Lecture 9
Interval Estimation

Chao Song

College of Ecology

Lanzhou University

November 6, 2025

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.

Point and interval estimation

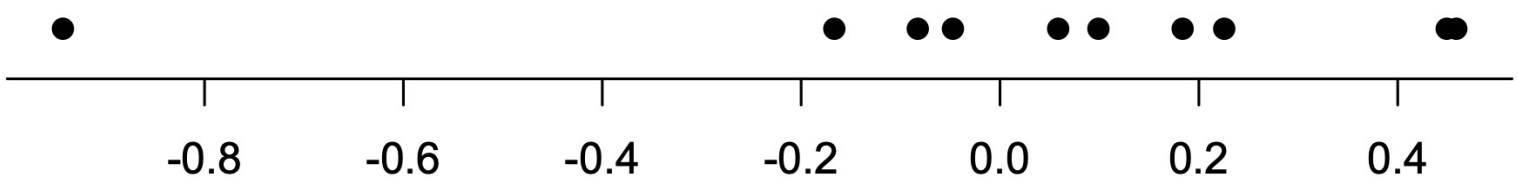
A point estimate gives us a single value estimate of the parameter of interest. But the probability that the point estimate is exactly the true value of the parameter is 0 and we do not know how close the point estimate is to the true value of the parameter.

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.

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 $\alpha/2$.

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

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

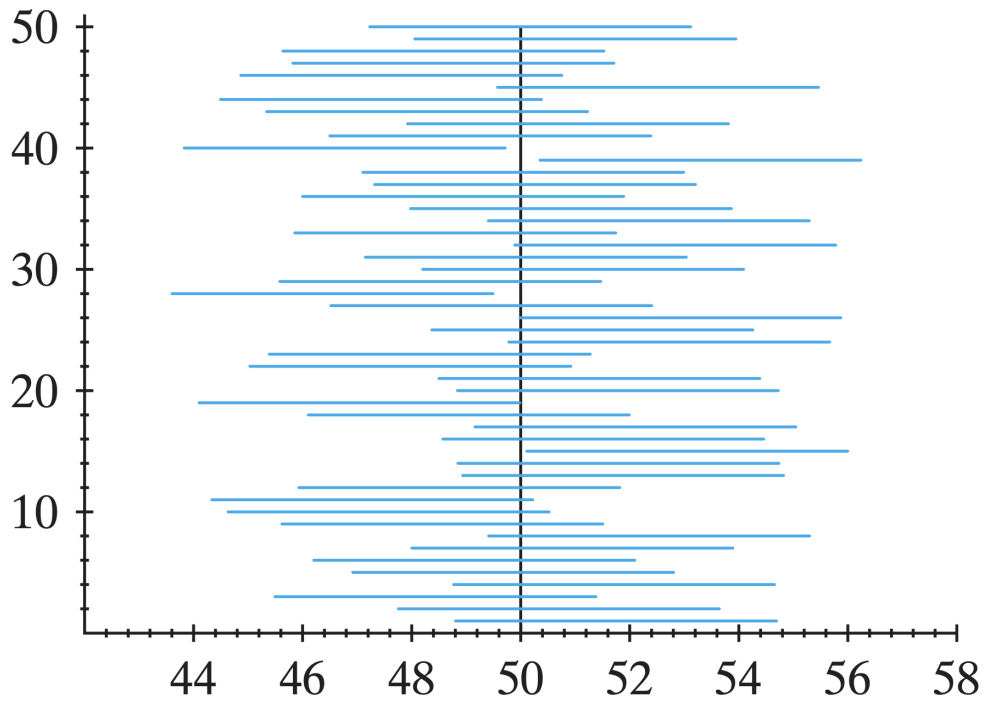
contains μ is $1-\alpha$. We call this interval the $100(1-\alpha)\%$ **confidence interval** for μ and $1-\alpha$ is the **confidence level** or **confidence coefficient**.

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:

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.

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

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

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 $\mu_X - \mu_Y$?

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,

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

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 $\overline{X} - \overline{Y}$.

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 $\mu_X - \mu_Y$?

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?

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.

