Confidence Interval and Uncertainty Analysis

MAE301 Applied Experimental Statistics

Yi Ren, Yabin Liao

School for Engineering of Matter, Transport Energy
Arizona State University

October 6, 2015

Outline

Estimation

Confidence interval

Uncertainty analysis

Summary

point estimate

We call the statistic $\hat{\Theta}$ an unbiased estimator of a parameter θ when

$$E(\hat{\Theta}) = \theta. \tag{1}$$

Let $\hat{\theta}$ be a realization of $\hat{\Theta}$. $\hat{\theta}$ is a **point estimate** of θ.

An estimator $\hat{\Theta}$ has a distribution. It is called **the most efficient estimator** of θ when the distribution has the smallest variance.

Cramér-Rao lower bound for random sample

Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be the n samples, $\hat{\Theta}$ a statistic calculated based on $\mathbf{X}$, θ the distribution parameter, then

$$Var(\hat{\Theta}) \geq I(\theta)^{-1}, \tag{2}$$

where $I(\theta)$ is Fisher information value, and is defined as

$$I(\theta) = -E_X\left(\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\right). \tag{3}$$

Here $f(x;\theta)$ is the pdf for X with parameter θ.

Cramér-Rao lower bound for random sample (cont.)

Consider the example where $X \sim N(\mu, \sigma^2)$. Let the parameter of interest be $\theta = \mu$, $\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (the sample mean). Below we show that the variance of the sample mean meets the Cramér-Rao lower bound, and thus it is a **uniformly minimum variance unbiased estimator (UMVUE)** of μ.

Here we have $\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} = -\frac{n}{\sigma^2}$. Thus $I(\theta) = -E_X\left(-\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}$. The Cramér-Rao bound becomes

$$Var(\hat{\Theta}) \geq \frac{\sigma^2}{n}. \tag{4}$$

Since $Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$, the variance of the sample mean meets the bound.

exercise

Let X follow a Bernoulli distribution with parameter p:

$$f(x) = p^x(1-p)^{(1-x)},\ x \in \{0, 1\}. \tag{5}$$

And $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be n samples from X. Calculate the Cramér-Rao lower bound. $(p(1-p)/n)$

correlation

Let X and Y be two random variables. The correlation between the two is

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}, \tag{6}$$

where $Cov(X, Y) = E(XY) - E(X)E(Y)$ is the covariance. X and Y are uncorrelated if $\rho(X, Y) = 0$, i.e., $E(XY) = E(X)E(Y)$.

Show that if X and Y are independent, then they are uncorrelated. Is the converse true?

best linear unbiased estimators

Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ be a sequence of observable real-valued random variables that are uncorrelated and have the same unknown mean μ, but possibly different standard deviations $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$. Define the following estimator for μ:

$$\hat{\Theta} = \sum_{i=1}^{n} c_i X_i \tag{7}$$

We have the following:

- $\hat{\Theta}$ is unbiased if and only if $\sum_{i=1}^{n} c_i = 1$.
- The variance of $\hat{\Theta}$ is $\sum_{i=1}^{n} c_i^2 \sigma_i^2$.
- The variance is minimized, subject to the unbiased constraint, when $c_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2}$, $i = 1, 2, \cdots, n$.

best linear unbiased estimators (cont.)

This shows how to construct **the Best Linear Unbiased Estimator (BLUE)** of μ, assuming that the vector of standard deviations σ is known.

This also shows that the sample mean is the BLUE of μ when the standard deviations are the same, and that moreover, we do not need to know the value of the standard deviation.

confidence interval

Even the most efficient unbiased estimator is unlikely to estimate the population parameter exactly.

There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the population parameter. Such an interval is called an **interval estimate**.

Mathematically, an interval is determined from the selected sample such that

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha, \tag{8}$$

where $\hat{\Theta}_L < \theta < \hat{\Theta}_U$ is called the **confidence interval** (CI) for the population parameter at a confidence level of $100(1-\alpha)\%$, and $\hat{\Theta}_L$ and $\hat{\Theta}_R$ are called the lower and upper confidence limits.

confidence interval (cont.)

Let $\bar{X}$ be the sample mean and normally distributed: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Assume that σ^2 is known. We have

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha. \tag{9}$$

To calculate the confidence limits, we can rewrite this as

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{10}$$

Therefore, for given $\bar{x}$ and σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

exercise

The average zinc concentration recovered from a sample of zinc measurements in 36 different locations is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3.

interpretation of the confidence interval

Following the presented estimation procedure, $100(1-\alpha)\%$ of time the confidence interval of the population mean will cover μ.

