant? Is algorithm A better than B?
- What is the confidence interval for p ?
- is this estimator likeli or not?

Survey: confidence interval

- Search for an interval $\widehat{I} = [\widehat{p} - \delta, \widehat{p} + \delta]$ such that $\mathbb{P}(p \in \widehat{I}) \geq 0.95 \Leftrightarrow$ search for δ such that $\mathbb{P}\left[\,|\widehat{p} - p| \;> \delta\right] \leq 0.05$
- Constituent: **Tchebyschev** inequality

$$\boxed{\forall \delta > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| > \delta) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\delta^2}}$$

For $X = \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ we know that $\mathbb{E}(\widehat{p}) = p$ and $\mathbb{V}\mathrm{ar}(\widehat{p}) = \frac{p(1-p)}{n}$:

$$\forall p \in (0,1), \forall \delta > 0, \quad \mathbb{P}(|\widehat{p} - p| > \delta) \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

Application: for a CI of 95%, find δ such that $\frac{1}{4n\delta^2} = 0.05$, *i.e.* $\delta = (0.2n)^{-1/2}$. For $n = 1000$, $\widehat{p} = 55\%$:

$$\delta = 0.07 \; ; \quad \widehat{I} = [0.48, 0.62]$$

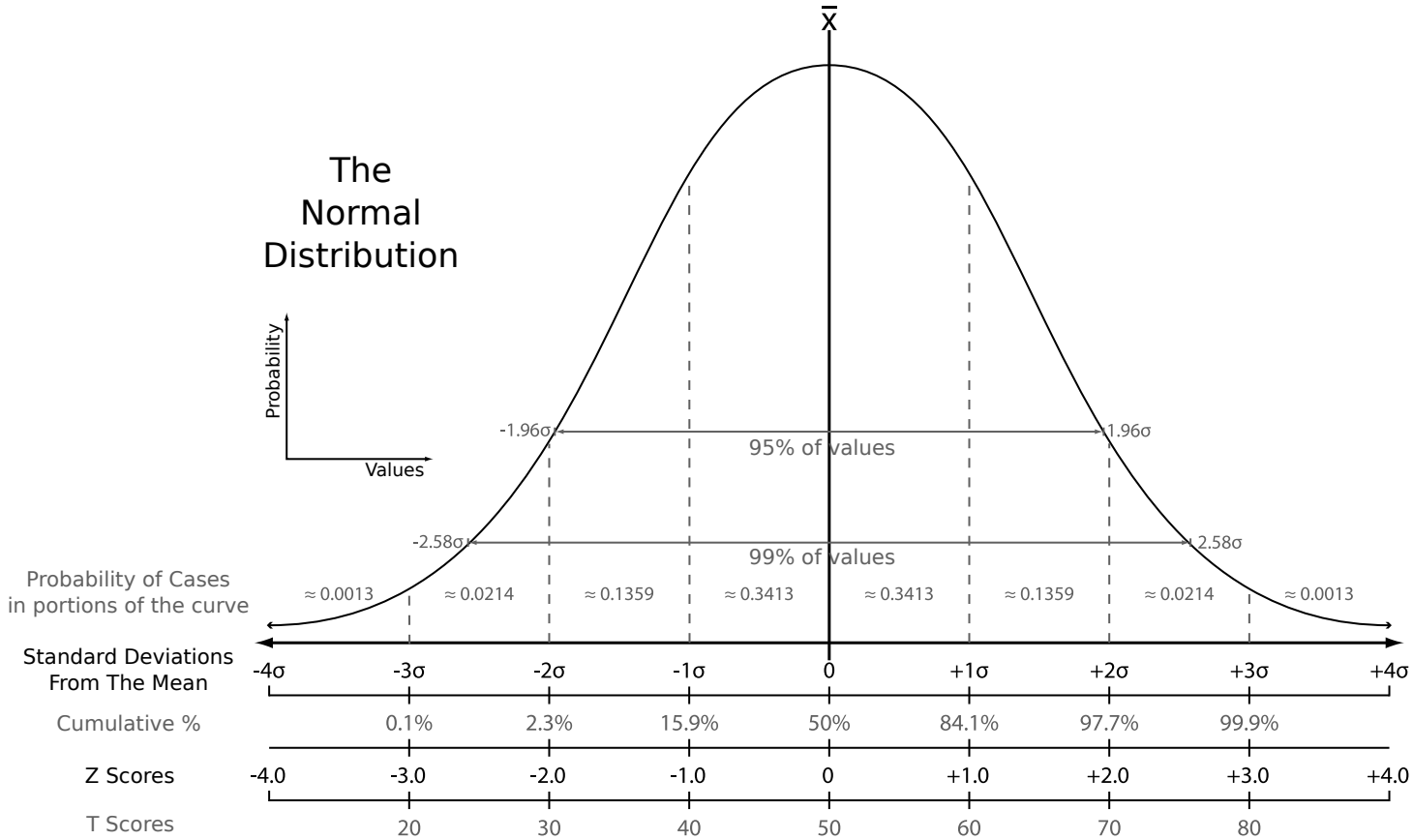
example for a rv iid N(0,1)