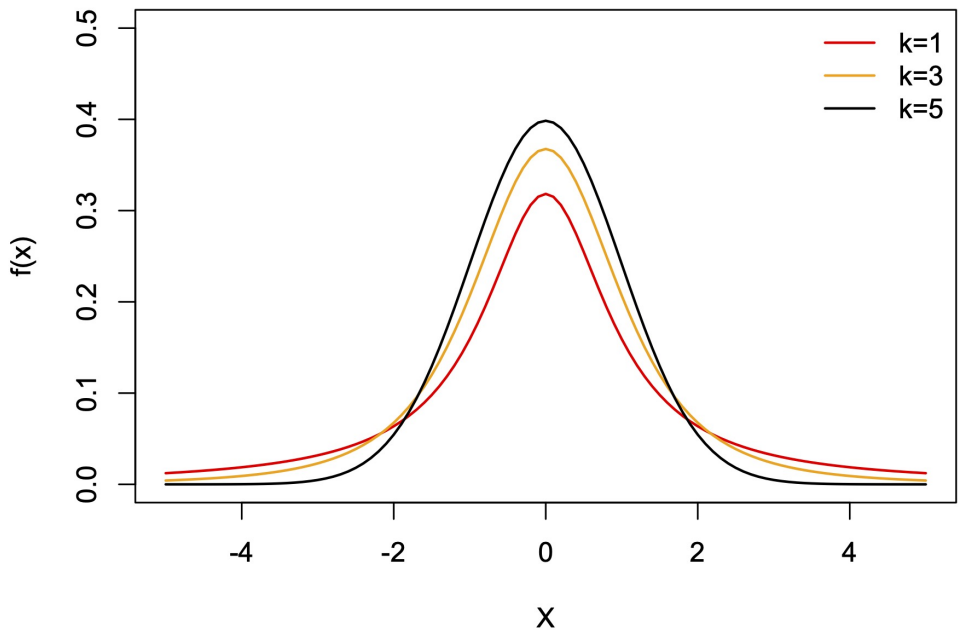
$$T = \frac{X}{\sqrt{v/k}}$$

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 only provide a sketch proof of the distribution properties.

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 only provide a sketch proof of the distribution properties. 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, using moment generating function of χ^2 distribution, it is easy to show that

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

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.

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,

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

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 $\sigma_X = \sigma_Y$.

Interval estimation

We have used the **pivotal method** for constructing confidence intervals. However, it is sometimes impossible to find a pivotal quantity. Under these situations, we need other techniques to find confidence intervals.

- Asymptotic distribution of statistics;
- Distribution free confidence intervals;
- Resampling based confidence intervals.

Confidence interval of maximum likelihood estimate

Recall that the maximum likelihood estimate $\hat{\theta}$ for a parameter θ asymptotically has a **normal** distribution

$$\hat{\theta} \sim N(\theta, I(\theta)^{-1})$$

where $I(\theta)$ is the Fisher information defined as

$$\begin{aligned} I(\theta) &= E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right] \\ &= -E\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right) = -nE\left(\frac{\partial^2 \ln f(x|\theta)}{\partial \theta^2}\right) \end{aligned}$$

Confidence interval of maximum likelihood estimate

The asymptotic properties of maximum likelihood estimates provide a generally applicable approach to deriving confidence interval.

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{I(\theta)^{-1}}} < z_{\alpha/2}) \approx 1 - \alpha$$

Thus, the $100(1 - \alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I(\theta)}}$$

Confidence interval of maximum likelihood estimate

The asymptotic properties of maximum likelihood estimates provide a generally applicable approach to deriving confidence interval.

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{I(\theta)^{-1}}} < z_{\alpha/2}) \approx 1 - \alpha$$

Thus, the $100(1 - \alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I(\theta)}}$$

Because θ is unknown, $I(\theta)$ is approximated by the observed Fisher information $I(\hat{\theta})$, i.e. Fisher information evaluated at the maximum likelihood estimate $\hat{\theta}$.

Confidence interval of maximum likelihood estimate

Previously, we can analytically derive the distribution of sample mean based on a sample from a normal distribution to construct its confidence interval. Now, we consider using the maximum likelihood framework to do so.

Confidence interval of maximum likelihood estimate

Previously, we can analytically derive the distribution of sample mean based on a sample from a normal distribution to construct its confidence interval. Now, we consider using the maximum likelihood framework to do so.

