Confidence regions

the notion of confidence regions is another way to quantify the precision of an estimation.

I Definition and properties

def: Let $\{(\Omega, \mathcal{F}), (\mathcal{X}, \mathcal{Z}), \{P_\theta, \theta \in \Theta\}, Z\}$ be a statistical model. Let $g : \Theta \to \mathbb{R}^p$ and $\alpha \in]0,1[$. A function $\mathcal{C} : \mathcal{Z} \to \mathcal{P}(\mathbb{R}^p)$ is called a confidence region at level $1-\alpha$ for $g(\theta)$ if $\forall \theta \in \Theta$: $\{z \in \mathcal{Z} : g(\theta) \in \mathcal{C}(z)\} \subset \mathcal{Z}$ and

$$\inf_{\theta \in \Theta} P_\theta\left(\{z \in \mathcal{Z} : g(\theta) \in \mathcal{C}(z)\}\right) \geq 1-\alpha$$

If $p=1$ then we talk about confidence intervals, and they take one of the 3 following forms:

(i) $\mathcal{C}(z) = [m(z) ; +\infty[$
(ii) $\mathcal{C}(z) =]-\infty ; M(z)]$
(iii) $\mathcal{C}(z) = [m(z) ; M(z)]$

Example: let us go back to the survey.

Il faudra avoir fait l'exo du chap d'avant avant de faire ce cours.

Let $X_1, \dots, X_n$ iid $\mathcal{B}(\theta)$

We have seen that with $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ (the average number of success)

$$\sup_{\theta \in \Theta} E_\theta\left[(\hat{\theta}_n - \theta)^2\right] = \frac{1}{4n}$$

which writes also $\sup_{\theta \in \Theta} E_\theta\left[\sqrt{n}(\hat{\theta}_n - \theta)^2\right] = \frac{1}{4}$

Lemme de Bienaymé-Tchebychev:

Soit Z une v.a. tq $E[Z^2] < \infty$. Alors $\forall \delta > 0$:

$$P(|Z - E(Z)| > \delta) \leq \frac{Var(Z)}{\delta^2}$$

We now construct a confidence interval.

Let $\delta > 0$ and $\theta \in \Theta$

thanks to the Bienaymé-Tchebychev inequality

$$P_\theta\left(|\hat{\theta}_n - \theta| \geq \delta\right) \leq \frac{1}{\delta^2} Var_\theta(\hat{\theta}_n) \leq \frac{1}{4n\delta^2}$$

If we choose $\delta_{n,\alpha}$ such that $\frac{1}{4n\delta_{n,\alpha}^2} = \alpha$ ie $\delta_{n,\alpha} = \frac{1}{2\sqrt{n\alpha}}$ and write

$$I_{n,\alpha} = \left[\hat{\theta}_n - \frac{1}{2\sqrt{n\alpha}} ; \hat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}\right] \text{ then :}$$

$$\forall \theta \in \Theta \qquad P_\theta(\theta \in I_{n,\alpha}) \geq 1-\alpha$$

the quality of this interval is given by its length : $|I_{n,\alpha}| = \frac{1}{\sqrt{n\alpha}}$

Note that $|I_{n,\alpha}| \to +\infty$ when $\alpha \to 0$. This is a tradeoff that one can easily understand. Indeed, one has to choose between the precision of the estimation ($|I_{n,\alpha}|$ small) and the risk we accept (α small)

II Classical confidence regions:

Let us first introduce some classical distributions.

1) χ^2_k : the Chi-square with k degrees of freedom:

def: Let $(X_1, \dots, X_k)$ k iid random variables such that $X_1 \sim \mathcal{N}(0,1)$. then $U = \sum_{i=1}^{k} X_i^2$ follows a centered χ^2 distribution with k degrees of freedom.

