Probability Theory & Statistics

Interval Estimation

Confidence Intervals

Suppose $X_1, \ldots, X_n$ are random variables with some joint distribution depending on a fixed parameter θ that may be real or vector-valued. In previous sessions we have described point estimators of θ (see Appendix).

The obvious problem with point estimation is the fact that typically $P_\theta(\widehat{\theta} = \theta)$ is small (if not 0) for a given point estimator θ. In practice, we usually attach to any point estimator an estimator of its variability (for example, its standard error); however, this raises the question of exactly how to interpret such an estimate of variability.

An alternative approach to estimation is interval estimation. Rather than estimating θ by a single statistic, we instead give a range of values for θ that we feel are consistent with observed values of $X_1, \ldots, X_n$, in the sense, that these parameter values could have produced (with some degree of plausibility) the observed data.

We will start by considering interval estimation for a single (that is, real-valued) parameter.

Confidence interval. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of random variables with joint distribution depending on a real-valued parameter θ and let $L(\boldsymbol{X}) < U(\boldsymbol{X})$ be two statistics. Then the (random) interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called a $100p\%$ *confidence interval* for θ if

$$P_\theta\left(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\right) \geq p \tag{1}$$

for all θ with equality for at least one value of θ.

The number p is called the *coverage probability* (or simply *coverage*) or *confidence level* of the confidence interval. In many cases, we will be able to find an interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ with

$$P_\theta\left(L(\boldsymbol{X}) \leq \theta \geq U(\boldsymbol{X})\right) = p \tag{2}$$

for all θ. We can also define upper and lower confidence bounds for θ. For example, suppose that

$$P_\theta\left(\theta \geq L(\boldsymbol{X})\right) = p \tag{3}$$

for some statistic $L(\boldsymbol{X})$ and for all θ; then $L(\boldsymbol{X})$ is called a $100p\%$ *lower confidence bound* for θ. Likewise, if

$$P_\theta\left(\theta \leq U(\boldsymbol{X})\right) = p \tag{4}$$

for some statistic $U(\boldsymbol{X})$ and all θ then $U(\boldsymbol{X})$ is called a $100p\%$ *upper confidence bound* for θ.

It is easy to see that if $L(\boldsymbol{X})$ is a $100p_1\%$ lower confidence bound, and $U(\boldsymbol{X})$ is a $100p_2\%$ upper confidence bound for θ, then the interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is a $100p\%$ confidence interval for θ where $p = p_1 + p_2 - 1$ (provided that $L(\boldsymbol{X}) < U(\boldsymbol{X})$).

The interpretation of confidence intervals is frequently misunderstood. Much of the confusion stems from the fact that confidence intervals are defined in terms of the distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ but, in practice, are stated in terms of the observed values of these random variables leaving the impression that a probability statement is being made about θ rather than about the random interval.

However, given data $X = x$, the interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ will either contain the true value of θ or will not contain the true value of θ; under repeated sampling, $100p\%$ of these intervals will contain the true value of θ. This distinction is important but poorly understood by many non-statisticians.

In many problems, it is difficult or impossible to find an exact confidence interval; this is particularly true if a model is not completely specified. However, it may be possible to find an interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ for which

$$P_\theta \left(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X}) \right) \approx p, \tag{5}$$

in which case the resulting interval is called an *approximate* $100p\%$ *confidence interval* for θ.

Example

Suppose that $X_1, \ldots, X_n$ are i.i.d. Normal random variables with mean μ and variance 1. Then $\sqrt{n}(\widehat{X} - \mu) \sim N(0, 1)$ and so

$$P_\mu \left(-1.96 \leq \sqrt{n}(\widehat{X} - \mu) \leq 1.96 \right) = 0.95. \tag{6}$$

The event $\{-1.96 \leq \sqrt{n}(\widehat{X} - \mu) \leq 1.96\}$ is clearly the same as the event $\{\widehat{X} - 1.96/\sqrt{n} \leq \mu \leq \widehat{X} + 1.96/\sqrt{n}\}$ and so we have

$$P_\mu \left(\widehat{X} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \widehat{X} + \frac{1.96}{\sqrt{n}} \right) = 0.95. \tag{7}$$

Thus the interval whose endpoints are $\widehat{X} \pm 1.96/\sqrt{n}$ is a *95% confidence interval for* μ.

Note in this example, if we assume only that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance 1 (not necessarily normally distributed), we have (by the CLT),

$$P_\mu \left(-1.96 \leq \sqrt{n}(\widehat{X} - \mu) \leq 1.96 \right) \approx 0.95. \tag{8}$$

if n is sufficiently large. Using the same argument used above, it follows that the interval whose endpoints are $\widehat{X} \pm 1.96/\sqrt{n}$ is an *approximate 95% confidence interval for* μ.

Interpretation

Confidence intervals and levels are frequently misunderstood, and published studies have shown that even professional scientists often misinterpret them.