For a sample $X_1, X_2, \ldots, X_n$, the log-likelihood is

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma^2}$$

$$\frac{\partial^2 \ln L(\mu, \sigma^2)}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$I(\mu) = -E\left(\frac{\partial^2 \ln L(\mu, \sigma^2)}{\partial \mu^2}\right) = \frac{n}{\sigma^2}$$

Confidence interval of maximum likelihood estimate

From previous lectures, we know that the maximum likelihood estimate for μ is $\overline{X}$. The confidence interval can thus be approximated by using observed Fisher information for deriving the variance.

$$\overline{X} \pm z_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}$$

where $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$ is the variance of MLE evaluated at MLE $\overline{X}$.

Confidence interval of maximum likelihood estimate

Recall the problem of calculating confidence interval for $\mu_X - \mu_Y$ when we cannot assume that $\sigma_X = \sigma_Y$. We can use the maximum likelihood approach to deriving its confidence interval. Recall that

$$(X - Y) \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Thus, using the result of confidence interval for the mean of a normal distribution, the confidence interval is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{s_X^2/n + s_Y^2/m}$$

Confidence interval of maximum likelihood estimate

Recall the problem of calculating confidence interval for $\mu_X - \mu_Y$ when we cannot assume that $\sigma_X = \sigma_Y$. We can use the maximum likelihood approach to deriving its confidence interval. Recall that

$$(X - Y) \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Thus, using the result of confidence interval for the mean of a normal distribution, the confidence interval is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{s_X^2/n + s_Y^2/m}$$

Note here that the variance of the MLE should be its variance evaluated at the MLE. Thus $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $s_Y^2 = \frac{1}{n}\sum_{i=1}^{m}(X_i - \overline{X})^2$

Confidence interval of maximum likelihood estimate

Comments on confidence interval of the difference of means:

- Depending on whether we can assume equal variance of X and Y, the confidence interval for $\mu_X - \mu_Y$ differs. It is thus critical to assess whether the equal variance assumption is valid or not.
- The confidence interval based on equal variance performs very poorly when variances are not actually equal and sample size of X and Y differ substantially.
- if the sample variance differs a lot between X and Y and their respective sample size are vastly different, it is safer and more robust to construct confidence interval assuming unequal variance.

Confidence interval for proportions

Let X be the number of success in n independent Bernoulli trials. How do we construct confidence interval for the success probability p?

Confidence interval for proportions

Let X be the number of success in n independent Bernoulli trials. How do we construct confidence interval for the success probability p?

The log-likelihood is

$$\begin{aligned}
\ln L(p) &= \ln\left(\mathbf{C}_n^k p^X (1-p)^{n-X}\right) \\
&= \ln \mathbf{C}_n^X + x \ln p + (n-x)\ln(1-p) \\
\frac{d \ln L(p)}{dp} &= \frac{x}{p} - \frac{n-x}{1-p} \\
\frac{d^2 \ln L(p)}{dp^2} &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \\
I(p) &= -E\left(\frac{d^2 \ln L(p)}{dp^2}\right) = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}
\end{aligned}$$

Confidence interval for proportions

The maximum likelihood estimate $\hat{p} = X/n$ is obtained by

$$\frac{d \ln L(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

Using the asymptotic properties of maximum likelihood estimate

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2} \right) \approx 1 - \alpha$$

Confidence interval for proportions

The maximum likelihood estimate $\hat{p} = X/n$ is obtained by

$$\frac{d \ln L(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

Using the asymptotic properties of maximum likelihood estimate

$$P\Big(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2} \Big) \approx 1 - \alpha$$

Using observed Fisher information, the $100(1-\alpha)\%$ confidence interval is

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

Confidence interval for proportions

We use the asymptotic properties of MLE to construct confidence interval. When n is large and p is not too small, the coverage probability is approximately correct. But if n is not sufficiently large or if p is fairly close to 0 or 1, improvements are needed.

Confidence interval for proportions

We use the asymptotic properties of MLE to construct confidence interval. When n is large and p is not too small, the coverage probability is approximately correct. But if n is not sufficiently large or if p is fairly close to 0 or 1, improvements are needed.

