Lecture 8.1

Confidence Intervals: continued

Confidence intervals: Comparing two populations

Recall: When dealing with one population, we constructed the CI for the mean or the proportion of that population by first choosing an estimator of the mean or the proportion, then considering

$$\frac{\text{Estimator} - \text{True Parameter}}{\text{SD or SE of Estimator}} \sim \text{Some distribution}$$

and finally forming

$$\text{Estimator} \pm c^\star \times (\text{SD or SE of Estimator}),$$

where $c^\star$ is the appropriate critical value of the underlying distribution and it depends on the desired level of confidence.

Confidence intervals: Comparing two populations

We now look at constructing confidence intervals for comparing particular aspects of **two populations**.

Specifically, we will develop confidence intervals for comparing **two population means** and **two population proportions**.

Confidence intervals: Comparing two populations

Throughout, we make the following assumptions:

1. $X_1, \ldots, X_{n_X}$ are iid with mean μ_X and variance σ_X^2,
2. $Y_1, \ldots, Y_{n_X}$ are iid with mean μ_Y and variance σ_Y^2, and
3. All X and Y samples are independent of one another.

Comparing two means

Suppose $X_1, \ldots, X_{n_X}$ is a simple random sample (i.e. independent and identically distributed) from $\mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_{n_Y}$ is a simple random sample from $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

We would like to construct a confidence interval for $\mu_X - \mu_Y$.

Let $\bar{X}$ and $\bar{Y}$ be the respective sample means. The natural estimator for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$.

Under the iid normality assumption on the samples, both $\bar{X}$ and $\bar{Y}$ are also normally distributed, and so is $\bar{X} - \bar{Y}$.

Comparing two means

So, we just need to compute the mean and variance of $\bar{X} - \bar{Y}$:

We have

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}\bar{X} - \mathbb{E}\bar{Y} = \mu_X - \mu_Y,$$

and

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}.$$

So, $\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_X - \mu_Y$ and

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right).$$

Comparing two means (known σ_X and σ_Y)

Hence,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1).$$

As a result, just like before, the random interval

$$(\bar{X} - \bar{Y}) \pm z^\star \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}},$$

is the $(1 - \alpha) \times 100\%$ **exact** CI for $\mu_X - \mu_Y$ when both σ_X and σ_Y are known, where $z^\star$ is the $(1 - \alpha/2)$-quantile of $\mathcal{N}(0, 1)$.

Comparing two means (unknown $\sigma_X = \sigma_Y$)

When σ_X and σ_Y are unknown, we can use their estimator S_X and S_Y instead.

Under the iid normality assumption on the samples, if $\sigma_X = \sigma_Y$, then always

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)}} \sim t_{n_X + n_Y - 2},$$

where

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2},$$

is the **pooled** estimator of the common variance.

Comparing two means (unknown $\sigma_X = \sigma_Y$)

An $(1-\alpha) \times 100\%$ **exact** confidence interval for $\mu_X - \mu_Y$, when $\sigma_X = \sigma_Y$ is unknown, is then

$$(\bar{X} - \bar{Y}) \pm t^\star \sqrt{S_p^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)}$$

where $t^\star$ is the $(1-\alpha/2)$-quantile of the $t_{n_X+n_Y-2}$ distribution.

The numerical CI is also computed similarly.

Comparing two means (unknown $\sigma_X = \sigma_Y$)

Rule of Thumb for Checking if $\sigma_X = \sigma_Y$:

$$\frac{\max\{S_X^2, S_Y^2\}}{\min\{S_X^2, S_Y^2\}} < \begin{cases} 5, & n_x, x_y \approx 7 \\ 3, & n_x, x_y \approx 15 \\ 2, & n_x, x_y \approx 30 \end{cases}$$

Comparing two means (unknown $\sigma_X \neq \sigma_Y$)

But what about when $\sigma_X \neq \sigma_Y$ and they are unknown? Under the iid normality assumption on the samples, does

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}$$

have an exact t-distribution? No ☹

Smith-Satterthwaite Approximation: The distribution of T is approximately a t_ν-distribution with (Welch) degrees of freedom

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1}\left(\frac{s_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1}\left(\frac{s_Y^2}{n_Y}\right)^2}.$$

Comparing two means (unknown $\sigma_X \neq \sigma_Y$)

An **approximately** $(1-\alpha)\times 100\%$ confidence interval for $\mu_X - \mu_Y$, when both σ_X and σ_Y are unknown, is then

$$(\bar{X} - \bar{Y}) \pm t^\star_\nu \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

where $t^\star_\nu$ is the $(1-\alpha/2)$-quantile of the t_ν distribution. The numerical CI is also computed similarly.

