Statistics

Luc Veldhuis

November 2016

Confidence intervals

Point estimates vs interval

Idea: not a point estimate, but an interval for θ. This interval quantifies the (un)certainty of the estimate.

Example 1

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 unknown.

Define $T = \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim N(0,1)$ with $P(\xi_{\frac{\alpha}{2}} < T < \xi_{1-\frac{\alpha}{2}}) = 1-\alpha$

or $P(\xi_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X}+\mu}{\sigma} < \xi_{1-\frac{\alpha}{2}}) = P(\overline{X} - \frac{\sigma}{\sqrt{n}}\xi_{\frac{\alpha}{2}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}}\xi_{\frac{\alpha}{2}})$

A $(1-\alpha)\%$ confidence interval for μ: $[\overline{X} - \frac{\sigma}{\sqrt{n}}\xi_{\frac{\alpha}{2}}; \overline{X} + \frac{\sigma}{\sqrt{n}}\xi_{\frac{\alpha}{2}}]$

Confidence intervals

Definition

Let X be a random variable with probability distribution p_θ, $\theta \in \Theta$. The mapping $X \mapsto G_x \subset \Theta$ is the confidence interval/region for θ with uncertainty α if $P(\theta \in G_x) \geq 1 - \alpha$ for all $\theta \in \Theta$

Remark

for case at hand, θ is fixed and data is given. So is G_x then either θ is in G_x or it is not, but this is unknown. Interpretation: if we would repeat the experiment x times, and reconstruct the G_x, we expect on average at least $x(1-\alpha)$ of these G_x to contain our θ

Confidence intervals

Remark

A confidence interval for a parameter is generally not unique. E.g. $P(\xi_{\frac{\alpha}{2}} < T < \xi_{\frac{1-\alpha}{2}}) = 1 - \alpha$. One needs to specify whether the interval is symmetric of minimum. (Not sure if I got the last sentence right)

Confidence intervals

It is often practical to use a pivot for confidence intervals

Definition

A pivot is a function $(X, \theta) \to T(X, \theta)$ such that $T(X, \theta)$ has, when θ is the true parameter, a known distribution that does <u>not</u> depend on θ.
E.g. $P_\theta(T(X, \theta) \in B)$ is known for all B. Effectively it cancels both θ's in $P_\theta(T(X, \theta) \in B)$.

Confidence intervals

Constructing a confidence interval with a pivot

Use of a pivot for construction of exact confidence intervals

- Given a point $T(X, \theta)$ find c_1 and c_2 such that $P_\theta(c_1 < T(X, \theta) < c_2) = 1 - \alpha$
- Solve inequalities (in the probabilistic) to arrive at the confidence interval.

$\{\theta \in \Theta : c_1 < T(X, \theta) < c_2\}$
Observe c_1 and c_2 do not depend on the unknown θ and the probability can be calculated.

Remark

'exact' and 'confidence' depends heavilly on the assumptions made.

Example

The pivot $T = \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim N(0,1)$ with known distribution independent of μ.
Choose $B = [\xi_{\frac{\alpha}{2}}, \xi_{1-\frac{\alpha}{2}}]$ for any $\alpha \in (0,1)$

Example

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 unknown.
Pivot: $T = \frac{\overline{X}-\mu}{S_x} \sim t_{n-1}$. Thus
$P_\mu(-t_{n-1;1-\frac{\alpha}{2}} < T < t_{n-1;1-\frac{\alpha}{2}}) = 1-\alpha$
Rewrite: $P_\mu(\overline{X} - \frac{S_x}{\sqrt{n}}t_{n-1;1-\frac{\alpha}{2}} < \mu < \overline{X} + \frac{S_x}{\sqrt{n}}t_{n-1;1-\frac{\alpha}{2}}) = 1-\alpha$
Thus our confidence interval is: $\overline{X} \pm \frac{S_x}{\sqrt{n}}t_{n-1;1-\frac{\alpha}{2}}$
Compare with previous example: this confidence interval is wider, due to heavier tails of t compared to $N(0,1)$.

