Lecture 17: Introduction to Interval Estimation

Mathematical Statistics II, MATH 60062/70062

Tuesday April 5, 2022

Reference: Casella & Berger, 9.1

Interval estimation problem

We observe $\boldsymbol{X} = (X_1, \ldots, X_n) \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$.

A **set estimation** makes a statement that "$\boldsymbol{\theta} \in C$," where $C \subset \Theta$ and $C = C(\boldsymbol{x})$ is a set determined by the value of the data $\boldsymbol{X} = \boldsymbol{x}$.

In classical statistics, $\boldsymbol{\theta}$ is regarded as fixed and unknown; $\boldsymbol{\theta} \in C$ shall be interpreted that the set *contains* $\boldsymbol{\theta}$, not $\boldsymbol{\theta}$ is in the set.

For a real-valued parameter θ, we usually prefer the set estimate C to be an **interval**.

Interval estimate and interval estimator

An **interval estimate** of a real-valued parameter θ is any pair of functions

$$L(\boldsymbol{x}) = L(x_1, \ldots, x_n)$$
$$U(\boldsymbol{x}) = U(x_1, \ldots, x_n),$$

satisfying $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. When $\boldsymbol{X} = \boldsymbol{x}$ is observed, the inference

$$L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$$

is made. The random version of the interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an **interval estimator**.

Interval estimate and interval estimator

An **interval estimate** of a real-valued parameter θ is any pair of functions

$$L(\boldsymbol{x}) = L(x_1, \ldots, x_n)$$
$$U(\boldsymbol{x}) = U(x_1, \ldots, x_n),$$

satisfying $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. When $\boldsymbol{X} = \boldsymbol{x}$ is observed, the inference

$$L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$$

is made. The random version of the interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an **interval estimator**.

Sometimes a **one-sided interval** estimate may be formed.

- If $L(\boldsymbol{x}) = -\infty$, then the interval is $(-\infty, U(\boldsymbol{x})]$ with the assertion that $\theta \leq U(\boldsymbol{x})$.
- If $U(\boldsymbol{x}) = \infty$, then the interval is $[L(\boldsymbol{x}), \infty)$ with the assertion that $\theta \geq L(\boldsymbol{x})$.

Coverage probability

For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the **coverage probability** of the interval is

$$P_\theta(L(\boldsymbol{X}) \le \theta \le U(\boldsymbol{X})).$$

Important notes:

- It is the endpoints of the interval $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$ that are random; not the parameter θ (which is fixed).
- The coverage probability is a function of θ. That is, the probability that $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ contains θ may be different for different values of $\theta \in \Theta$.

Confidence coefficient

For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the **confidence coefficient** of the interval is

$$\inf_{\theta \in \Theta} P_\theta(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})).$$

An interval estimator with confidence coefficient equal to $1 - \alpha$ is called a $1 - \alpha$ **confidence interval**.

Similarly, when working with set estimators, a set estimator $C(\boldsymbol{X})$ with confidence coefficient equal to $1 - \alpha$ is called a $1 - \alpha$ **confidence set**.

Uniform interval estimators

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Consider two interval estimators:

1. $(aX_{(n)}, bX_{(n)})$, where $1 \leq a < b$
2. $(X_{(n)} + c, X_{(n)} + d)$, where $0 \leq c < d$.

Uniform interval estimators

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Consider two interval estimators:

1. $(aX_{(n)}, bX_{(n)})$, where $1 \leq a < b$
2. $(X_{(n)} + c, X_{(n)} + d)$, where $0 \leq c < d$.

The PDF of $X_{(n)}$ is

$$\begin{aligned} f_{X_{(n)}}(x) = nf_X(x)[F_X(x)]^{n-1} &= n\left(\frac{1}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1} I(0 < x < \theta) \\ &= \frac{nx^{n-1}}{\theta^n} I(0 < x < \theta). \end{aligned}$$

Uniform interval estimators

Suppose that $X_1, \dots, X_n$ are iid $\mathrm{Unif}(0, \theta)$, where $\theta > 0$. Consider two interval estimators:

1. $(aX_{(n)}, bX_{(n)})$, where $1 \le a < b$
2. $(X_{(n)} + c, X_{(n)} + d)$, where $0 \le c < d$.

The PDF of $X_{(n)}$ is

$$f_{X_{(n)}}(x) = nf_X(x)[F_X(x)]^{n-1} = n\left(\frac{1}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1} I(0 < x < \theta)$$
$$= \frac{nx^{n-1}}{\theta^n} I(0 < x < \theta).$$

The PDF of $T = \frac{X_{(n)}}{\theta}$ is

$$f_T(t) = nt^{n-1} I(0 < t < 1).$$

That is, $T \sim \mathrm{Beta}(n, 1)$.

