Statistics
MDI220
6. Confidence intervals

Thomas Bonald
Institut Polytechnique de Paris

2022 – 2023

Consider a parametric model with unknown parameter $\theta \in \Theta$. A point estimator $\hat{\theta}(x)$ provides a single value for θ, given some observation x. There is no information about the quality of this estimation. In this lecture, we will see how to provide a confidence region (or interval) for θ, that is a subset of Θ instead of a single value.

1 Confidence region

The objective is to find some decision function δ that returns a subset of Θ for each observation.

The decision function δ is a mapping from the set of observations $\mathcal{X}$ to subsets of Θ:

$$\forall x, \quad \delta(x) \subset \Theta.$$

Informally, we would like to ensure that $\theta \in \delta(x)$ with high probability. More formally, we will focus on $\mathrm{P}_\theta(\theta \in \delta(X))$, for each parameter θ. Note that the randomness comes from the observation X and *not* from the parameter θ.

Confidence region

We say that δ provides a confidence region at level $1-\alpha$ if

$$\forall \theta \in \Theta, \quad P_\theta(\theta \in \delta(X)) \geq 1 - \alpha.$$

When $\theta \in \mathbb{R}$, the confidence region might be:

- a confidence interval, if $\delta(x) = [m(x), M(x)]$,
- a lower confidence bound, if $\delta(x) = [m(x), +\infty)$,
- an upper confidence bound, if $\delta(x) = (-\infty, M(x)]$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$. For n i.i.d. observations, the Maximum Likelihood Estimator (MLE) is:*

$$\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Typical confidence regions will be:

- *a confidence interval $\delta(x) = [m(x), M(x)]$, with $m(x) < \hat{\theta}(x) < M(x)$,*
- *a lower confidence bound (LCB) $\delta(x) = [m'(x), +\infty)$, with $\hat{\theta}(x) > m'(x)$,*
- *an upper confidence bound (UCB) $\delta(x) = (-\infty, M'(x)]$, with $\hat{\theta}(x) < M'(x)$.*

We expect the LCB to provide a more precise lower bound on θ than the confidence interval in the sense that

$$m'(x) > m(x).$$

Similarly, we expect the UCB to provide a more precise upper bound on θ than the confidence interval in the sense that

$$M'(x) < M(x).$$

2 Pivot function

Confidence regions can be constructed through pivot functions.

Pivot function

We say that φ_θ is a pivot function (from the set of observations) if the distribution of the random variable $\varphi_\theta(X)$ is independent of θ.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$. Let:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We have $\bar{X} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n})$, that is $\bar{X} \sim \frac{\sigma}{\sqrt{n}} Z + \theta$ with $Z \sim \mathcal{N}(0,1)$. A pivot function is:

$$\varphi_\theta(x) = \frac{\sqrt{n}}{\sigma}(\bar{x} - \theta) \quad \text{with } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Indeed, we have $\varphi_\theta(X) \sim \mathcal{N}(0,1)$.

To get a confidence region at level $1-\alpha$ using the pivot function φ_θ, let A be some set such that:

$$P(Z \in A) = 1 - \alpha,$$

with $Z = \varphi_\theta(X)$. Then:

$$\forall \theta, \quad P_\theta(\varphi_\theta(X) \in A) = 1 - \alpha,$$

so that a confidence region is:

$$\delta(x) = \{\theta \subset \Theta : \varphi_\theta(x) \in A\}.$$

Observe that this confidence region depends on the choice of A, which is not unique. In particular, we might obtain a confidence interval, a LCB or a UCB depending on the choice of A.

Pivot function → Confidence region

Let A be such that $P(\varphi_\theta(X) \in A) = 1 - \alpha$.
Then $\delta(x) = \{\theta : \varphi_\theta(x) \in A\}$ is a confidence region at level $1 - \alpha$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$. Let $c = Q(1 - \frac{\alpha}{2})$ with Q the quantile function of the normal distribution, $Z \sim \mathcal{N}(0, 1)$. We have:*

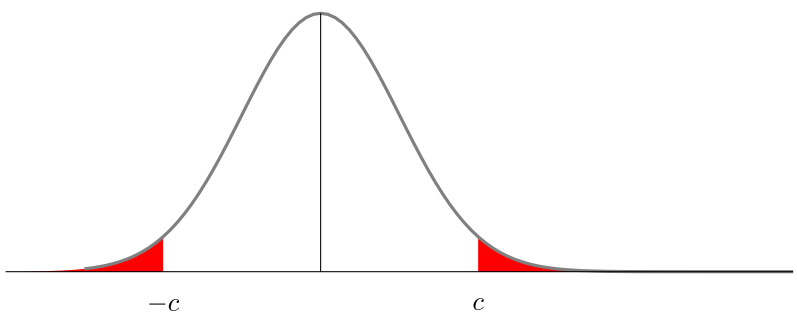
$$P(Z \in [-c, c]) = 1 - \alpha,$$

that is,

$$\forall \theta, \quad P_\theta\left(\theta \in \left[\bar{X} - \frac{c\sigma}{\sqrt{n}}, \bar{X} + \frac{c\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

