STA2002: Probability and Statistics II

Confidence Interval II

Fangda Song and Ka Wai Tseng

School of Data Science, CUHK(SZ)

September, 2025

Introduction

In this lecture we will continue our discussion on confidence intervals.

Our focus today is on:

- Confidence intervals for difference of two means
- Confidence intervals for proportions
- Sample size determination for a given margin of error

Suggested reading: Chapter 7.2, 7.3 and 7.4 of the textbook.

CI for the Difference of Two Means

- Suppose that we have two populations from which we draw i.i.d. random samples:

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$$

- How to construct confidence intervals for $\mu_X - \mu_Y$, the difference between the two population means?
- Consider three cases:
  1. Pooled t-interval: $X_1, \ldots, X_n$ and $Y_1, \cdots, Y_m$ are independent. $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ but σ^2 is unknown.
  2. Welch's t-interval: $X_1, \ldots, X_n$ and $Y_1, \cdots, Y_m$ are independent. $\sigma_X^2 \neq \sigma_Y^2$ are both unknown.
  3. Paired t-interval: X_i and Y_i are dependent for each i, but pairs (X_i, Y_i) are independent. σ_X^2 and σ_Y^2 are unknown.

Case 1: Pooled t-interval

Theorem

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma^2)$ *and* $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma^2)$ *be independent random variables. Then a* $100(1-\alpha)\%$ *CI for* $\mu_X - \mu_Y$ *is*

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

where S_p^2 *is the pooled estimator of* σ^2*:*

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2}.$$

It can be shown that S_p^2 *is an unbiased estimator of* σ^2*.*

Case 1: Pooled t-interval

Example

Let $X \sim N(\mu_X, \sigma^2)$ be the score on a standardized test in a large high school, and $Y \sim N(\mu_Y, \sigma^2)$ be the score on a standardized test in a small high school. We have

$$n = 9, \quad \bar{x} = 81.31, \quad s_x^2 = 60.76$$
$$m = 15, \quad \bar{y} = 78.61, \quad s_y^2 = 48.24.$$

Solution. To construct a 95% confidence interval for $\mu_X - \mu_Y$, we can obtain $t_{0.025}(22) = 2.074$

$$s_p = \sqrt{\frac{8(60.76) + 14(48.24)}{22}}$$

The two-sided 95% confidence interval for $\mu_X - \mu_Y$ is

$$81.31 - 78.61 \pm 2.074 \sqrt{\frac{8(60.76) + 14(48.24)}{22}} \sqrt{\frac{1}{9} + \frac{1}{15}} = [-3.65, 9.05].$$

Case 2: Welch's t-interval

Theorem (Welch's t-interval)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$ *and* $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$ *be independent. Then an approximate* $100(1-\alpha)\%$ *CI for* $\mu_X - \mu_Y$ *is:*

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}(r)\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}},$$

where

$$r = \left\lfloor \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2} \right\rfloor.$$

Note that $\lfloor x \rfloor$ is the floor function (greatest integer less than or equal to x), say $\lfloor 3.4 \rfloor = \lfloor 3.9 \rfloor = 3$.

Case 2: Welch's t-interval

Example

This time we assume $\sigma_X^2 \neq \sigma_Y^2$. Let $X \sim N(\mu_X, \sigma_X^2)$ be the score on a standardized test in a large high school, and $Y \sim N(\mu_Y, \sigma_Y^2)$ be the score on a standardized test in a small high school. We have

$$n = 9 \quad \bar{x} = 81.31 \quad s_X^2 = 60.76$$

$$m = 15 \quad \bar{y} = 78.61 \quad s_Y^2 = 48.24.$$

Case 2: Welch's t-interval

Solution. To construct a 95% Welch's t-interval for $\mu_X - \mu_Y$, we calculate

$$r = \left\lfloor \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^2}{m}\right)^2} \right\rfloor = 15$$

At the same time,

$$t_{0.025}(r) = 2.131, \qquad \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = \sqrt{\frac{60.76}{9} + \frac{78.61}{15}} = 3.4629$$

The computed 95% Welch's t-interval for $\mu_X - \mu_Y$ is

$$81.31 - 78.61 \pm 2.131 \times 3.4629 = [-4.6794, 10.0794].$$

In comparison, the two-sample pooled t-interval is $[-3.65, 9.05]$.

Case 3: Paired t-interval

Definition: paired samples

Two samples are said to be paired when each data point in the first sample is matched and is related to a unique data point in the second sample.