The **Wilson score method** directly solve the inequality

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

to obtain the confidence interval

$$\frac{\hat{p} + z^2_{\alpha/2}/(2n) \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/(4n^2)}}{1 + z^2_{\alpha/2}/n}$$

Confidence interval for proportions

Agresti and Coull (1998) suggested that we use $\tilde{p} = (X + 2)/(n + 4)$ as an estimator for p when n is small or if X is close to 0 or n. The confidence interval is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1 - \tilde{p})/(n + 4)}$$

Confidence interval for proportions

Agresti and Coull (1998) suggested that we use $\tilde{p} = (X + 2)/(n + 4)$ as an estimator for p when n is small or if X is close to 0 or n. The confidence interval is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$$

If we form 95% confidence interval, $z_{\alpha/2} = 1.96 \approx 2$. The 95% confidence interval using the Wilson score method, it is centered at

$$\frac{\hat{p} + z^2_{\alpha/2}/(2n)}{1 + z^2_{\alpha/2}/n} = \frac{X + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}} \approx \frac{X+2}{n+4}$$

Thus it is roughly consistent with the Agresti and Coull method.

Distribution-free confidence intervals

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be a random sample sorted from the smallest to the largest. We call $X_{(j)}$ the jth order statistics of the random sample. For the 100p% percentile of the distribution m, we have

$$P(X_{(i)} < m < X_{(j)}) = \sum_{k=i}^{j-1} \mathbf{C}_n^k p^k (1-p)^{n-k} = 1 - \alpha$$

Distribution-free confidence intervals

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be a random sample sorted from the smallest to the largest. We call $X_{(j)}$ the jth order statistics of the random sample. For the 100p% percentile of the distribution m, we have

$$P(X_{(i)} < m < X_{(j)}) = \sum_{k=i}^{j-1} \mathbf{C}_n^k p^k (1-p)^{n-k} = 1 - \alpha$$

This approach only uses the order statistics to construct confidence intervals. Little is assumed about the underlying distribution, except that the distribution is continuous. Thus, these confidence intervals are called **distribution-free confidence intervals**.

Distribution-free confidence intervals

Example: Suppose we have a sample $X_1 < X_2 < X_3 < X_4 < X_5$. One confidence interval of the median m is

$$P(X_1 < m < X_5) = \sum_{k=1}^{4} \mathbf{C_5^k} (\frac{1}{2})^k (\frac{1}{2})^{5-k} = 0.9375$$

$$P(X_2 < m < X_4) = \sum_{k=2}^{3} \mathbf{C_5^k} (\frac{1}{2})^k (\frac{1}{2})^{5-k} = 0.625$$

Distribution-free confidence intervals

Example: Suppose we have a sample $X_1 < X_2 < X_3 < X_4 < X_5$. One confidence interval of the median m is

$$P(X_1 < m < X_5) = \sum_{k=1}^{4} \mathbf{C_5^k} (\frac{1}{2})^k (\frac{1}{2})^{5-k} = 0.9375$$

$$P(X_2 < m < X_4) = \sum_{k=2}^{3} \mathbf{C_5^k} (\frac{1}{2})^k (\frac{1}{2})^{5-k} = 0.625$$

The interval (X_1, X_n) tends to get wider as n increases, thus we are not "pinning down" m very well. However, if we used the interval (X_2, X_{n-1}) or (X_3, X_{n-2}), we would obtain shorter intervals, but also smaller confidence coefficient.

Distribution-free confidence intervals

As you can see, confidence interval based on order statistic has a prominent shortcoming: we can calculate the confidence coefficient of an interval, but we cannot construct an interval with a pre-specified confidence coefficient.

Distribution-free confidence intervals

As you can see, confidence interval based on order statistic has a prominent shortcoming: we can calculate the confidence coefficient of an interval, but we cannot construct an interval with a pre-specified confidence coefficient.

This approach, therefore, is not widely used in practice. Only use it if there are not other available approach to calculate confidence interval.

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

- Sample from $X_1, X_2, \ldots, X_n$ with replacement;

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

- Sample from $X_1, X_2, \ldots, X_n$ with replacement;
- Calculate the statistic of interest from the sample drawn;

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

- Sample from $X_1, X_2, \ldots, X_n$ with replacement;
- Calculate the statistic of interest from the sample drawn;
- Repeat the above procedures many times to obtain a empirical distribution of the statistic;

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

- Sample from $X_1, X_2, \ldots, X_n$ with replacement;
- Calculate the statistic of interest from the sample drawn;
- Repeat the above procedures many times to obtain a empirical distribution of the statistic;
- Obtain the confidence interval of the statistic from its empirical distribution