`Rstudio` will use the Welch degrees of freedom when constructing a confidence interval for the difference of two means. When constructing this confidence interval by hand, we use the conservative approximation to the degrees of freedom

$$\nu = \min(n_X - 1, n_Y - 1).$$

Example

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

The survey comprised responses from 25 male students and 19 female students. The male students spent an average of 6.96 hours per week playing video games, with a sample standard deviation of 7.42 hours. The 19 female students spent an average of 2.16 hours per week playing video games, with a sample standard deviation of 4.15 hours.

Construct a 99% confidence interval for the population mean difference in time spent playing video games between males and females.

Example

We have

$$n_m = 25, \quad \bar{x}_m = 6.96, \quad s_m = 7.42,$$
$$n_f = 19, \quad \bar{x}_f = 2.16, \quad s_f = 4.15.$$

For ν, we can use $\min\{25-1, 19-1\} = 18$. So,

$$T_{18} = \frac{(\bar{X}_m - \bar{X}_f) - (\mu_m - \mu_f)}{\sqrt{\frac{S_m^2}{n_m} + \frac{S_f^2}{n_f}}} \sim t_{18}.$$

We need 99% confidence interval, so we find the critical value as

$\mathbb{P}(T_{18} \leq t^\star) = 0.995 \implies t^\star = 2.8784$ (In R: `qt(0.995, df = 18)`)

So, we get

$$(6.96 - 2.16) \pm 2.8784 \sqrt{\frac{7.42^2}{25} + \frac{4.15^2}{19}}.$$

Comparing Proportions

Suppose $X_1, \ldots, X_{n_X}$ is a simple random sample (iid) from $\text{Ber}(p_X)$ and $Y_1, \ldots, Y_{n_Y}$ is a simple random sample (iid) from $\text{Ber}(p_Y)$. Further, suppose all X and Y samples are independent of each other.

We would like to construct a confidence interval for $p_X - p_Y$.

Consider the respective sample proportions as

$$\hat{P}_X = \frac{1}{n_X}\sum_{i=1}^{n_X} X_i, \quad \text{and} \quad \hat{P}_Y = \frac{1}{n_Y}\sum_{i=1}^{n_Y} Y_i.$$

Then $\mathbb{E}(\hat{P}_X - \hat{P}_Y) = p_X - p_Y$, and

$$\text{Var}(\hat{P}_X - \hat{P}_Y) = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}.$$

Comparing Proportions

So, by CLT,

$$\frac{\left(\hat{P}_X - \hat{P}_Y\right) - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$$

As for the case of single population, we can replace p_X and p_Y in the denominator by $\hat{P}_X$ and $\hat{P}_X$, respectively, and get that an **approximately** $(1-\alpha) \times 100\%$ CI for $p_X - p_Y$ is

$$\hat{P}_X - \hat{P}_Y \pm z^* \sqrt{\frac{\hat{P}_X(1-\hat{P}_X)}{n_X} + \frac{\hat{P}_Y(1-\hat{P}_Y)}{n_Y}},$$

where z^* is the $(1-\alpha/2)$ quantile of $\mathcal{N}(0,1)$.

Comparing two populations: large samples

Comparing means: If the populations are non-normal, then we can use the previous CIs approximately if the sample sizes are large enough; see slide 17 of Lecture 7.3 and apply to both n_X and n_Y.

- **Populations have the same variance:** We can use the pooled variance estimator and we approximately have a t-distribution with $n_X + n_y - 2$ degree of freedom.

- **Populations have possibly unequal variances:** We can use the Smith-Satterthwaite approximation (or the conservative variant) and have a t-distribution with ν degree of freedom.

Comparing proportions: use the normal approximation when

$$n_X \cdot \min\{\hat{p}_X, 1 - \hat{p}_X\} \geq 8, \quad \text{and} \quad n_Y \cdot \min\{\hat{p}_Y, 1 - \hat{p}_Y\} \geq 8.$$

Summary

Our confidence intervals (in this course) are often of the form

$$\text{Estimator} \pm \text{Multiplier} \times (\text{SD or SE of Estimator}),$$

where the "Multiplier" (t or z) is determined by the level of confidence.