Lecture 14
Interval Estimation

Chao Song

College of Ecology

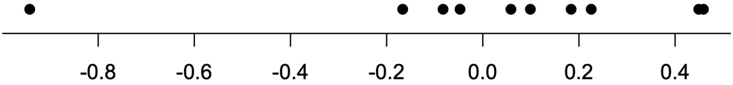
Lanzhou University

November 7, 2024

Point and interval estimation

A point estimate gives us a single value estimate of the parameter of interest. But the probability that the point estimate is exactly the true value of the parameter is 0 and we do not know how close the point estimate is to the true value of the parameter.

The point estimate is a function of the random sample and will typically vary from one sample to another. Thus, based on the uncertainty of the point estimate, we may be able to give plausible range of values that may contain the true value of the parameter.



A motivating example

Given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$. Suppose that σ^2 is known. We know that the sample mean $\overline{X}$ is $N(\mu, \sigma^2/n)$. Thus, based on the properties of normal distribution, we have

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

where α is a pre-specified probability and $z_{\alpha/2}$ is the quantile of a standard normal distribution with tail probability $\alpha/2$. Rearrange the probability statement above, we get

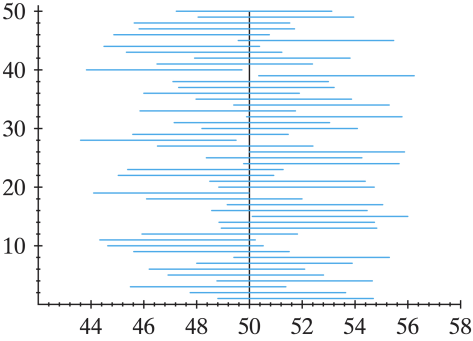
$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

A motivating example

The probability that the random interval

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

contains μ is $1 - \alpha$. We call this interval the $100(1-\alpha)\%$ **confidence interval** for μ and $1-\alpha$ is the **confidence level** or **confidence coefficient**.



Confidence interval

What confidence interval means is that if we repeat the experiment and collect the same kind of data many times, and calculate the confidence interval each time, $100(1-\alpha)\%$ of all these intervals contain the true value of the parameter.

- Confidence coefficient tells us the probability that a confidence interval covers the true value before the sample is drawn;
- Roughly speaking, confidence is about the method of calculating confidence interval;
- Once an interval is calculated based on a particular sample, it is incorrect to state how much likely this particular interval contains the true value. There should be no probability statement made about a particular realized interval.

Constructing confidence interval

In the motivating example, we derive the confidence interval based on the fact that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution. Notice two properties of this quantity:

- It is a function of the sample measurements and the unknown parameter μ, and μ is the only unknown quantity.
- It has a known probability distribution and the distribution does not depend on the unknown parameter μ.

A quantity that possesses these two properties are called a **pivotal quantity**. Pivotal method is a useful way to find confidence interval.

Confidence interval for the mean

If we drawn a random sample from a normal distribution with **known** variance σ^2, the confidence interval for the mean can be constructed as

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Example: Let X equal the length of life of a 60-watt light bulb by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\overline{X} = 1478$ hours, then the 95% confidence interval for the mean μ is

$$\left[1478 - 1.96 \times \frac{\sqrt{1296}}{\sqrt{27}}, 1478 + 1.96 \times \frac{\sqrt{1296}}{\sqrt{27}}\right] = [1464.42, 1491.58]$$

Confidence interval for the difference of two means

Suppose that we are interested in comparing the means of two normal distributions. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes n and m from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Suppose for now the variances are known. What is the confidence interval of $\mu_X - \mu_Y$?

The sample mean from a normal distribution also has a normal distribution, i.e., $\overline{X} \sim N(\mu_X, \sigma_X^2/n)$ and $\overline{Y} \sim N(\mu_Y, \sigma_Y^2/m)$. Because X and Y are independent, $\overline{X} - \overline{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$. Thus,

$$P\Big(-z_{\alpha/2} < \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} < z_{\alpha/2} \Big) = 1 - \alpha$$

Confidence interval for the difference of two means

The confidence interval for the difference of two means from normal distributions with known variances is

$$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sigma_W,\ \overline{X} - \overline{Y} + z_{\alpha/2}\sigma_W\right]$$

where $\sigma_W = \sqrt{\sigma_X^2/n + \sigma_Y^2/m}$ is the standard deviation of $\overline{X} - \overline{Y}$.

Example: Suppose we have two samples, let $n = 15$, $m = 8$, $\overline{X} = 70.1$, $\overline{Y} = 75.3$, $\sigma_X^2 = 60$, $\sigma_Y^2 = 40$. What is the 90% confidence interval of $\mu_X - \mu_Y$?

Here, $\alpha = 0.1$ and $z_{0.05} = 1.645$. $\sigma_w = \sqrt{(60/15) + (40/8)} = 3$, $\overline{X} - \overline{Y} = 70.1 - 75.3 = -5.2$. The confidence interval is thus

$$[-5.2 - 1.645 \times 3, -5.2 + 1.645 \times 3] = [-10.135, -0.265]$$

Confidence intervals with unknown variance

Our construction of confidence interval so far relies on known variance, i.e.,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a pivotal quantity only if σ^2 is known. But in practice, the variance is often unknown, how do we deal with this?