There are many concentration inequalities

- Markov's inequality $\mathsf{Pr}(X \geq \delta) \leq \dfrac{\mathsf{E}(X)}{\delta}$
- Hoeffding's inequality $\mathsf{Pr}\left[n^{-1}\sum_n |S_n - E_n| > \delta\right] < 2\exp\left(-\frac{2n\delta^2}{\sum_{i=1}^n c_i^2}\right)$

Applications in sample complexity:

Gaussian case



1. Confidence interval: reminder
   - Definition
   - Limit theorems

2. Confidence intervals for linear model

Central limit theorem

Let

- $y_1, y_2, \ldots,$ *i.i.d.* square integrable random variables.
- μ and σ their theoretical mean and variance.

Central limit theorem (CLT) Properly normalized average $\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right)$ converges towards a standard normal distribution $\mathcal{N}(0,1)$.

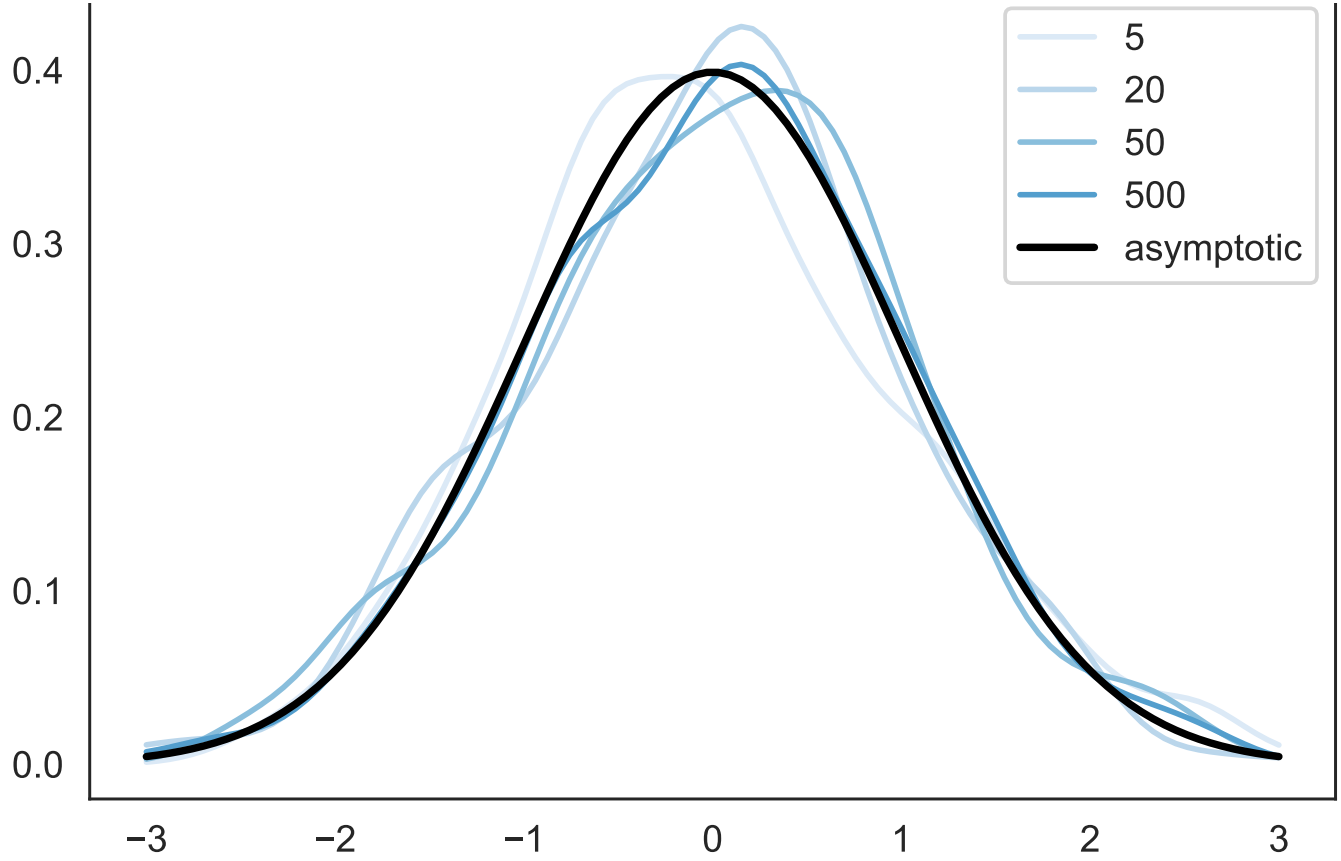
- σ is known.