Prop:
- the pdf of U is $f_k(x) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)} x^{k/2-1} \exp\left(-\frac{x^2}{2}\right) \mathbb{1}_{\{x>0\}}$
- $\mathbb{E}[U] = k \qquad Var(U) = 2k$

def: If $X_i \sim \mathcal{N}(\mu_i, 1)$ then $U = \sum_{i=1}^{k} X_i$ follows a non-centered χ^2 distribution with k degrees of freedom denoted by $\chi^2_k(\gamma)$ where $\gamma = \frac{1}{2}\sum_{i=1}^{k} \mu_i^2$

In this case $\mathbb{E}(U) = \gamma + k$ and $Var(U) = 2k + 4\gamma$

2) Student distribution

def: Let X and Y be 2 independent r.v such that:
- $X \sim \mathcal{N}(0,1)$
- $Y \sim \chi^2_n$ (centered)

then $T = \frac{X}{\sqrt{Y/n}} \sim t(n)$ a Student distribution with n degrees of freedom.

Prop: $f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{(n\pi)^{1/2}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$

theorem: Let $X_1, \dots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ then:

(i) $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (the empirical mean) and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ the variance of the sample are independent

(ii) $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

(iii) $(n-1)\,S_n^2 \sim \chi^2_{n-1}$

(iii) $T_n = \sqrt{n}\,\dfrac{\bar{X}_n - \mu}{S_n^2} \sim t(n-1)$

3) Fisher's distribution:

def: Let X, Y 2 independent r.v such that

- $X \sim \chi^2_q$
- $Y \sim \chi^2_r$

then $W = \dfrac{X/q}{Y/r}$ ~ $F(q,r)$ a fisher's distribution with $(q;r)$ degrees of freedom

Prop: $f_{(q,r)}(w) = \dfrac{\Gamma\left(\frac{q+r}{2}\right)}{\Gamma\left(\frac{q}{2}\right)\Gamma\left(\frac{r}{2}\right)} \left(\frac{q}{r}\right)^{q/2} \dfrac{w^{\frac{q}{2}-1}}{\left(1+\left(\frac{q}{r}\right)w\right)^{\frac{q+r}{2}}} \mathbb{1}_{\{w>0\}}$

4) Confidence interval for a normal sample with known variance:

Let $(X_1, \dots, X_n)$ i.i.d samples of $N(\mu, \sigma^2)$, with σ^2 known.

We want to construct a confidence interval for μ of the form $[m(X_1,\dots,X_n); M(X_1,\dots,X_n)]$ of level $1-\alpha$: ie: we aim at finding m and n such that

$$\mathbb{P}_\mu\left(\mu \in [m(X_1,\dots,X_n); \Pi(X_1,\dots,X_n)]\right) = 1-\alpha$$

For all $\mu \in \mathbb{R}$ $G(X_1,\dots,X_n,\mu) = \dfrac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \sim \mathcal{N}(0,1)$ thanks to the central limit theorem.

Let Φ the cumulative distribution function of $\mathcal{N}(0,1)$ distribution. For all $\alpha \in]0,1[$, let z_α be the α-quantile of Φ given by $\Phi(z_\alpha) = \alpha$

In particular $z_{1-\frac{\alpha}{2}} = 1{,}96$ for $\alpha = 0{,}05$ and $z_{1-\frac{\alpha}{2}} = 3$ for $\alpha = 0{,}01$

This yields: $\mathbb{P}_\mu\left(\dfrac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \in [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]\right) = 1-\alpha$

Note that $I_\alpha = [-z_{1-\frac{\alpha}{2}}; z_{1-\frac{\alpha}{2}}]$ is independent of μ.

therefore $I_{\alpha,n} = \left[\bar{X}_n - \dfrac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} ; \bar{X}_n + \dfrac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$ is the confidence interval for μ and $|I_{\alpha,n}| \xrightarrow[n\to+\infty]{} 0$