Resampling based confidence intervals

Suppose that we need to find the distribution of some statistic, but we do not know its sampling distribution. We observed the values of $X_1, X_2, \ldots, X_n$. The empirical distribution found by placing weight $1/n$ on each X_i is a best estimate of that distribution. A resampling based confidence interval can be constructed using the following steps:

- Sample from $X_1, X_2, \ldots, X_n$ with replacement;
- Calculate the statistic of interest from the sample drawn;
- Repeat the above procedures many times to obtain a empirical distribution of the statistic;
- Obtain the confidence interval of the statistic from its empirical distribution

This approach is also referred to as **bootstrapping.** It allows us to substitute computation for theory for statistical inference.

Resampling based confidence interval

Example: We have a random sample of size 10:

$0.17, -0.27, -1.70, 0.89, -0.14, 0.88, -0.87, 0.25, -1.65, -0.45.$

Use bootstrapping to find the 95% confidence interval of the mean μ.

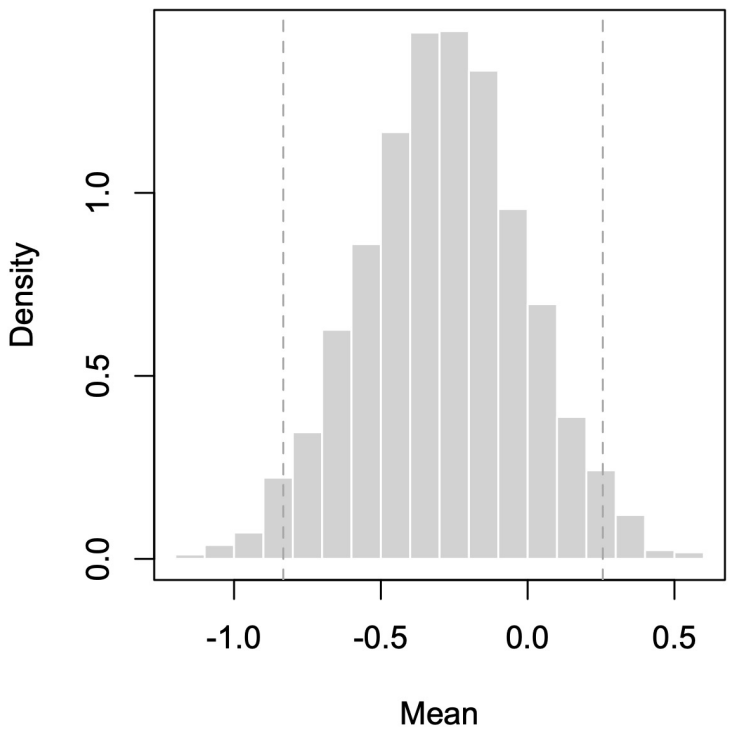
Resampling based confidence interval

Example: We have a random sample of size 10:

$0.17, -0.27, -1.70, 0.89, -0.14, 0.88, -0.87, 0.25, -1.65, -0.45.$

Use bootstrapping to find the 95% confidence interval of the mean μ.

Using 5000 iterations of resampling and the percentile method, we obtain the 95% confidence interval as $(-0.832, 0.256)$.



Resampling based confidence interval

A few comments about resampling based confidence interval

Resampling based confidence interval

A few comments about resampling based confidence interval

- After obtaining the empirical distribution of the statistic, there are alternative methods in addition to the percentile method. The resulting CI can differ depending on which method you choose.

Resampling based confidence interval

A few comments about resampling based confidence interval

- After obtaining the empirical distribution of the statistic, there are alternative methods in addition to the percentile method. The resulting CI can differ depending on which method you choose.
- Resampling approach is effective when the original sample size is big. After all, the method relies on using the empirical distribution of data to approximate the underlying distribution.

Resampling based confidence interval

A few comments about resampling based confidence interval

- After obtaining the empirical distribution of the statistic, there are alternative methods in addition to the percentile method. The resulting CI can differ depending on which method you choose.
- Resampling approach is effective when the original sample size is big. After all, the method relies on using the empirical distribution of data to approximate the underlying distribution.
- Iterations of resampling should be sufficiently large to obtain reliable empirical distribution of the statistic. If computation is not too time consuming, it is better to have large number of iterations just to be safe.