Confidence Intervals and Bootstrap

Maxim Panov

Skoltech

November, 2020

Outline

Confidence Estimation

Definition

Confidence interval for the parameter for the parameter θ with the confidence level α is a random interval $C_n = (a_n, b_n)$, where

- $a_n = a(X_1, \ldots, X_n)$,
- $b_n = b(X_1, \ldots, X_n)$

are two such functions of the data that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$$

for all $\theta \in \Theta$.

Remark: $C_n = (a, b)$ is random, while θ is a **fixed** unknown quantity.

Remark: If θ is a vector, then C_n is called a *confidence set*.

Confidence Estimation: Remarks

A confidence interval is not a probabilistic statement about θ because θ is not a random variable, but a fixed unknown.

One can use the following interpretations:

- repeating the same experiment many times:
  - $(1 - \alpha)$ * 100% of the time the unknown parameter will fall within the interval;
- constructing confidence intervals for multiple unrelated quantities using the same procedure:
  - $1 - \alpha$ portion of the constructed intervals will contain their corresponding unknown parameters.

Example: Bernoulli

Example

Let $X_1, \dots, X_n$ be $i.i.d.$ having Bernoulli distribution with parameter p.
As $\widehat{p}_n = n^{-1} \sum_{i=1}^{n} X_i$, then

$$C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n),$$

where

$$\epsilon_n^2 = \frac{\log(2/\alpha)}{2n}.$$

Example: Bernoulli

Definition

Hoeffding Inequality.
$Y_1, \ldots, Y_n$ are $i.i.d.$, such that $\mathbb{E}(Y_i) = 0$ and with probability **1**: $a_i \leq Y_i \leq b_i$. Then for any $t > 0$ and $\epsilon > 0$ hold true:

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2 (b_i - a_i)^2 / 8}.$$

Hence, we get:

$$\mathbb{P}(p \in C_n) \geq 1 - \alpha.$$

Confidence intervals for MLE

Theorem

Assume that $\widehat{\theta}_n \rightsquigarrow \mathcal{N}(\theta, \widehat{se}^2)$.
Let Φ *be the standard normal distribution function,* $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, *so*

$$\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$$

and

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha,$$

where $Z \sim \mathcal{N}(0,1)$,

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{se},\ \widehat{\theta}_n + z_{\alpha/2}\widehat{se}).$$

Hence,

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha.$$

Confidence intervals for MLE

Proof

Indeed, let

$$Z_n = (\widehat{\theta}_n - \theta)/\widehat{se}.$$

Then according to the assumption $Z_n \rightsquigarrow Z$*, where* $Z \sim \mathcal{N}(0, 1)$*:*

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta(\widehat{\theta}_n - z_{\alpha/2}\widehat{se} < \theta < \widehat{\theta}_n + z_{\alpha/2}\widehat{se}) \\
&= \mathbb{P}_\theta\Big(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{se}} < z_{\alpha/2}\Big) \\
&\to \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - \alpha.
\end{aligned}$$

A confidence interval of this type is a pointwise asymptotic confidence interval.

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 **is known**.

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 **is known**.

$$s(X;\theta) = (X - \theta)/\sigma^2;$$
$$s'(X;\theta) = -1/\sigma^2;$$
$$I_1(\theta) = 1/\sigma^2;$$
$$\widehat{\theta}_n = \overline{X}_n;$$

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 **is known**.

$$
\begin{aligned}
s(X;\theta) &= (X - \theta)/\sigma^2; \\
s'(X;\theta) &= -1/\sigma^2; \\
I_1(\theta) &= 1/\sigma^2; \\
\widehat{\theta}_n &= \overline{X}_n;
\end{aligned}
$$

As a result we obtain

$$\overline{X}_n \approx \mathcal{N}(\theta, \sigma^2/n). \tag{1}$$

It turns out that in (1) the distribution is exactly normal.

Example: Bernoulli via MLE

Example

Let $X_1, \ldots, X_n \sim Bernoulli(p)$, $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$\mathbb{V}(\widehat{p}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) =$$
$$= \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n},$$

$$se = \sqrt{p(1-p)/n},$$
$$\widehat{se} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}.$$

Example: Bernoulli via MLE

Example (continued)

According to Central Limit Theorem

$$\widehat{p}_n \approx \mathcal{N}(p, \widehat{se}^2).$$

Then the approximate confidence interval with confidence probability $1-\alpha$ has the form

$$\widehat{p}_n \pm z_{\alpha/2}\widehat{se} = \widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

Skoltech
Skolkovo Institute of Science and Technology

Example: Poisson

Example

Let $X_1, \ldots, X_n \sim Poisson(\lambda)$, then

$$\widehat{\lambda}_n = \overline{X}_n \quad \text{and} \quad I_1(\lambda) = 1/\lambda.$$

From which it follows:

$$\widehat{se} = \frac{1}{\sqrt{nI(\widehat{\lambda}_n)}} = \sqrt{\frac{\widehat{\lambda}_n}{n}}.$$

Hence,

$$\widehat{\lambda}_n \pm z_{\alpha/2}\sqrt{\widehat{\lambda}_n/n}.$$

give the bounds of an approximate $1-\alpha$ confidence interval.