- A 95% confidence level *does not mean* that for a given realized interval there is a 95% probability that the population parameter lies within the interval (i.e., a 95% probability that the interval covers the population parameter). According to the frequentist interpretation, once an interval is calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval.
- A 95% confidence level *does not mean* that 95% of the sample data lie within the confidence interval.
- A 95% confidence level *does not mean* that there is a 95% probability of the parameter estimate from a repeat of the experiment falling within the confidence interval computed from a given experiment

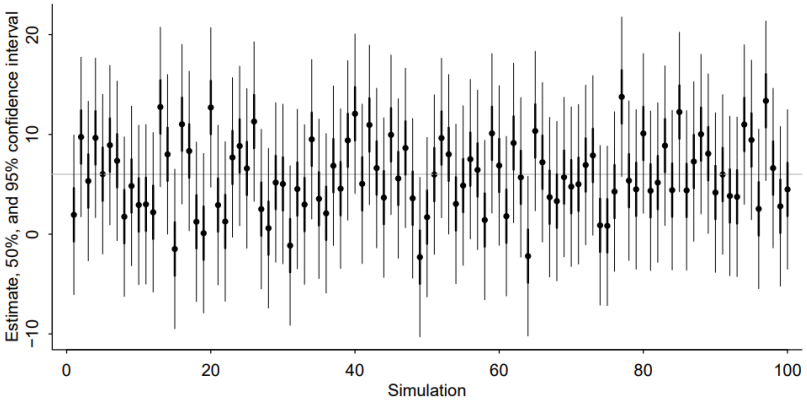

Figure 1: Simulation of coverage of confidence intervals: the horizontal line shows the true parameter value, and dots and vertical lines show estimates and confidence intervals obtained from 100 random simulations from the sampling distribution. If the model is correct, 50% of the 50% intervals and 95% of the 95% intervals should contain the true parameter value, in the long run.

A confidence interval can be interpreted as follows (taking the 95% confidence interval as an example in the following).

- The confidence interval can be expressed in terms of a long-run frequency in repeated samples (or in resampling): *were this procedure to be repeated on numerous samples, the proportion of calculated 95% confidence intervals that encompassed the true value of the population parameter would tend toward 95%.*

- The confidence interval can be expressed in terms of probability with respect to a single theoretical (yet to be realized) sample: *there is a 95% probability that the 95% confidence interval calculated from a given future sample will cover the true value of the population parameter.* This essentially reframes the repeated samples interpretation as a probability rather than a frequency.
- The confidence interval can be expressed in terms of statistical significance, e.g.: *the 95% confidence interval represents values that are not statistically significantly different from the point estimate at the .05 level.*

Counterexample

Since confidence interval theory was proposed, a number of counter-examples to the theory have been developed to show how the interpretation of confidence intervals can be problematic, at least if one interprets them naïvely.

Suppose that X_1 and X_2 are independent observations from a uniform $(\theta - 1/2, \theta + 1/2)$ distribution. Then the optimal 50% confidence procedure for θ is then

$$\widehat{X} \pm \begin{cases} \frac{|X_1 - X_2|}{2}, & \text{if } |X_1 - X_2| < 1/2 \\ \frac{1 - |X_1 - X_2|}{2}, & \text{if } |X_1 - X_2| \geq 1/2. \end{cases} \tag{9}$$

A fiducial or objective Bayesian argument can be used to derive the interval estimate

$$\widehat{X} \pm \frac{|X_1 - X_2|}{4}, \tag{10}$$

which is also a 50% confidence procedure. for every $\theta_1 \neq \theta$, the probability that the first procedure contains θ_1 is *less than or equal* to the probability that the second procedure contains θ_1 . The average width of the intervals from the first procedure is less than that of the second. Hence, the first procedure is preferred under classical confidence interval theory.

However, when $|X_1 - X_2| \geq 1/2$, intervals from the first procedure are guaranteed to contain the true value θ. Therefore, the nominal 50% confidence coefficient is unrelated to the uncertainty we should have that a specific interval contains the true value. The second procedure does not have this property.

Moreover, when the first procedure generates a very short interval, this indicates that X_1 and X_2 are very close together and hence only offer the information in a single data point. Yet the first interval will exclude almost all reasonable values of the parameter due to its short width. The second procedure does not have this property.

The two counter-intuitive properties of the first procedure – 100% coverage when X_1 and X_2 are far apart and almost 0% coverage when X_1 and X_2 are close together – balance out to yield 50% coverage on average. However, despite the first procedure being optimal, its intervals offer neither an assessment of the precision of the estimate

nor an assessment of the uncertainty one should have that the interval contains the true value.

This counter-example is used to argue against naïve interpretations of confidence intervals. If a confidence procedure is asserted to have properties beyond that of the nominal coverage (such as relation to precision, or a relationship with Bayesian inference), those properties must be proved; *they do not follow from the fact that a procedure is a confidence procedure.*

Practice

1) Show Equation 6 is true for the random variable $Z \sim N(0, 1)$.