Slutsky lemma Distribution of the "studentized" average $\sqrt{n}\left(\frac{\bar{y}_n - \mu}{\widehat{\sigma}}\right)$ converges towards a standard normal distribution $\mathcal{N}(0,1)$ when $\widehat{\sigma} \to \sigma$

Reformulation: $\bar{y}_n \simeq \mathcal{N}(\mu, \widehat{\sigma}^2/n)$

Illustration

https://github.com/sukhdev01/Central-Limit-Theorem/blob/master/Central_Limit_Theorem.ipynb



CLT: convergence of $\bar{y}$ density w.r.t. the number of observations

Asymptotic confidence intervals

- The survey example: $y_i \in \{0, 1\}$, $n = 1000$,

$$\widehat{p} = n^{-1} \sum_{i=1}^{n} y_i = 0.55$$

- We assume that n is sufficiently large, such that

$$\sqrt{n} \left(\frac{\widehat{p} - p}{\widehat{\sigma}} \right) \sim \mathcal{N}(0, 1)$$

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - \widehat{p})^2 = \widehat{p} - \widehat{p}^2$$

- We know the quantiles of the normal distribution (numerically) $q(1 - 0.05/2) \simeq 1.96$
- Following the CLT and approximation of the Gaussian quantiles

$$\mathbb{P} \left[-1.96 < \sqrt{n} \, \frac{0.55 - p}{\widehat{\sigma}} < 1.96 \right] \approx 0.95$$

new CI: $\widehat{I} = [0.52, 0.58]$: better! (**more optimistic**)

In Python

Data generation

```
import numpy as np
from scipy.stats import norm

n = 1000
x = np.random.binomial(1, .5, n)
```

Computation of the CI

```
pchap = np.mean(x)
sig = np.sqrt(pchap * (1 - pchap))
alpha = .05
q = norm.ppf(1 - alpha/2)
borneinf = pchap - sig * q / np.sqrt(n)
bornesup = pchap - sig * (1 - q) / np.sqrt(n)
print('IC = [' + str(borneinf) +
      ',' + str(bornesup) + ']')
```

1. Confidence interval: reminder
   - Definition
   - Limit theorems

2. Confidence intervals for linear model
   - CI for the linear model under the Gaussian assumption
   - CI for the non-Gaussian case

Gaussian model

Proposition

Under model with Gaussian noise, whenever the matrix X has full rank, we have

(i) $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ are independent random variables

(ii) $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim \mathcal{N}(0, \sigma^2(X^\top X/n)^{-1})$ for every n