Example: Uniform

Example

- Let $X_1, \ldots, X_n \sim Unif(0, \theta)$.
- Find MLE of the parameter θ.

Example: Uniform

Example

- Let $X_1, \dots, X_n \sim Unif(0, \theta)$.
- Find MLE of the parameter θ.
- Density of the uniform distribution:

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in (0,\theta), \\ 0, & x \notin (0,\theta). \end{cases}$$

Example: Uniform

Example

- Let $X_1, \ldots, X_n \sim Unif(0, \theta)$.
- Find MLE of the parameter θ.
- Density of the uniform distribution:

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in (0,\theta), \\ 0, & x \notin (0,\theta). \end{cases}$$

- Fix $\theta < X_i$ for some i,
- then $f(X_i;\theta) = 0$, and so likelihood function becomes $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) = 0$.
  (continued on the next slide)

Example: Uniform

Example (continued)

- From that, $\mathcal{L}_n(\theta) = 0$, if $\theta < X_i$ for at least one i.
- This can be expressed as $\mathcal{L}_n(\theta) = 0$ for $\theta < X_{(n)}$, where $X_{(n)} = \max\{X_1, \ldots, X_n\}$.
- Consider an arbitrary $\theta \geq X_{(n)}$, then $f(X_i; \theta) = 1/\theta$ for any i. Then $\mathcal{L}_n(\theta) = \prod_i f(X_i; \theta) = \theta^{-n}$.
- This produces:

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & \theta \geq X_{(n)}, \\ 0, & \theta < X_{(n)}. \end{cases}$$

- Since $\mathcal{L}_n(\theta)$ is a strictly decreasing function of θ on the interval $[X_{(n)}; \infty)$, then $\widehat{\theta}_n = X_{(n)}$.

Problem Statement

- Model:
  - Given *i.i.d.* sample $X_1, \ldots, X_n \subset \mathbb{R}$ from probability distribution F.
  - Given a functional $T_n = T_n(X_1, \ldots, X_n)$.

- Problem: estimate variance $\mathbb{V}_F(T_n)$ which depends on unknown distribution F.

Example

- $T_n = \overline{X}_n$.
- Get $\mathbb{V}_F(T_n) = \sigma^2/n$, where $\sigma^2 = \int (x-\mu)^2 dF(x)$ and $\mu = \int x dF(x)$.
- Hence, variance T_n is a function of F.

Why do we care?

- We have CLT $\rightarrow$ we can construct confidence intervals automatically!
- Not really, CLT has limitations!
- In practice, normal approximation might be very bad in the case of limited data and complex functional.

Bootstrap Idea

Step 1: Estimate $\mathbb{V}_F(T_n)$ using $\mathbb{V}_{\widehat{F}_n}(T_n)$.

Step 2: Approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ through simulation.

Example

- For $T_n = \overline{X}_n$, $\mathbb{V}_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$, where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.
- In this case, Step 1 is sufficient.
- However, it is not often possible to write explicitly $\mathbb{V}_{\widehat{F}_n}(T_n)$. In this case, apply Step 2.

Variance Estimation

Let $Y_1, \ldots, Y_B$ be sampled *i.i.d.* from G. According to law of averages,

$$\overline{Y}_n = \frac{1}{B} \sum_{j=1}^{B} Y_j \xrightarrow{P} \int y dG(y) = \mathbb{E}(Y), \quad B \to \infty.$$

Hence, we can use $\overline{Y}_n$ to approximate $\mathbb{E}(Y)$ when B is sufficiently large. Besides, for any functional h with finite mathematical expectation we get:

$$\frac{1}{B} \sum_{j=1}^{B} h(Y_j) \xrightarrow{P} \int h(y) dG(y) = \mathbb{E}(h(Y)), \quad B \to \infty.$$

Variance Estimation

In particular, it means that it is possible to model variance:

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y}_n)^2 = \frac{1}{B}\sum_{j=1}^{B}(Y_j)^2 - \left(\frac{1}{B}\sum_{j=1}^{B}Y_j\right)^2 \xrightarrow{P}$$

$$\xrightarrow{P} \int y^2 dG(y) - \left(\int y dG(y)\right)^2 = \mathbb{V}(Y), \quad B \to \infty.$$

- In this way, we can use sample for variance estimation.
- This procedure allows to find $\mathbb{V}_{\widehat{F}_n}(T_n)$ – "variance T_n with sample distributed over $\widehat{F}_n$".