The confidence interval is:

$$\delta(x) = \left[\bar{x} - \frac{c\sigma}{\sqrt{n}}, \bar{x} + \frac{c\sigma}{\sqrt{n}}\right], \quad \textit{with } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$



Now let $c' = Q(1 - \alpha)$. We have:

$$P(Z \leq c') = 1 - \alpha,$$

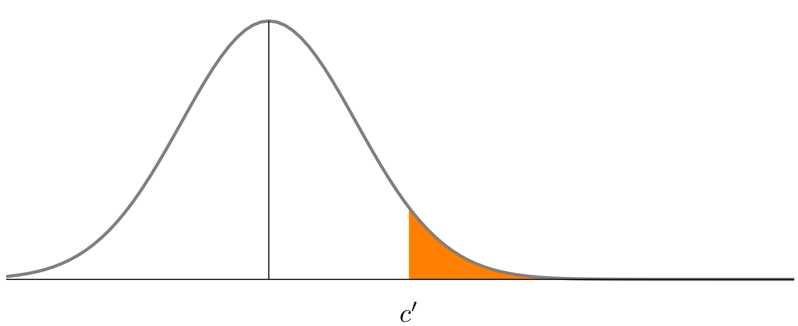
that is,

$$\forall \theta, \quad P_\theta\left(\theta \geq \bar{X} - \frac{c'\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

We get a lower confidence bound:

$$\delta(x) = \left[\bar{x} - \frac{c'\sigma}{\sqrt{n}}, +\infty\right).$$

Observe that $c' < c$ so that the lower bound is more precise than that of the confidence interval.



3 Link with hypothesis testing

The problems of confidence regions and hypothesis testing are dual. Consider the test of the null hypothesis $H_0 = \{\theta = \theta_0\}$ against the alternative hypothesis $H_1 = \{\theta \neq \theta_0\}$, for any given $\theta_0 \in \Theta$. Assume that for each $\theta_0 \in \Theta$, you have a test at level α, denoted by δ_{θ_0}. Then:

$$\forall \theta \in \Theta, \quad P_\theta(\delta_\theta(X) = 1) \leq \alpha.$$

A confidence region at level $1 - \alpha$ is given by:

$$\delta(x) = \{\theta \in \Theta : \delta_\theta(x) = 0\}.$$

Conversely, if δ is a confidence region at level $1 - \alpha$, a test at level α of the null hypothesis $H_0 = \{\theta = \theta_0\}$ against the alternative hypothesis $H_1 = \{\theta \neq \theta_0\}$ is given by:

$$\delta'(x) = 1_{\{\theta_0 \notin \delta(x)\}}.$$

Confidence region $\leftrightarrow$ Hypothesis testing

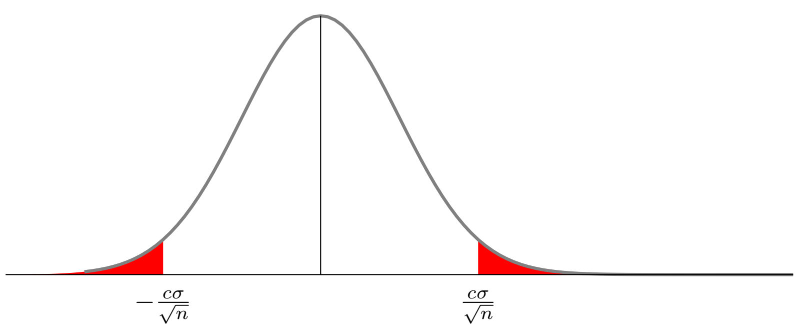
Finding a confidence region at level $1 - \alpha$ is equivalent to test the null hypothesis $H_0 = \{\theta = \theta_0\}$ against the alternative hypothesis $H_1 = \{\theta \neq \theta_0\}$ at level α, for each $\theta_0 \in \Theta$.

Example. *Consider the Gaussian model $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$. Let $c = Q(1 - \frac{\alpha}{2})$ with Q the quantile function of the normal distribution. A test of the null hypothesis $H_0 = \{\theta = 0\}$ against the alternative hypothesis $H_1 = \{\theta \neq 0\}$ is given by:*

$$\delta'(x) = 1_{\left\{0 \in \left[\bar{x} - \frac{c\sigma}{\sqrt{n}}, \bar{x} + \frac{c\sigma}{\sqrt{n}}\right]\right\}},$$

that is

$$\delta'(x) = 1_{\{|\bar{x}| > \frac{c\sigma}{\sqrt{n}}\}}.$$



4 Gaussian model

A practically interesting case for the search of confidence regions is the Gaussian model with unknown mean and variance, $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\mu, \sigma^2), \quad \theta = (\mu, \sigma^2)\}$. We would like to get a confidence interval on the mean, μ, from n i.i.d. observations. Note that confidence intervals considered in the previous examples are not eligible as σ^2 is unknown.