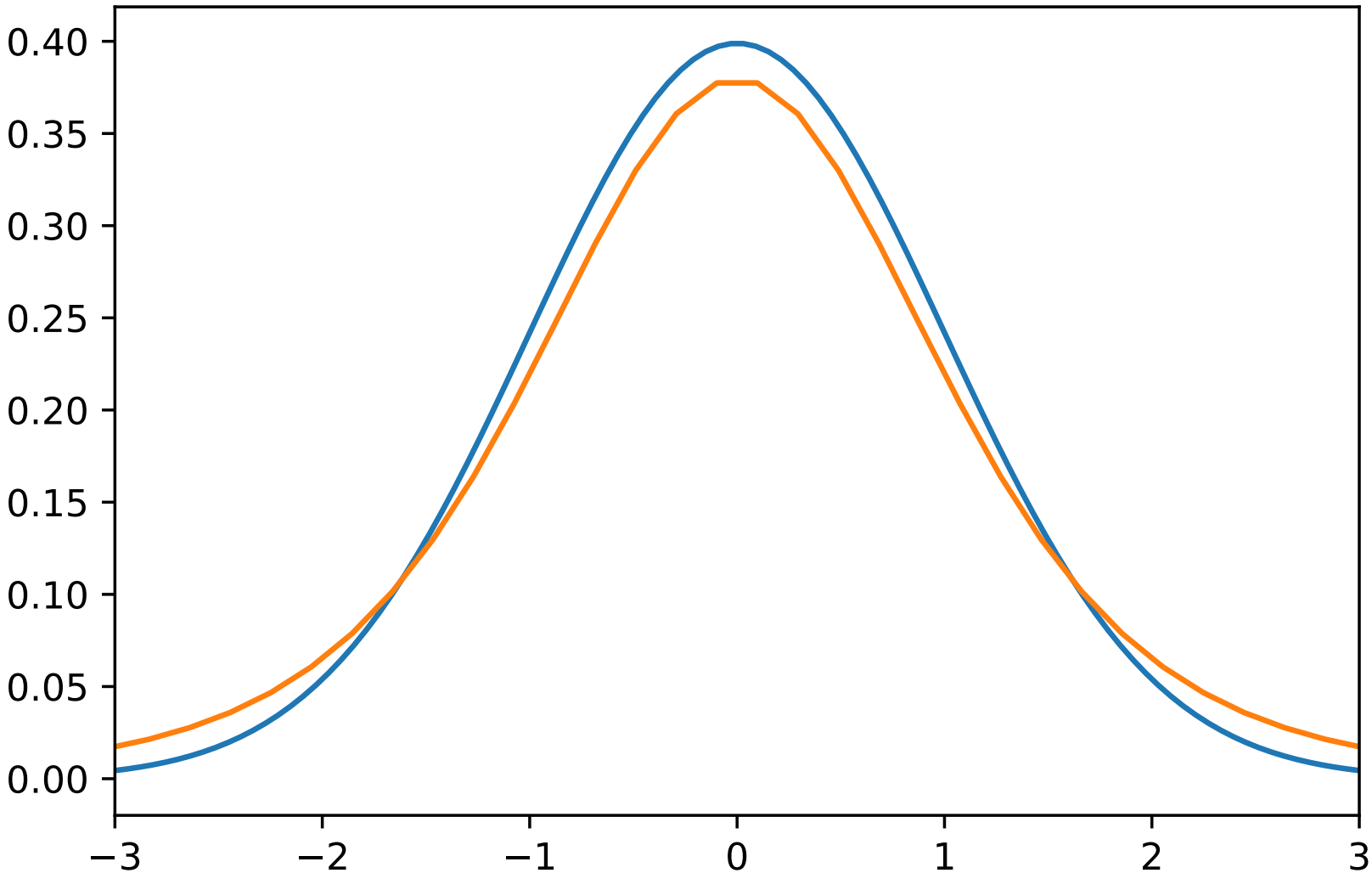
(iii) $(n - rank(X))\frac{\widehat{\sigma}^2}{\sigma^{*2}} \sim \chi^2_{n-rank(X)}$ for every n

(iv) Let $\widehat{s}_k = (X^\top X/n)^{-1}_{k,k}$,

$$\sqrt{n}\left(\frac{\widehat{\theta} - \theta^*}{\sqrt{\widehat{s}_k \widehat{\sigma}^2}}\right) \sim \mathcal{T}_{n-rank(X)}$$

where $\mathcal{T}_{n-rank(X)}$ stands for a student distribution with $n - rank(X)$ degrees of freedom

Gaussian vs t-student



CI for the regression coefficients (I)

Reminder: given $X \in \mathbb{R}^{n \times p}$, then $\widehat{\sigma}^2 = \|\mathsf{y} - X\widehat{\boldsymbol{\theta}}\|_2^2/(n - \mathsf{rang}(X))$ is the unbiased estimator for the variance. In addition (cf. Poly):

$$\text{If } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathsf{Id}_n), \text{ then } \boxed{T_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{[(X^\top X)^{-1}]_{j,j}}} \sim \mathcal{T}_{n-\mathsf{rang}(X)}}$$

where $\mathcal{T}_{n-\mathsf{rang}(X)}$ is a Student-t distribution with $n - \mathsf{rang}(X)$ degrees of freedom.