Confidence intervals

Remark

Confidence interval is larger when:

- σ increases
- n decreases
- α decreases

Find a balance between precision and confidence

Confidence intervals

Example

$X \sim Bin(n,p)$ with $p \in (0,1)$ unknown.
$\hat{p}_{ml} = \frac{x}{n}$ point estimate. Pivot does not exist here.
But there is an approximate pivot.
$T = \frac{x-np}{\sqrt{np(1-p)}} \approx\sim N(0,1)$ as $Bin(n,p) \approx N(np, np(p-1))$ for large n
Find a $1-\alpha$ confidence interval for p. Derive from
$P_p(-\xi_{1-\frac{\alpha}{2}} < \frac{x-np}{\sqrt{np(1-p)}} < \xi_{1-\frac{\alpha}{2}}) = 1-\alpha$ for $x = 102$, $n = 500$, $\alpha = 0.10$.
Then using the ABC-formula: $(102-500)^2 \leq 500p(1-p)(1,96)^2$ solves for $p = 0.187 \vee p = 0.223$. They approximate a 90% confidence interval for $p \in [0.187; 0.223]$

Confidence intervals

Example

Let $X_1, \ldots, X_n \sim U[0, \theta]$ then all $\frac{X_1}{\theta}, \ldots, \frac{X_n}{\theta} \in U[0, 1]$. Thus any function of any of these variables is a pivot. In particular $\frac{X_{(n)}}{\theta}$ as $X_{(n)}$ is the ML estimator of θ.

Confidence intervals

When T_n is an estimator of $g(\theta)$, then often T_n is asymptotically normally distributed (by the Central Limit Theorem):

$$\frac{T_n - g(\theta)}{\sigma_{n,\theta}} \tag{1}$$

if $n \to \infty$ and $\sigma_{n,\theta}$ is the standard deviation of T_n.
Then (1) is an approximate pivot.
In the third example, $T_n = \frac{X}{n}$, $g(\theta) = p$ and $\theta_{n,p} = \sqrt{\frac{1}{n}p(1-p)}$.
Then an approximate $(1-\alpha)\%$ confidence interval for $g(\theta)$: $T_n \pm \sigma_{n,\theta}\xi_{1-\frac{\alpha}{2}}$
Often $\sigma_{n,\theta}$ needs to be estimated. E.g. a ML estimator.

Confidence intervals

Definition

Score function is defined as:

$$\theta \mapsto \frac{\delta}{\delta\theta} log(p_\theta(X)) = l_\theta(X)$$

Definition

Fisher information is defined as:

$$i_\theta = \mathbb{V}(l_\theta(X))$$

Confidence intervals

Lemma

Under some regularly conditions:

$$i_\theta = -\mathbb{E}(\ddot{l}_\theta(X))$$

where $\ddot{l}_\theta = \frac{\delta^2}{\delta^2\theta} log(p_\theta(X))$
Fisher-information $\approx$ curvature of log-likelihood
High curvature = low variance = good
Low curvature = high variance = bad

Confidence intervals

Example

Let $X_1, \dots, X_n \sim Geo(\theta)$ with $\theta \in (0,1)$ unknown.

$$P_\theta(X) = (1-\theta)^{x-1}\theta \quad \mathbb{E}(X_1) = \frac{1}{\theta}$$

$$\begin{aligned}
\mathbb{V}(X_1) &= \frac{1-\theta}{\theta^2} \\
\dot{l}_\theta(X) &= \frac{\delta}{\delta\theta}(log((1-\theta)^{x-1}\theta)) \\
&= \frac{\delta}{\delta\theta}((x-1)log(1-\theta) + log(\theta)) \\
&= -\frac{x-1}{1-\theta} + \frac{1}{\theta} \\
\ddot{l}_\theta(X) &= \frac{x-1}{(1-\theta)^2} - \frac{1}{\theta^2}
\end{aligned}$$

Confidence intervals

Example (continued)

$$\begin{aligned} i_\theta &= \mathbb{V}(-\frac{x-1}{1-\theta} + \frac{1}{\theta}) \\ &= \mathbb{V}(-\frac{x-1}{1-\theta}) \\ &= \frac{1}{(1-\theta)^2}(\frac{1-\theta}{\theta^2}) \\ &= \frac{1}{\theta^2(1-\theta)} \end{aligned}$$

Confidence intervals

Theorem

Let $X_1, \ldots, X_n \sim P_\theta(X)$ and $i_\theta < \infty$.
Define $\hat{\theta}_n$ as the maximum likelihood estimator based on $X_1, \ldots, X_n$, then (under conditions):

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, i_\theta^{-1})$$

if $n \to \infty$
In words, the ML estimator is unbiased asymptotically and normally distributed for large n, thus:

$$\hat{\theta}_n \approx\sim N(\theta, n^{-1} i_\theta^{-1})$$

From this note that variance of the ML estimator decrease as n increases.

Confidence intervals

Theorem (continued)

Approximage pivot:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{n} i_\theta^{-1}}} = \sqrt{n i \theta}(\hat{\theta}_n - \theta) \approx\sim N(0,1)$$

Approximate $(1-\alpha)$ confidence interval for θ is:

$$\theta = \hat{\theta}_n \pm \frac{1}{n i_\theta} \xi_{1-\frac{\alpha}{2}}$$