Variance Estimation

Now, there are the following views.

"Reality point of view":

$$F \Rightarrow X_1, \dots, X_n \Rightarrow T_n = g(X_1, \dots, X_n)$$

"Bootstrap point of view":

$$\widehat{F}_n \Rightarrow X_1^*, \dots, X_n^* \Rightarrow T_n^* = g(X_1^*, \dots, X_n^*)$$

Problem: how to get $X_1^*, \dots, X_n^*$ from $\widehat{F}_n$?

Solution: estimating mathematical expectation with $\widehat{F}_n$ we used equal weigths $\frac{1}{n}$. It means, that getting a point $\widehat{F}_n$ is equivalent to choosing a random point from initial sample.

Algorithm

Algorithm of variance estimation using bootstrap:

1. Take $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Take $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat Steps 1 and 2 until you get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\Big(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\Big)^2.$$

As a result:

$$\mathbb{V}_F(T_n) \approx \mathbb{V}_{\widehat{F}_n}(T_n) \approx v_{boot}.$$

Parametric Bootstrap

- Let $F(x) \in \mathfrak{F} = \{f(x, \theta) \colon \theta \in \Theta \subset \mathbb{R}^d\}$.

- Find parameter θ with likelihood maximization:

$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}_n(\vec{X}, \theta).$$

- Besides, you can use method of moments instead of MLE.

Further use the scheme for non-parametric bootstrap described before.

Normal Interval

Let's assume that points distributed normally.
In this case, consider the following confidence interval:

$$(T_n - z_{\alpha/2}\widehat{se}_{boot}, T_n + z_{\alpha/2}\widehat{se}_{boot}),$$

where

- z_α satisfies the condition $F_{\mathcal{N}(0,1)}(z_\alpha) = \Phi(z_\alpha) = 1 - \alpha$,
- $\widehat{se}_{boot} = \sqrt{v_{boot}}$.

Central Interval

- Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$.
- Define $R_n = \widehat{\theta}_n - \theta$.
- Denote $H(r) = \mathbb{P}_F(R_n \leq r)$.
- Denote $C_n^* = (a_n, b_n)$, where

$$a_n = \widehat{\theta}_n - H^{-1}(1 - \alpha/2), \; b_n = \widehat{\theta}_n - H^{-1}(\alpha/2).$$

Hence, we get following chain of equations:

$$\mathbb{P}(a_n \leq \theta \leq b_n) = \mathbb{P}(a_n - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b_n - \widehat{\theta}_n) =$$
$$= \mathbb{P}(\widehat{\theta}_n - b_n \leq \widehat{\theta}_n - \theta \leq \widehat{\theta}_n - a_n) = \mathbb{P}(\widehat{\theta}_n - b_n \leq R_n \leq \widehat{\theta}_n - a_n) =$$
$$= H\left(H^{-1}(1 - \alpha/2)\right) - H\left(H^{-1}(\alpha/2)\right) = 1 - \alpha.$$

As result, C_n^* is $(1 - \alpha)$ is confidence interval for θ.

Central Interval

- Unfortunately, a_n and b_n depends on unknown distribution H, but we can estimate them using bootstrap:

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R^*_{n,b} \le r),$$

where $R^*_{n,b} = \widehat{\theta}^*_{n,b} - \widehat{\theta}_n$.
$\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ from iterations of bootstrap steps 1 and 2.
- Let r^*_β denote β-quantile for $(R^*_{n,1}, \ldots, R^*_{n,B})$
- Let θ^*_β denote β-quantile for $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$.
- Spot that $r^*_\beta = \theta^*_\beta - \widehat{\theta}_n$. Then $(1-\alpha)$-confidence interval is $C_n = (\widehat{a}_n, \widehat{b}_n)$, where

$$\widehat{a}_n = \widehat{\theta}_n - \widehat{H}^{-1}(1-\alpha/2) = \widehat{\theta}_n - r^*_{1-\alpha/2} = 2\widehat{\theta}_n - \theta^*_{1-\alpha/2};$$
$$\widehat{b}_n = \widehat{\theta}_n - \widehat{H}^{-1}(\alpha/2) = \widehat{\theta}_n - r^*_{\alpha/2} = 2\widehat{\theta}_n - \theta^*_{\alpha/2}.$$

Central Interval

Thus, central $(1-\alpha)$-confidence interval:

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}\right).$$

Theorem

With some soft conditions on $T(F)$

$$\mathbb{P}_F(T(F) \in C_n) \to 1-\alpha,\ n \to \infty,$$
$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}\right).$$

Jackknife Method

Let $T_n = (X_1, \ldots, X_n)$.