Its density, quantiles, etc..., can be computed numerically and are accessible in any software.

CI for the regression coefficients (II)

Under the Gaussian assumption, since

$$T_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-\mathsf{rang}(X)}$$

and noting $t_{1-\alpha/2}$ a quantile of order $1-\alpha/2$ of the distribution $\mathcal{T}_{n-\mathsf{rang}(X)}$,

$$\left[\widehat{\theta}_j - t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}, \widehat{\theta}_j + t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

for the quantity θ_j^*.

Rem: $\mathbb{P}(|T_j| < t_{1-\alpha/2}) = 1-\alpha$ since the Student-t distribution is symmetric.

CI for the predicted values

Now we would like to construct a CI for the predicted value at a single (new) given point $x = (1, x_1, ..., x_p)^\top \in \mathbb{R}^{p+1}$.

The predicted value at x (under the true model) is defined as

$$y^* = x^\top \boldsymbol{\theta}^* .$$

Under the Gaussian assumption, with the same notation, the following confidence interval is of level $1 - \alpha$

$$\left[x^\top \widehat{\boldsymbol{\theta}} - t_{1-\alpha/2} \widehat{\sigma} \sqrt{x^\top (X^\top X)^{-1} x}, x^\top \widehat{\boldsymbol{\theta}} + t_{1-\alpha/2} \widehat{\sigma} \sqrt{x^\top (X^\top X)^{-1} x} \right]$$

for the quantity y^*.

CI for the new predicted values

The CI from above is for the regression hyperplane, i.e. it is reflecting uncertainty of the fitted values.

How to build a CI for a new predicted predicted value at a single (new) given point $x = (1, x_1, ..., x_p)^\top \in \mathbb{R}^{p+1}$?

A new predicted value at x (under the true model) is defined as

$$y = y^* + \epsilon\,.$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

One can show that, in this case, and with the same notation, the following confidence interval is of level α

$$\left[x^\top \widehat{\boldsymbol{\theta}} - t_{1-\alpha/2}\widehat{\sigma}\sqrt{1 + x^\top (X^\top X)^{-1} x}, x^\top \widehat{\boldsymbol{\theta}} + t_{1-\alpha/2}\widehat{\sigma}\sqrt{1 + x^\top (X^\top X)^{-1} x}\right]$$

for the quantity y.

Example: Investment data (I)

Following classical example follows Greene (2012).