To get a pivot function, define:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{V_0}{n}}}, \tag{1}$$

with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad V_0 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Observe that V_0 corresponds to an unbiased estimator of the variance.

The random variable Z has the Student's distribution with $n-1$ degrees of freedom.

Confidence intervals on the mean can then be constructed using the quantile function of the Student's distribution (also known as the t-distribution). Details about this distribution and the above result are given in the appendix. The corresponding hypothesis test is known as Student's test.

Appendix

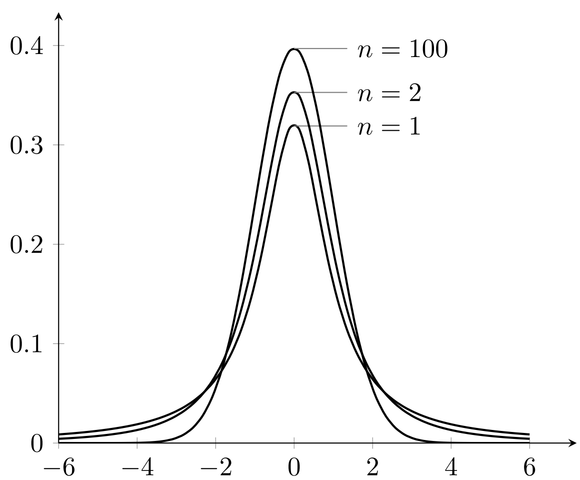
A Student's distribution

A random variable T has the Student's distribution with n degrees of freedom, denoted by $T \sim \mathrm{St}(n)$ if:

$$T \sim \frac{X}{\sqrt{\frac{Y}{n}}},$$

where $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(n)$ are independent. Its density is given by:

$$f(t) \propto \left(\frac{1}{1 + \frac{t^2}{n}}\right)^{\frac{n+1}{2}}.$$



For large n, we have $Y \approx \mathcal{N}(n, 2n)$ so that $\frac{Y}{n} \approx \mathcal{N}(1, \frac{2}{n})$ and $T \approx \mathcal{N}(0,1)$.

B Gaussian model

To prove that the random variable Z given by (1) has a Student's distribution, we need the following result.

Theorem 1 (Cochran's theorem)**.** *Let $X_1, \dots, X_n$ be i.i.d. standard normal random variables. The random variables $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S = \sum_{i=1}^n (X_i - \bar{X})$ are independent, with respective distributions $\mathcal{N}(0, \frac{1}{n})$ and $\chi^2(n-1)$.*

Proof. First observe that $\bar{X} \sim \mathcal{N}(0, \frac{1}{n})$. Since $(\bar{X}, X_1 - \bar{X}, \dots, X_n - \bar{X})$ is a Gaussian vector and

$$\mathrm{cov}(\bar{X}, X_1 - \bar{X}) = \mathrm{cov}(\bar{X}, X_1) - \mathrm{var}(\bar{X}) = 0,$$

we conclude that $\bar{X}$ is independent of $X_1 - \bar{X}, \dots, X_n - \bar{X}$ and thus of S.

To get the distribution of S, we write:

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

that is:

$$S + n\bar{X}^2 = \sum_{i=1}^n X_i^2.$$

The corresponding Laplace transforms satisfy:

$$L_S(t) L_{n\bar{X}^2}(t) = L_{X_1^2}(t)^n.$$

Since $\sqrt{n}\bar{X} \sim \mathcal{N}(0,1)$, we obtain:

$$L_S(t) = L_{X_1^2}(t)^{n-1}.$$

Thus S has distribution $\chi^2(n-1)$. $\square$

Now let $X'_1 = \frac{1}{\sigma}(X_1 - \mu), \dots, X'_n = \frac{1}{\sigma}(X_n - \mu)$. These are i.i.d. standard normal random variables. Observe that:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{V_0}{n}}} = \frac{\bar{X}'}{\sqrt{\frac{V'_0}{n}}},$$

with

$$\bar{X}' = \frac{1}{n}\sum_{i=1}^n X'_i, \quad V'_0 = \frac{1}{n-1}\sum_{i=1}^n (X'_i - \bar{X}')^2.$$

The fact that Z has a Student's distribution follows from Cochran's theorem, on observing that:

$$Z = \frac{\sqrt{n}\bar{X}'}{\sqrt{\frac{S'}{n-1}}},$$

with $S' = \sum_{i=1}^n (X'_i - \bar{X}')^2$.