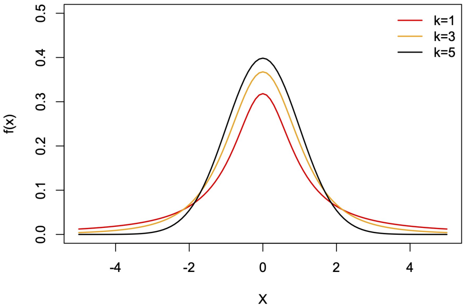
We need to find a new pivotal quantity that we know its distribution without knowing σ^2. To do that, we need to introduce **t-distribution**, and some properties about sample mean and variance from a normal distribution.

t-distribution

Definition: If $X \sim N(0,1)$, $v \sim \chi^2(k)$ and the X and v are independent, than the random variable T defined below follows a $t-$distribution with k degrees of freedom.

$$T = \frac{X}{\sqrt{v/k}}$$

As the degrees of freedom increases, a t-distribution converge to a normal distribution.



Sample mean and variance of a normal distribution

Proposition: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$, then

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

and $\overline{X}$ and s^2 are independent.

Sample mean and variance of a normal distribution

Proof: Proving the independence of $\overline{X}$ and s^2 is beyond the scope of this course. Here, we provide a sketch proof. Note that s^2 is a function of all $X_i - \overline{X}$. If $X_i - \overline{X}$ and $\overline{X}$ are independent, s^2 and $\overline{X}$ are independent.

$$Cov(X_i - \overline{X}, \overline{X}) = Cov(X_i, \overline{X}) - Cov(\overline{X}, \overline{X}) = 0$$

because here,

$$\begin{aligned} Cov(X_i, \overline{X}) &= Cov(X_i, \sum_{j=1}^{n} \frac{X_j}{n}) = \sum_{j=1}^{n} Cov(X_i, \frac{1}{n} X_j) \\ &= Cov(X_i, \frac{1}{n} X_i) = \frac{\sigma^2}{n} \\ Cov(\overline{X}, \overline{X}) &= Var(\overline{X}) = \frac{\sigma^2}{n} \end{aligned}$$

A zero covariance between $X_i - \overline{X}$ and $\overline{X}$ shows that they are independent.

Sample mean and variance of a normal distribution

From the lecture on transformation of random variables, we have seen that if $X_i \sim N(\mu, \sigma^2)$, the sample mean $\overline{X} \sim N(\mu, \sigma^2/n)$. Now, we prove that $(n-2)s^2/\sigma^2$ has a chi-square distribution. Consider a random variable W

$$\begin{aligned}
W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 &= \sum_{i=1}^{n} \left[\frac{(X_1 - \overline{X}) + (\overline{X} - \mu)}{\sigma} \right]^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \\
&= \frac{(n-1)s^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}
\end{aligned}$$

because the cross product term is equal to

$$2 \sum_{i=1}^{n} \frac{(X_i - \overline{X})(\overline{X} - \mu)}{\sigma^2} = \frac{2(\overline{X} - \mu)}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

Sample mean and variance of a normal distribution

Recall that the square of a standard normal variable is $\chi^2(1)$ and the sum of k independent $\chi^2(1)$ distributed variables follows $\chi^2(k)$. Here, $X_i \sim N(\mu, \sigma^2)$ and $\overline{X} \sim N(\mu, \sigma^2/n)$. Thus,

$$W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

$$\frac{n(\overline{X} - \mu)^2}{\sigma^2} = \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi^2(1)$$

These two terms are independent because $\overline{X}$ and s^2 are independent. Thus

$$\frac{(n-1)s^2}{\sigma^2} = W - \frac{n(\overline{X} - \mu)^2}{\sigma^2} \sim \chi^2(n - 1.)$$

Confidence interval with unknown variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. The following quantity

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

has a t-distribution with $n - 1$ degrees of freedom.

Proof: Using the properties of the sample mean and variance of a normal distribution, we have

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

and they are independent. Using the definition of t-distribution,

$$\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2} \Big/ (n-1)}} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

Confidence interval with unknown variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. If σ^2 is unknown, the $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}(n-1)$ is the quantile in a t-distribution with $n-1$ degrees of freedom and tail probability $\alpha/2$.

Confidence interval with unknown variance

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes n and m from two normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. When $\sigma_X = \sigma_Y$, the confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}(n + m - 2) S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where S_p is the pooled estimator of the common standard deviation

$$S_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}$$

Confidence interval with unknown variance

Proof: Since both X and Y follow normal distributions and they are independent,

$$\overline{X} - \overline{Y} \sim N\Big(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\Big)$$

Using properties of the mean and variance of normal distribution,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$$

$$\frac{(n-1)s_X^2}{\sigma_X^2} + \frac{(m-1)s_Y^2}{\sigma_Y^2} \sim \chi^2(n+m-2)$$

The latter coming from the fact that the sum of independent χ^2 distributed random variables also follows a χ^2 distribution.

Confidence interval with unknown variance

Using the definition of t-distribution, we have

$$\frac{\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}}{\sqrt{\frac{(n-1)s_X^2}{\sigma_X^2}+\frac{(m-1)s_Y^2}{\sigma_Y^2}\Big/(n+m-2)}} \sim t(n+m-2)$$

If $\sigma_X = \sigma_Y$, the variance term in the numerator and denominator cancels out, we have

$$\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim t(n+m-2)$$

This allows us to construct a confidence interval for $\mu_X - \mu_Y$, assuming $\sigma_X = \sigma_Y$.