Lecture 7.3

Confidence Intervals: continued

Confidence Interval for mean

Suppose we are seeking to construct a CI for the mean of a population with certain confidence level.

Let us first assume that $X_1, \ldots, X_n$ are independent random variables, each having a $\mathcal{N}(\mu, \sigma^2)$ distribution.

We will see later what we need to do when we relax the normality assumption.

Normal distribution has two parameters, namely μ and σ. Constructing the appropriate CI for μ depends on whether we know σ or not. We treat these two cases separately.

Confidence Interval for mean (known σ^2)

Suppose $X_1, \ldots, X_n$ are independent random variables, each having a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

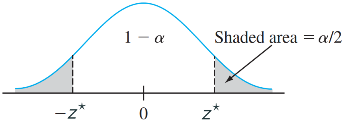
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

This result allows us to construct a $(1 - \alpha) \times 100\%$ confidence interval for μ, assuming σ^2 is known.

Confidence interval for mean (known σ^2)

Let $z^\star$ be the $(1-\alpha/2)$ quantile of the standard normal distribution, i.e.,

$$\mathbb{P}(Z \le z^\star) = 1 - \alpha/2.$$



By symmetry about 0, we also have $\mathbb{P}(Z \le -z^\star) = \alpha/2$. Then

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-z^\star \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z^\star\right) \\ &= \mathbb{P}\left(-z^\star \frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z^\star \frac{\sigma}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(-\bar{X} - z^\star \frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X} + z^\star \frac{\sigma}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\bar{X} - z^\star \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z^\star \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Confidence interval for mean (known σ^2)

When $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the random interval

$$\left[\bar{X} - z^\star \frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + z^\star \frac{\sigma}{\sqrt{n}}\right],$$

is the $(1-\alpha) \times 100\%$ **exact** CI for μ when σ is known.

We say that "*we are* $(1-\alpha) \times 100\%$ *confident*" that the population mean is in the numerical interval

$$\left[\bar{x} - z^\star \frac{\sigma}{\sqrt{n}},\ \bar{x} + z^\star \frac{\sigma}{\sqrt{n}}\right].$$

But what if we don't know σ?

Confidence interval for mean (unknown σ^2)

The random variable

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

has a standard Normal distribution.

If we don't know σ and have to use the estimator of the standard deviation S, where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then, this becomes what is known as the *t-statistic*:

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

What distribution does the random variable T have?

Student's t Distribution

When $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the random variable

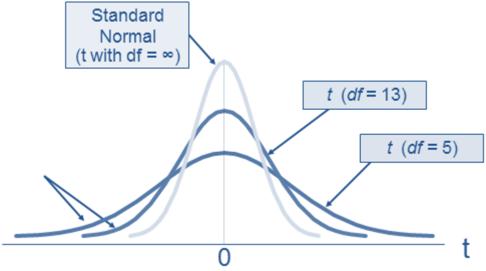
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

has the distribution known as *Student's t distribution* with $n-1$ *degrees of freedom*.

The distribution of T is parametrised by the degrees of freedom. Importantly, the distribution of T does not depend on μ or σ^2.

The random variable T is an example of a **pivot variable**: (a) it depends on all the data and on the parameter to be estimated, but (2) its distribution does not depend on any unknown parameters.

Student's t Distribution



Student's t distribution is

- symmetric,
- bell-shaped, like the normal distribution,
- but has heavier tails than the standard normal distribution,
- does not depend on the parameters of the original normal distribution, i.e., only depends on its degree of freedom.

Confidence interval for mean (unknown σ^2)

Let the critical value $t^\star$ be such that

$$\mathbb{P}\left(T_{n-1} \leq t^\star\right) = 1 - \alpha/2.$$

When $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the random interval

$$\left[\bar{X} - t^\star \frac{S}{\sqrt{n}}\,,\, \bar{X} + t^\star \frac{S}{\sqrt{n}}\right],$$

is the $(1-\alpha) \times 100\%$ **exact** CI for μ when σ is <u>unknown</u>. We say that "*we are* $(1-\alpha) \times 100\%$ *confident*" that the population mean is in the numerical interval

$$\left[\bar{x} - t^\star \frac{s}{\sqrt{n}},\ \bar{x} + t^\star \frac{s}{\sqrt{n}}\right].$$

The quantity $s/\sqrt{n}$ and $t^\star s/\sqrt{n}$ are, respectively, called the standard error and the $(1-\alpha) \times 100\%$ *margin of error*.

