MAP 531: Risk and confidence regions

28 septembre 2017

Bias and transformation

- Applying a non linear transformation to an unbiased estimator does not produce in general an unbiased estimator, i.e. if T is an unbiased estimator of θ, $g(T)$ is not in general an unbiased estimator of $g(\theta)$.
- For example, of T is an unbiased estimator of θ, then T^2 is not an unbiased estimator of θ^2. Indeed,

$$\mathbb{E}_\theta[T^2] = \mathrm{Var}_\theta(T) + (\mathbb{E}_\theta[T])^2 = \mathrm{Var}_\theta(T) + \theta^2$$

- Therefore, unless T is concentrated on a unique point (! ! !), $\mathrm{Var}_\theta(T) > 0$ and T^2 is positively biased, for all $\theta \in \Theta$,

$$\mathbb{E}_\theta[T^2] - \theta^2 > 0$$

.

- On the other hand, if T^2 is an unbiased estimator of θ^2, then

$$|\mathbb{E}_\theta[T]| < |\theta|.$$

Estimation of a uniform distribution support

Let $X_1, X_2, \dots, X_n$ be independent random variables which follow $\mathrm{Unif}([0,\theta])$, with $\theta \in \Theta = \mathbb{R}_+^*$. We denote$X_{n:n} = \max(X_1, \dots, X_n)$.
For all $\theta \in \Theta$ et $x \in [0,\theta]$, we have

$$\mathbb{P}_\theta(X_{n:n} \le x) = \mathbb{P}_\theta(\max(X_1, \dots, X_n) \le x) = \prod_{i=1}^n \mathbb{P}_\theta(X_i \le x) = (x/\theta)^n \;.$$

The p.d.f. of $X_{n:n}$ is therefore given for $x \in [0,\ \theta]$, by

$$n\frac{x^{n-1}}{\theta^n}$$

This yileds :

$$\mathbb{E}_\theta[X_{n:n}] = \int_0^\theta x.n\frac{x^{n-1}}{\theta^n}\mathrm{d}x = \frac{n}{n+1}\frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta \;,$$
$$\mathbb{E}_\theta[X_{n:n}^2] = \int_0^\theta x^2.n\frac{x^{n-1}}{\theta^n}\mathrm{d}x = \frac{n}{n+2}\frac{\theta^{n+2}}{\theta^n} = \frac{n}{n+2}\theta^2 \;.$$

Estimation of a uniform distribution support

Let $X_1, X_2, \ldots, X_n$ be independent random variables which follow $\mathrm{Unif}([0, \theta])$, with $\theta \in \Theta = \mathbb{R}_+^*$. We denote$X_{n:n} = \max(X_1, \ldots, X_n)$.

$$\mathbb{E}_\theta[X_{n:n}] = \frac{n}{n+1}\theta \ , \quad \mathbb{E}_\theta[X_{n:n}^2] = \frac{n}{n+2}\theta^2 \ .$$

The estimator $(n+1)/nX_{n:n}$ is an unbias estimator of the parameter θ. The quadratic risk of $a_n X_{n:n}$ is

$$\begin{aligned}
\mathbb{E}_\theta[(a_n X_{n:n} - \theta)^2] &= a_n^2 \mathbb{E}_\theta[X_{n:n}^2] - 2a_n\theta\mathbb{E}_\theta[X_{n:n}] + \theta^2 \\
&= \frac{na_n^2}{n+2}\theta^2 - \frac{2a_n n}{n+1}\theta^2 + \theta^2 = \theta^2\left\{\frac{na_n^2}{n+2} - \frac{2a_n n}{n+1} + 1\right\}
\end{aligned}$$

This quadratic risk is minimum if we take $a_n = (n+2)/(n+1)$ and the minimum of the risk is

$$\mathbb{E}_\theta\left[\left(\frac{n+2}{n+1}X_{n:n} - \theta\right)^2\right] = \frac{\theta^2}{(n+1)^2}$$