Consider n subsamples: $T_{(-i)} = \frac{1}{n-1} \sum\limits_{j \neq i} X_j$.

Let $\overline{T}_n = \frac{1}{n} \sum\limits_{i=1}^{n} T_{(-i)}$.

Build the following estimation $\mathbb{V}(T_n)$:

$$v_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(-i)} - \overline{T}_n\right)^2$$

Then, standard error estimation with Jackknife method takes form of $\widehat{se}_{jack} = \sqrt{v_{jack}}$.

It can be shown that $v_{jack}/\mathbb{V}(T_n) \xrightarrow{P} 1$.

Example 1: Skewness

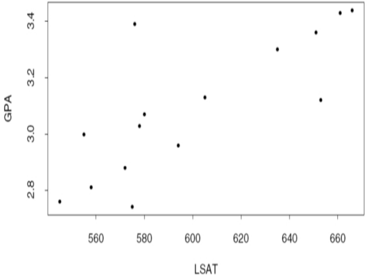
Data: time series of impulses passing nerve fiber.

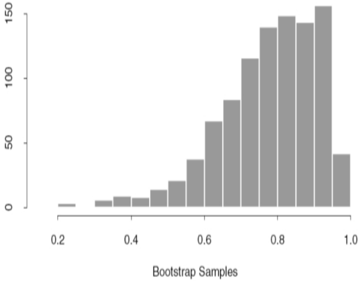
$\theta = T(F) = \int \frac{(x-\mu)^3}{\sigma^3} dF(x)$ – skewness.

1. Variance estimation with non-parametric bootstrap:
$$\widehat{\theta}_n = T(\widehat{F}_n) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3}{\widehat{\sigma}^3} = 1.76.$$
2. $\widehat{\mathbb{V}}^{\text{boot}}_{\widehat{F}_n}(T_n) = (0.16)^2$, with $N = 1000$.
3. 95% skewness interval:
   - normal interval: $(1.44, 2.09)$;
   - central interval: $(1.48, 2.11)$;
   - percentile-based interval: $(1.42, 2.03)$.

Example 2: Two Random Variables Correlation

Data about LSAT(Law School Admissible Test) and GPA(Grade Point Average).





We are interested in correlation between them.

Example 2: Two Random Variables Correlation

1. $\widehat{r}(LSAT, GPA) = \frac{\sum_i (LSAT_i - \overline{LSAT})(GPA_i - \overline{GPA})}{\sqrt{[\sum_i (LSAT_i - \overline{LSAT})^2][\sum_i (GPA_i - \overline{GPA})^2]}} = 0.776.$

2. $\widehat{\mathbb{V}}(\widehat{r}(LSAT, GPA)) = 0.137^2$, with $N = 1000$.

3. 95% skewness interval:
   - normal interval: $(0.51, 1)$;
   - percentile-based interval: $(0.46, 0.96)$.

Example 3: Ratio between Mathematical Expectations of Two Random Variables

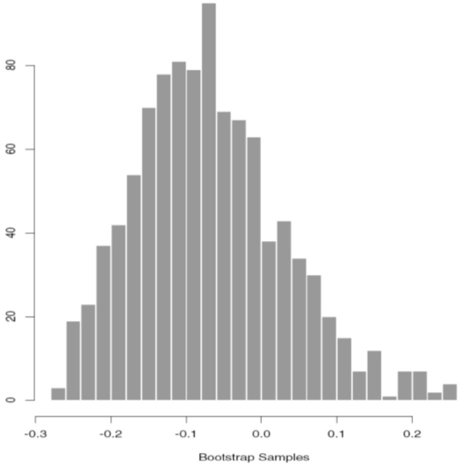
Data on the effectiveness of old and new drugs.

# experiment	placebo	old	new	old - placebo	new - old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8456	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

It is necessary to determine whether the new drug is equivalent to the old one or not.

A new drug is called equivalent to the old one if $\theta = \left|\frac{\mathbb{E}Y}{\mathbb{E}Z}\right| < 0.2$, where $Y = \text{new} - \text{old}$ and $Z = \text{old} - \text{placebo}$.

Example 3



Skoltech
Skolkovo Institute of Science and Technology

Example 3: Ratio between Mathematical Expectations of Two Random Variables

We get the following estimates:

1. $\widehat{\theta} = \frac{\overline{Y}}{\overline{Z}} = -0.0713$.
2. $\widehat{\mathbb{V}}(\widehat{\theta}) = 0.105^2$, with $N = 1000$.
3. 95% interval for $\widehat{\theta}$:
    - Central interval: $(-0.24, 0.15)$.

Therefore, with such precision, we cannot say that they are equivalent, since $(-0.24, 0.15) \notin (-0.2, 0.2)$.

Thank you for your attention!