The **coverage probability** for the first interval is

$$
\begin{aligned}
P_\theta(aX_{(n)} \le \theta \le bX_{(n)}) &= P_\theta\left(\frac{1}{bX_{(n)}} \le \frac{1}{\theta} \le \frac{1}{aX_{(n)}}\right) \\
&= P_\theta\left(\frac{1}{b} \le \boxed{\frac{X_{(n)}}{\theta}} \le \frac{1}{a}\right) \\
&= \int_{1/b}^{1/a} nt^{n-1}dt \\
&= \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n,
\end{aligned}
$$

Finding the distribution of this term

which is the identical for all $\theta \in \Theta = \{\theta : \theta > 0\}$.

The **confidence coefficient** of the interval $(aX_{(n)}, bX_{(n)})$ is

$$
\inf_{\theta>0}\left[\left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n\right] = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.
$$

On the other hand, the **coverage probability** for the second interval is

$$\begin{aligned} P_\theta(X_{(n)} + c \le \theta \le X_{(n)} + d) &= P_\theta(c \le \theta - X_{(n)} \le d) \\ &= P_\theta\left(\frac{c}{\theta} \le 1 - \frac{X_{(n)}}{\theta} \le \frac{d}{\theta}\right) \\ &= P_\theta\left(1 - \frac{d}{\theta} \le \frac{X_{(n)}}{\theta} \le 1 - \frac{c}{\theta}\right) \\ &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n, \end{aligned}$$

which depends on θ.

The **confidence coefficient** of the interval $(X_{(n)} + c, X_{(n)} + d)$ is

$$\inf_{\theta > 0}\left[\left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n\right] = \lim_{\theta \to \infty}\left[\left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n\right] = 0.$$

Bernoulli interval estimator

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(p)$, where $0 < p < 1$. A "$1-\alpha$ confidence interval" commonly taught in Introductory Statistics courses is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard Normal distribution, and $\hat{p}$ is the sample proportion. That is

$$\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where $Y = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$.

This is a large-sample "Wald-type" confidence interval based on the CLT.

The coverage probability of the Wald interval is

$$
\begin{aligned}
&P_p\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
&= E_p\left[I\left(\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}}\right)\right] \\
&= \sum_{y=0}^{n} I\left(\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}} \le p \le \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right)\binom{n}{y}p^y(1-p)^{n-y}.
\end{aligned}
$$

Here Y is hard to be computed directly.

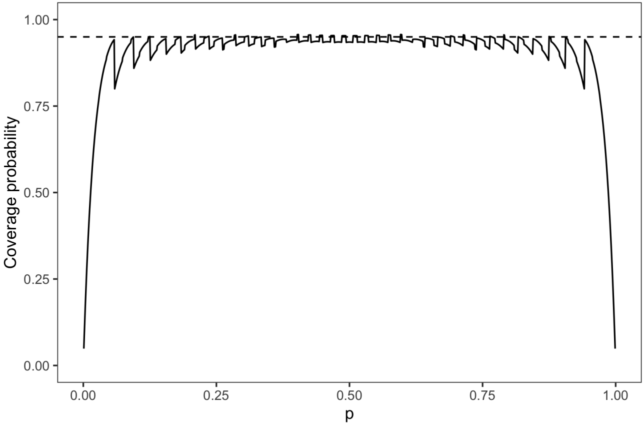
The coverage probability of the Wald interval is

$$
\begin{aligned}
&P_p\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
&= E_p\left[I\left(\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{Y}{n}(1-\frac{Y}{n})}{n}}\right)\right] \\
&= \sum_{y=0}^{n} I\left(\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}} \le p \le \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{\frac{y}{n}(1-\frac{y}{n})}{n}}\right)\binom{n}{y}p^y(1-p)^{n-y}.
\end{aligned}
$$

Ideally, the coverage probability of a 95% Wald-type confidence interval should be close to 0.95. How does this interval perform?

Coverage probability of the Wald interval

Sample size $n = 50$, $\alpha = 0.05$



The coverage probability rarely attains the 0.95 level.

Statistical Science
2001, Vol. 16, No. 2, 101–133

Interval Estimation for a Binomial Proportion

Lawrence D. Brown, T. Tony Cai and Anirban DasGupta

Abstract. We revisit the problem of interval estimation of a binomial proportion. The erratic behavior of the coverage probability of the standard Wald confidence interval has previously been remarked on in the literature (Blyth and Still, Agresti and Coull, Santner and others). We begin by showing that the chaotic coverage properties of the Wald interval are far more persistent than is appreciated. Furthermore, common textbook prescriptions regarding its safety are misleading and defective in several respects and cannot be trusted.

This leads us to consideration of alternative intervals. A number of natural alternatives are presented, each with its motivation and context. Each interval is examined for its coverage probability and its length. Based on this analysis, we recommend the Wilson interval or the equal-tailed Jeffreys prior interval for small n and the interval suggested in Agresti and Coull for larger n. We also provide an additional frequentist justification for use of the Jeffreys interval.

Interval estimation methods

Methods of finding interval estimators

- Inverting a test statistic
- Using pivotal quantities
- Pivoting a CDF
- Bayesian credible intervals

Methods of evaluating interval estimators

- Coverage probability
- Interval length