Estimation of a uniform distribution support

Let $X_1, X_2, \ldots, X_n$ be independent random variables which follow $\mathrm{Unif}([0,\theta])$, with $\theta \in \Theta = \mathbb{R}_+^*$. We denote $X_{n:n} = \max(X_1, \ldots, X_n)$. The estimator $(n+1)/nX_{n:n}$ is unbiased for θ but for all $\theta \in \Theta$

$$\mathbb{E}_\theta\left[\left(\frac{n+2}{n+1}X_{n:n} - \theta\right)^2\right] \le \mathbb{E}_\theta\left[\left(\frac{n+1}{n}X_{n:n} - \theta\right)^2\right]$$

This shows that $(n+1)/nX_{n:n}$ is inadmissible for the quadratic risk.

Confidence intervalle for a proportion

- Let $X_1, \ldots, X_n$ be independent random variables following a Bernoulli distribution with parameter $\theta \in \Theta = [0, 1]$. If $\bar{X}_n \triangleq n^{-1} \sum_{i=1}^n X_i$ is the average number of succes,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(\bar{X}_n - \theta)^2 \right] = \sup_{\theta \in \Theta} \frac{\theta(1-\theta)}{n} = \frac{1}{4n} \, .$$

Confidence intervalle for a proportion

- Let $X_1, \ldots, X_n$ be independent random variables following a Bernoulli distribution with parameter $\theta \in \Theta = [0,1]$. If $\bar{X}_n \triangleq n^{-1} \sum_{i=1}^n X_i$ is the average number of succes,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(\bar{X}_n - \theta)^2 \right] = \sup_{\theta \in \Theta} \frac{\theta(1-\theta)}{n} = \frac{1}{4n} .$$

- For all $\delta > 0$ et $\theta \in \Theta$, by leariry of Bienayme-Tchebychev

$$\mathbb{P}_\theta \left(|\bar{X}_n - \theta| \geq \delta \right) \leq \delta^{-2} \operatorname{Var}_\theta \left(\bar{X}_n \right) \leq \frac{1}{4n\delta^2} .$$

Confidence intervalle for a proportion

- Let $X_1, \ldots, X_n$ be independent random variables following a Bernoulli distribution with parameter $\theta \in \Theta = [0, 1]$. If $\bar{X}_n \triangleq n^{-1} \sum_{i=1}^{n} X_i$ is the average number of succes,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(\bar{X}_n - \theta)^2 \right] = \sup_{\theta \in \Theta} \frac{\theta(1-\theta)}{n} = \frac{1}{4n} \; .$$

- For all $\delta > 0$ et $\theta \in \Theta$, by lineariry of Bienayme-Tchebychev

$$\mathbb{P}_\theta \left(|\bar{X}_n - \theta| \geq \delta \right) \leq \delta^{-2} \operatorname{Var}_\theta \left(\bar{X}_n \right) \leq \frac{1}{4n\delta^2} \; .$$

- For $\alpha \in (0, 1)$, we set $\delta_{n,\alpha} \triangleq 1/2\sqrt{n\alpha}$. For all $\theta \in \Theta$,

$$\mathbb{P}_\theta \left(\theta \in \mathcal{I}_{n,\alpha} \right) \geq 1 - \alpha \; , \quad \text{où} \quad \mathcal{I}_{n,\alpha} = \left[\bar{X}_n \pm \frac{1}{2\sqrt{n\alpha}} \right] \; .$$

Confidence intervalle for a proportion

- The quality of this intervalle is measured with its length, $|\mathcal{I}_{n,\alpha}|$, which here equals
$$|\mathcal{I}_{n,\alpha}| = \frac{1}{\sqrt{n\alpha}} \, .$$
- When $\alpha \to 0$ we have $|\mathcal{I}_{n,\alpha}| \to +\infty$.
- On can refine this intervalle using exponential inequalities.