Beyond normality assumption

Let $X_1, \ldots, X_n$ be iid from a distribution, not necessarily normal, with mean μ and standard deviation σ (both unknown). When n is large, the CLT coupled with LLN say that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

So the random interval $\left[\bar{X} - z^\star \frac{S}{\sqrt{n}}, \bar{X} + z^\star \frac{S}{\sqrt{n}}\right]$, is the **approximately** $(1 - \alpha) \times 100\%$ CI for μ. In other words, we just put "approximately" in front of the confidence level.

We say that we are approximately $(1 - \alpha) \times 100\%$ *confident* that the population mean is in the numerical interval

$$\left[\bar{x} - z^\star \frac{s}{\sqrt{n}},\ \bar{x} + z^\star \frac{s}{\sqrt{n}}\right].$$

Example

A 2010 study[1] examined the use of video games by Flemish secondary school students aged 12- 20 from over 20 schools.

A sample of 25 male students spent an average of 6.96 hours per week playing video games. The sample standard deviation was 7.42 hours.

Construct a 95% confidence interval for the mean time spent playing video games by the population school aged Flemish males.

[1] Adapted from Bourgonjon et al. (2010) Computes & Education, 54, 1145-1156.

Example

First, we need to find $z^\star$ such that $\mathbb{P}(Z \leq z^\star) = 1 - \alpha/2$ with $\alpha = 0.05$, which gives us $z^\star \approx 1.96$.

So the approximately 95% numerical CI for the mean time spent playing video games is

$$\left[6.96 - 1.96 \times \frac{7.42}{\sqrt{25}}\ ,\ 6.96 + 1.96 \times \frac{7.42}{\sqrt{25}}\right] = [4.05\ ,\ 9.87]\,.$$

Confidence interval for population proportion

Let p denote the proportion of individuals or objects in a population that have a specified property. A random sample of n individuals or objects is to be selected, and X is the number of individuals with that property. The natural estimator of p is $\widehat{P} = X/n$. It is not easy to find an exact confidence interval for p. However, since we can regard X as $X \overset{\text{approx}}{\sim}$ Bin(n,p), if $np \geq 5, n(1-p) \geq 5$, then X has approximately a normal distribution, and

$$\frac{\widehat{P} - p}{\sqrt{p(1-p)/n}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1).$$

We can construct an approximate CI as before but we need to solve for roots of a quadratic equation and we will get a complicated formula...

Confidence interval for population proportion

However, we can actually make our life simpler. From CLT and LLN, we get

$$\frac{\widehat{P} - p}{\sqrt{\widehat{P}(1-\widehat{P})/n}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1).$$

So, the random interval

$$\left[\widehat{P} - z^\star\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}, \widehat{P} + z^\star\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}\right],$$

is the **approximately** $(1-\alpha)\times 100\%$ CI for p, and the corresponding approximate numerical confidence interval is

$$\left[\hat{p} - z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Example

A 2010 study examined the use of video games by Flemish secondary school students aged 12- 20 from over 20 schools. Students were asked if they played video games regularly.

Out of 103 female students, 22 replied that they did not. Construct a 95% confidence interval for the proportion of female students who do not regularly play video games.

Example

First, let's compute the sample proportion:

$$n = 103, \quad \text{and} \quad x = 22 \implies \hat{p} = 22/103.$$

Again, we need to find $z^\star$ such that $\mathbb{P}(Z \leq z^\star) = 0.975$, which gives us $z^\star \approx 1.96$. So, we will have

$$\frac{22}{103} \pm 1.96\sqrt{\frac{22/103 \times (1 - 22/103)}{103}},$$

as our approximately 95% numerical confidence interval.

Beyond normality assumption

Some rules of thumb for how big n should be to use the normal distribution to construct an approximate CI are as follows,

CI for population mean:

- $n < 15$: if the data are close to symmetric.
- $15 \leq n < 40$: if there is no strong skewness in the data.
- $n \geq 40$: generally, this method is justified even in the presence of strong skewness.

CI for population proportion:

- use this approximation when $n\hat{p} \geq 8$ and $n(1 - \hat{p}) \geq 8$.

Beyond normality assumption

In some textbooks, even when $X_1, \ldots, X_n$ are iid from a distribution, not necessarily normal, with mean μ and standard deviation σ (both unknown), the t-statistic is modeled as

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\text{approx}}{\sim} t_{n-1}.$$

In this case, the approximate confidence interval is constructed using the same approach as when assuming normality for samples with an unknown variance.

However, for large enough n, approximating the distribution of the t-statistic with the standard normal or t_{n-1} distribution makes only minor difference.