Accuracy of the point estimate: If $\hat{x}$ is used as an estimate of μ, we can then be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

For given $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, we cannot say that μ has $100(1-\alpha)\%$ chance to lie between the two. This is because we assume μ to be a constant number. When the bounds are calculated, they either cover μ or not. There is a definitive answer which we don't know.

sample size for desired accuracy in the point estimate

If we do not want the error in the estimate to exceed a specified amount δ:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \delta, \tag{11}$$

the condition on the sample size is

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{\delta}\right)^2 \tag{12}$$

exercise

An efficiency expert wishes to determine the average time that it takes to drill three holes in a certain metal clamp. How large a sample will he need to be 95% confident that his sample mean will be within 15 seconds of the true mean. Assume that it is known from previous studies that $\sigma = 40$ seconds.

confidence interval when σ is unknown

If $\bar{x}$ and s are the sample mean and standard deviation a random sample of size n from a normal population with unknown variance, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{13}$$

where $t_\alpha/2$ is the t-value with $\nu = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

Again, it should be noted that the t-distribution is derived based on the assumption that the sample is obtained from a normal population (or approximately a normal population based on the central limit theorem).

exercise

An accelerometer is used to measure the acceleration of a moving cart. The measurements are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 m/s^2. Find a 90% confidence interval for the mean of accelerations, assuming an approximate normal distribution.

exercise

In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular experiment is measured. Past experience suggests that the variance in reaction time to these types of stimuli are 4 sec^2 and that reaction time is approximately normal. The average time for the subjects was 6.2 seconds.

Determine the upper 95% bound for the mean reaction time

exercise

An experiment was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon (MPG), was measured. Fifty experiments were conducted using engine type A and 75 experiments were done for engine type B. The average gas mileage for engine A was 36 MPG and the average for machine B was 42 MPG.

Construct a 96% confidence interval on $\mu_B - \mu_A$. Assume that the population standard deviations are 6 and 8 for machines A and B, respectively.

exercise

An experiment reported in Popular Science compared fuel economics for two types of similarly equipped diesel mini-trucks. Suppose that 12 Volkswagen and 10 Toyota trucks are used in the tests. If the 12 Volkswagen trucks average 38 MPG (miles per gallon) with a standard deviation of 2.35 MPG, and the 10 Toyota trucks average 26 with a standard deviation of 1.88 MPG.

Construct a 90% confidence interval for the difference between the average MPG of these two minivans. Assume approximately normal distributions with equal variances.

exercise

The following data represent the running times of films produced by two motion-picture companies:

Company 1	103	94	110	87	98		
Company 2	97	82	123	92	175	88	118

Compute a 90% confidence interval for the difference between the average running times of films produced by the two companies. Assume that the running-time differences are approximately normally distributed with unequal variances.

confidence interval for the variance

If s^2 is the sample variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}. \tag{14}$$

The following are the Youngs modulus measurements, in GPa, of 5 specimens of a metal distributed by a materials company: 46.4, 46.1, 45.8, 47.0, 46.1. Find a 95% confidence interval for the variance of all such metal distributed by this company, assuming a normal population.

confidence interval for the variance

If s_1^2 and s_2^2 are sample variances of independent samples of size n_1 and n_2, respectively, from normal populations, then a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is given by

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(\nu_1,\nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(\nu_2,\nu_1), \tag{15}$$

where $f_{\alpha/2}(\nu_1,\nu_2)$ is an f-value with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom leaving an area of $\alpha/2$ to the right.

exercise

A taxi company is trying to decide whether to purchase brand *A* or *B* tires for its fleet of taxis. An experiment was conducted using 12 of each brand. The tires are run until they wear out. The results (in miles) are:

Brand	Sample mean	Sample standard deviation
A	36,300	5,000
B	38,100	6,100

Construct a 90% confidence interval for σ_1^2/σ_2^2. Should the equal-variance assumption be used to study the difference in the means?

uncertainty

Uncertainty: another way of describing the confidence interval:

$$x = \bar{x} + u_{\bar{x}} \; (P\%), \tag{16}$$

where x is the true value, $\bar{x}$ is the measured value, $u_{\bar{x}}$ is the uncertainty in $\bar{x}$ at confidence level $P\%$. (Confidence level is the complement of significance level.)

For example, $m = 100 \pm 3kg$ (95%) has the same meaning as $P(97 \leq m \leq 103) = 0.95$.

propagation of uncertainty

The experimental result, R, is usually a function of several independent variable, $x_1, x_2, \cdots, x_n$:

$$R = R(x_1, x_2, \cdots, x_n). \tag{17}$$

The uncertainty in the result, u_R, can be calculated from the uncertainties in the independent variables

$$u_R = \sqrt{\left(\frac{\partial R}{\partial x_1} u_{x_1}\right)^2 + \left(\frac{\partial R}{\partial x_2} u_{x_2}\right)^2 + \cdots + \left(\frac{\partial R}{\partial x_n} u_{x_n}\right)^2}. \tag{18}$$

Note that the derivation of this formula when R is a nonlinear function of $x_1, x_2, \cdots, x_n$ relies on truncated Taylor expansion. Therefore it is a good approximation only when variances in x are small.

exercise

Power measurement $P = UI$, with voltage and current obtained as $U = 200 \pm 3(V)$ and $I = 2 \pm 0.1(A)$. Then the uncertainty in P is

$$u_P = \sqrt{\left(\frac{\partial P}{\partial U} u_U\right)^2 + \left(\frac{\partial P}{\partial I} u_I\right)^2} = 20.88. \qquad (19)$$

summary of the class

- unbiased estimator, point estimate
- Cramér-Rao lower bound, UMVUE, BLUE
- Confidence intervals for Z-statistic, T-statistic, χ^2-statistic, and F-statistic
- Uncertainty analysis