Hoeffding inequality

Théorème

Let $Y_1, \ldots, Y_n$ *be real valued independent random such that for all* $i \in \{1, \ldots, n\}$, $a_i \leq Y_i \leq b_i$. *For all* $t > 0$, *we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mathbb{E}[Y_i]\} \geq t\right) \leq \mathrm{e}^{-2n^2t^2/\tau_n^2} ,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mathbb{E}[Y_i]\} \leq -t\right) \leq \mathrm{e}^{-2n^2t^2/\tau_n^2} .$$

with $\tau_n^2 = \sum_{i=1}^n (b_i - a_i)^2$.

Note that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mathbb{E}[Y_i]\}\right| \geq t\right) \leq 2\mathrm{e}^{-2n^2t^2/\tau_n^2}$$

Confidence intervalle for a proportion

- Let $\delta > 0$,

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\left| \bar{X}_n - \theta \right| > \delta \right) \leq 2 \exp\left(-2n\delta^2 \right) .$$

- Let us take $\delta = \delta(\alpha, n)$ the solution of $2\exp(-2n\delta^2) = \alpha$, we define, for all $\alpha > 0$,

$$\mathcal{I}^\star_{n,\alpha} = \left[\bar{X}_n \pm \delta_{n,\alpha} \right] = \left[\bar{X}_n \pm \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right] .$$

- By construnction $\mathcal{I}^\star_{n,\alpha}$ is a confidence intervalle for θ at level $1-\alpha$. We have

$$|\mathcal{I}^\star_{n,\alpha}| / |\mathcal{I}_{n,\alpha}| = \sqrt{2\alpha \log(2/\alpha)} \to 0 \quad \text{when} \quad \alpha \to 0 .$$

  For $\alpha = 0.01$, $|\mathcal{I}^\star_{n,\alpha}| / |\mathcal{I}_{n,\alpha}| \approx 0.33$, that is to say a precision multiplied by 3...
- If $n = 1000$, , $|\mathcal{I}^\star_{n,\alpha}| = 0.04$ for $\alpha = 0.05$ and 0.05 for $\alpha = 0.01$...

Pivotal function

Définition

Let $\{(\Omega, \mathcal{F}), (\mathsf{Z}, \mathcal{Z}), \{\mathbb{P}_\theta\,,\ \theta \in \Theta\}\,, Z\}$ be a statistical model where $\Theta \in \mathcal{B}(\mathbb{R}^d)$. We say that a measurable function

$$
\begin{aligned}
G : (\mathsf{Z} \times \Theta, \mathcal{Z} \otimes \mathcal{B}(\Theta)) &\to (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)) \\
(z, \theta) &\mapsto G(z, \theta)
\end{aligned}
$$

is pivotal if, for all $\theta \in \Theta$, the function $z \mapsto G(z, \theta)$ is measurable and, for all $\theta \in \Theta$, the probability distribution $G(Z, \theta)$ does not depend on θ, i.e. for all $\theta, \vartheta \in \Theta$ and $A \in \mathcal{B}(\mathbb{R}^p)$,

$$
\mathbb{P}_\theta\left(G(Z, \theta) \in A\right) = \mathbb{P}_\vartheta\left(G(Z, \vartheta) \in A\right)
$$

Pivotal functions and confidence intervalle

- A pivotal function enables to construct confidence intervalles $\mathcal{C}(Z)$ for θ at a given level $(1-\alpha) \in (0,1)$.
- Let $A_\alpha \in \mathcal{B}(\mathbb{R}^p)$ be such that

$$\mathbb{P}_\theta\left(G(Z,\theta) \in A_\alpha\right) \geq 1-\alpha, \quad \text{for all} \quad \theta \in \Theta\,.$$

The left hand side of the previous inequality does not depend on $\theta \in \Theta$.

- For any $A_\alpha \in \mathcal{B}(\mathbb{R}^p)$ as such, the confidence region defined by

$$\mathcal{C}(Z) \triangleq \{\theta \in \Theta \,:\, G(Z,\theta) \in A_\alpha\}$$

is a confidence region at level $1-\alpha$.