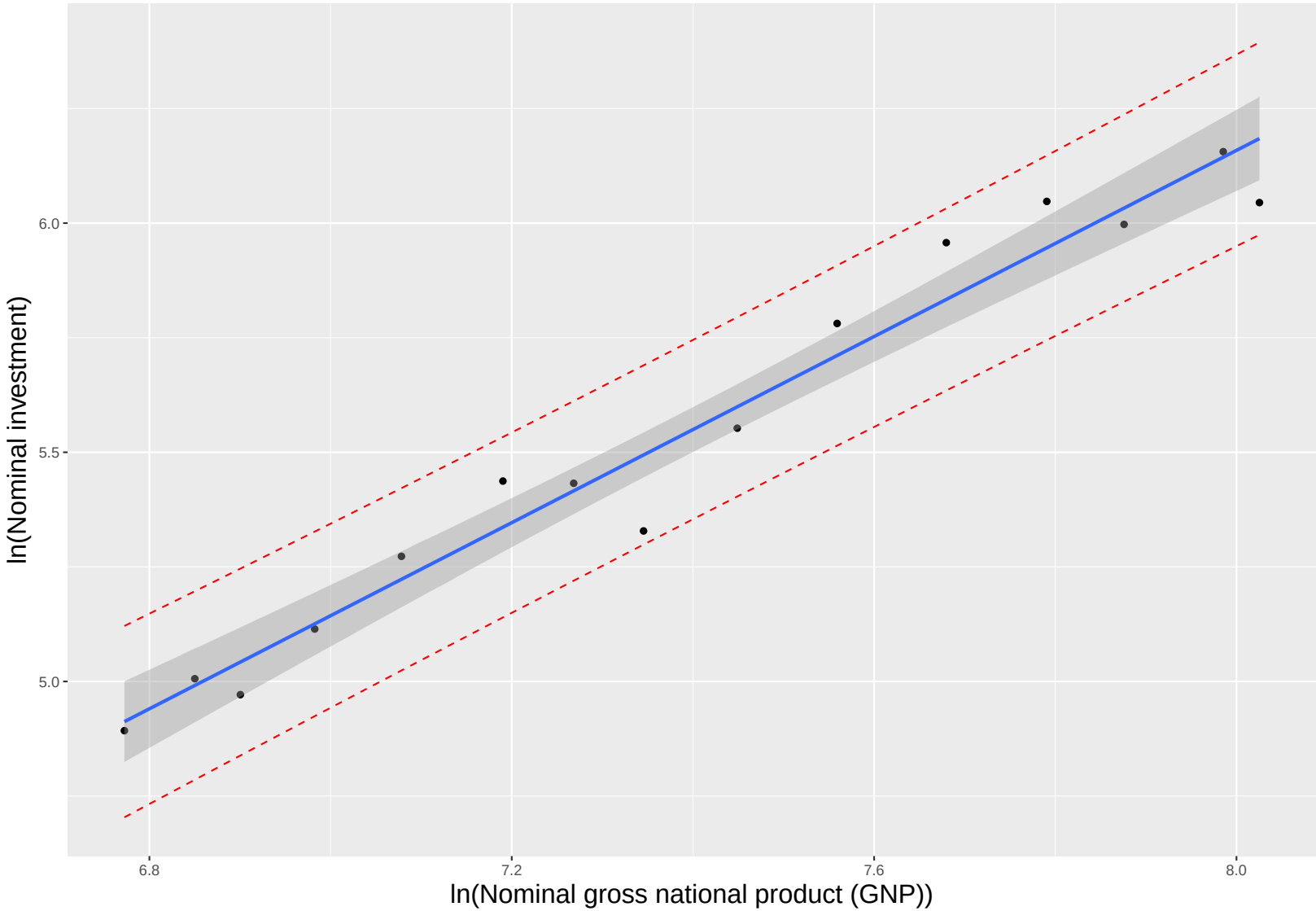
Regard the the data of the nominal investment and nominal gross national product records for 15 years.

First, we apply the logarithmic transform to both variables.

Then, we try to explain (the logarithm of) the nominal investment Y by (the logarithm of) the nominal gross national product X.

We also would like to provide inference about the estimated coefficient as well as the newly predicted values, at a 95% level.

Example: Investment data (II).



Exercise

How to use Bootstrap to simulate confidence intervals?

CI when no Gaussian assumption (I)

Regard now the random design model, where the aim is to estimate the best linear approximation of Y_1 made up with X_1 in terms of L_2-risk.

In this case we can still provide a CLT.

Reminder: Suppose that $\mathbb{E}[X_1 X_1^\top]$ and $\mathbb{E}[Y_1^2]$ exist and that $\mathbb{E}[X_1 X_1^\top]$ is invertible.

$$\boxed{T_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{[(X^\top X)^{-1}]_{j,j}}} \rightsquigarrow \mathcal{N}(0,1)}$$

This can be directly used to construct the CI.

CI when no Gaussian assumption (II)

If no normality of the noise term ϵ is assumed, the confidence interval from above is not reliable.

It remains valid asymptotically if we assume $\mathbb{E}[X_{1,j}^2] < \infty$ and $\mathbb{E}[Y_1^2] < \infty$ though, which is indicated by the following result.

The following confidence interval is of level $1 - \alpha$

$$\widehat{I}_j(1-\alpha) = \left[\widehat{\theta}_j - q_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}, \widehat{\theta}_j + q_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

asymptotically for the quantity θ_j^* (with q_α being the α-quantile of $\mathcal{N}(0,1)$), i.e.

$$\liminf_{n\to\infty} \mathbb{P}\big(\theta_j^* \in \widehat{I}_j(1-\alpha)\big) \geq 1 - \alpha\,.$$

Limitations of the CI previously seen

In the previous part, the reasoning is based on the Gaussian assumption, even if using **asymptotic approximation**.

Attention: if the model is (very) wrong or the sample is very small, the obtained CIs will not necessarily be valid.

Possible alternative: *bootstrap*, a non-parametric method based on the resampling, well suited (theoretically) for regular statistics like the average, the quantiles, etc... (but not for max or min!).

For further material see: Efron and Tibshirani (1994)

References I

- Some of these slides have been prepared by Josef Salmon, the authors express their gratitude for this.