5) Confidence interval for μ with unknown σ^2

In the same setting as above, one can consider

$$G(X_1, \dots, X_n, \mu) = \frac{\sqrt{n}(\overline{X_n} - \mu)}{S_n^2} \quad \text{where} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

thanks to the previous theorem $G(X_1, \dots, X_n, \mu) \sim t_{n-1}$

therefore $\forall A \subset \mathbb{R}$ and $\theta = (\mu, \sigma^2)$

$$P_\theta\left(G(X_1, \dots, X_n, \mu) \in A\right) = t_{n-1}(A)$$

Let $t_\alpha(n-1)$ the α quantile of t_{n-1}. Thanks to the symmetry of the Student's distribution, confidence intervals are symmetric as well and this writes

$$P_\theta\left(-t_{1-\frac{\alpha}{2}}(n-1) \le G(X_1, \dots, X_n, \mu) \le t_{1-\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

this leads to the following interval for μ:

$$I(X_1, \dots, X_n) = \overline{X_n} - \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1) \quad \text{and} \quad I(X_1, \dots, X_n) = \overline{X_n} + \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$$

Remark: If the sample is small $n \le 100$ then it is important to consider the Student's distribution. However as soon as $n \ge 100$ then the Student's distribution is very close to a normal one and one can use the normal quantiles.

6) Confidence Interval for the variance:

Let again $(X_1, \dots, X_n)$ iid $\mathcal{P}(\mu, \sigma^2)$ with unknown μ and σ^2.

Let: $G(X_1, \dots, X_n, \sigma^2) = \frac{(n-1)}{\sigma^2} S_n^2$

We know that $\frac{(n-1)}{\sigma^2} S_n^2 \sim \chi^2_{n-1}$. Let $\chi^2_\alpha(n-1)$ the α quantile of χ^2_{n-1} and fix $\alpha_1 + \alpha_2 = \alpha$ then: $\forall (\mu, \sigma^2)$

$$P_{(\mu,\sigma)}\left(\chi^2_{\alpha_1}(n-1) \le G(X_1, \dots, X_n, \sigma^2) \le \chi^2_{1-\alpha_2}(n-1)\right) = 1-\alpha$$

this yields $I_{\alpha,n} = \left[\frac{n-1}{\chi^2_{1-\alpha_2}(n-1)} S_n^2 \; ; \; \frac{n-1}{\chi^2_{\alpha_1}(n-1)} S_n^2\right]$ is a confidence interval of level $1-\alpha$ for σ^2

Note that the length of this interval is random as it depends on S_n^2

7) Confidence region for both mean and variance

Again let $X_1, \dots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$. We want a confidence region for both μ and σ^2 of level $1-\alpha$.

We have seen previously that

$$I_1(X_1, \dots, X_n) = \left[\bar{X}_n - S_n \frac{t_{1-\frac{\alpha}{2}}(n-1)}{\sqrt{n}} \; ; \; \bar{X}_n + S_n \frac{t_{1-\frac{\alpha}{2}}(n-1)}{\sqrt{n}} \right]$$

is a confidence interval for μ of level $1-\frac{\alpha}{2}$

Moreover $I_2(X_1, \dots, X_n) = \left[\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{4}}(n-1)} \; ; \; \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{4}}(n-1)} \right]$ is a confidence interval for σ^2 at level $1-\frac{\alpha}{2}$. Therefore

$$\begin{aligned} \mathbb{P}_{(\mu,\sigma^2)}\left((\mu,\sigma^2) \in I_1(X_1, \dots, X_n) \times I_2(X_1, \dots, X_n) \right) &\geq 1 - \mathbb{P}_{(\mu,\sigma^2)}\left(\mu \notin I_1(X_1, \dots, X_n) \right) \\ &\quad - \mathbb{P}_{(\mu,\sigma^2)}\left(\sigma^2 \notin I_2(X_1, \dots, X_n) \right) \\ &= 1-\alpha \end{aligned}$$

$I_1 \times I_2$ defines the confidence region of (μ, σ^2) at level $1-\alpha$.