Confiance region for the mean

- Let $X_1, \dots, X_n$ n be independent random variables following a Gaussienne distribution with unknown mean μ and known variance σ^2

Confiance region for the mean

- Let $X_1, \ldots, X_n$ n be independent random variables following a Gaussienne distribution with unknown mean μ and known variance σ^2
- The function

$$G(X_1, \ldots, X_n; \mu) = \frac{n^{-1/2} \sum_{i=1}^{n} (X_i - \mu)}{\sigma}$$

is pivotal : for all $\mu \in \mathbb{R}$, $G(X_1, \ldots, X_n; \mu)$ follows a standard normal distribution if $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, i.e. for all $\mu \in \mathbb{R}$ and $a < b$,

$$\mathbb{P}_\mu \left(G(X_1, \ldots, X_n; \mu) \in [a, b] \right) = \Phi(b) - \Phi(a)$$

where Φ is the cumulative distribution function of the standard normal distribution.

Confiance region for the mean

- Let $X_1, \ldots, X_n$ n be independent random variables following a Gaussienne distribution with unknown mean μ and known variance σ^2
- For $\beta \in (0,1)$, we set z_β the quantile of order β of $\mathrm{N}(0,1)$: $\Phi(z_\beta) = \beta$. As $z_{\alpha/2} = -z_{1-\alpha/2}$,

$$\begin{aligned}\mathbb{P}_\mu\left(G(X_1, \ldots, X_n; \mu) \in \left[-z_{1-\alpha/2}, z_{1-\alpha/2}\right]\right) &= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) \\ &= 1 - \alpha \,.\end{aligned}$$

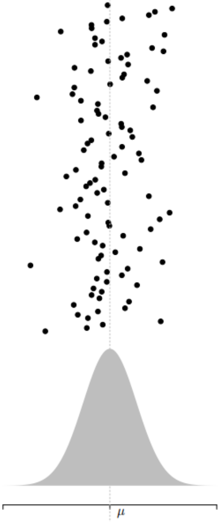
Confiance region for the mean

- Let $X_1, \ldots, X_n$ n be independent random variables following a Gaussienne distribution with unknown mean μ and known variance σ^2
- Confidence region : For all $\mu \in \mathbb{R}$,

$$\mathbb{P}_\mu \left(\mu \in \left[\bar{X}_n \pm \sigma z_{1-\alpha/2}/\sqrt{n}\right]\right) = 1 - \alpha \; .$$

Example

Each point represents the result of one experiement : we generate a sample from the standard normal and we evaluate its empirical mean.
We have represented 100 points resulting from 100 experiments.

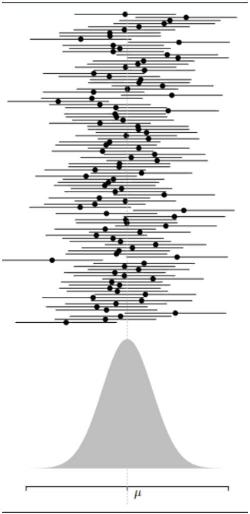


Example

For each experiment, we determine the confidence region at level $1-\alpha$ with $\alpha = 0.05$

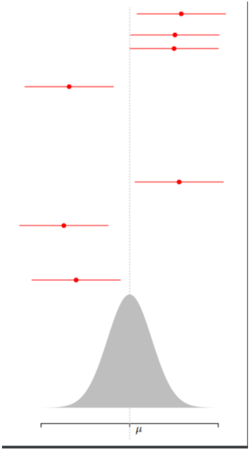
$$\bar{X}_n \pm 1.96\sigma/\sqrt{n}$$

Ẁhat do you expect ?



Example

Among the 100 experiments, 7 regions we have constructed do not contain the true mean μ. This is essential in the notion of confidence : we construct regions that contains the true value of the parameter with probability $1-\alpha$



A small "bug"...

- The construction is only partialy satisfying, as the situations where we know σ are not usual. Although we focus only on the mean, it is compulsory to have an estimated value of the variance.
- We have seen that we can estimate σ^2 either with the MLE

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

or with its unbiased version

$$\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

- Question : How do we adapt the construction of the region to take this into account ?

We compare here on 1000 experiements the distribution of

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

and

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n}$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and

$$S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

What do you observe ?