Paired samples commonly arise in many settings

- Efficacy of a weight loss drug: participants' weights before and after taking the drug
- Efficacy of an online course: participants' grades in the tests before and after taking the course

Example: New hair style for Mr. Trump


Let's try to recommend a new hair style to Mr. Trump, "going chestnut" or "daddy look". Design a study to seek public opinions on these two hair styles.

- Design 1: Randomly select 100 people and show them "going chestnut", and randomly select another 100 people and show them "daddy look". Ask the participants to score the hair style being shown to them.
- Design 2: Randomly select 100 people and show them both "going chestnut" and "daddy look". Ask the participants to score both hair styles being shown to them.

Is there any difference between the two designs? How would you conduct the comparisons?

Case 3: Paired t-interval

Theorem (Paired t-interval)

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be* n *pairs of dependent measurements. Let* $D_i = X_i - Y_i$. *Suppose that* $D_1, \ldots, D_n \overset{i.i.d.}{\sim} N(\mu_D, \sigma_D^2)$ *with* $\mu_D = \mu_X - \mu_Y$. *Then, a* $100(1-\alpha)\%$ *CI for* $\mu_X - \mu_Y$ *is:*

$$\bar{D} \pm t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}},$$

where

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \qquad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2.$$

Case 3: Paired t-interval

Proof: As

$$\bar{D} \sim N(\mu_D, \sigma_D^2/n)$$
$$\frac{(n-1)S_D^2}{\sigma_D^2} \sim \chi^2(n-1),$$

we have

$$\frac{\bar{D} - (\mu_X - \mu_Y)}{S_D/\sqrt{n}} \sim t(n-1)$$

$$\Pr\left(\bar{D} - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}} \leq \mu_X - \mu_Y \leq \bar{D} + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right) = 1-\alpha.$$

Case 3: Paired t-interval

Example

An experiment was conducted to compare people's reaction times to a red light versus a green light. When signaled with either the red or the green light, the subject was asked to hit a switch to turn off the light. When the switch was hit, a clock was turned off and the reaction time in seconds was recorded. The following results give the reaction times for eight subjects:

Subject	Red (x)	Green (y)	$d = x - y$
1	0.30	0.43	–0.13
2	0.23	0.32	–0.09
3	0.41	0.58	–0.17
4	0.53	0.46	0.07
5	0.24	0.27	–0.03
6	0.36	0.41	–0.05
7	0.38	0.38	0.00
8	0.51	0.61	–0.10

Case 3: Paired t-interval

Solution: Let X_i be the reaction time to red light, and Y_i be the reaction time to green light. Based on the data in the table, we have

$$n = 8, \quad \bar{d} = -0.0625, \quad s_d = 0.0765.$$

The 95% CI for $\mu_X - \mu_Y$ is

$$-0.0625 \pm t_{0.025}(7)\left(\frac{0.0765}{\sqrt{8}}\right) = [-0.1265, 0.0015].$$

Case 3: Paired t-interval

We can also build the pooled and Welch's t-interval on these data using $\bar{x} = 0.3700, \bar{y} = 0.4325, s_X = 0.1124, s_Y = 0.1173$

- Pooled t-interval: Using $t_{\alpha/2}(14) = 2.145$ and $s_p = \sqrt{\frac{7s_X^2+7s_Y^2}{14}} = 0.1149$, the 95% CI is given by

$$0.3700 - 0.4325 \pm 2.145\sqrt{\frac{1}{8}+\frac{1}{8}}0.1149 = [-0.1857, 0.0607]$$

- Welch's t-interval: Using $r = \left\lfloor \frac{(s_X^2/8+s_Y^2/8)^2}{\frac{1}{7}(s_X^2/8)^2+\frac{1}{7}(s_Y^2/8)^2} \right\rfloor = 13$, $t_{\alpha/2}(13) = 2.160$ and $\sqrt{\frac{s_X^2}{8}+\frac{s_Y^2}{8}} = 0.0574$, the 95% CI is given by

$$0.3700 - 0.4325 \pm 2.160 \times 0.0574 = [-0.1865, 0.0615]$$

Comparison of CI for difference in means

- Have the same center $\bar{d} = \bar{x} - \bar{y}$
- In general, the CI width:

  Paired t-interval $<$ Pooled t-interval $<$ Welch's t-interval
- When s_X^2 is close to s_Y^2, the widths of pooled t-interval and Welch's t-interval are similar.

Example: New hair style for Mr. Trump


Let's try to recommend a new hair style to Mr. Trump, "going chestnut" or "daddy look". Design a study to seek public opinions on these two hair styles.

- Design 1: Randomly select 100 people and show them "going chestnut", and randomly select another 100 people and show them "daddy look". Ask the participants to score the hair style being shown to them. (Pooled or Welch's t-interval)
- Design 2: Randomly select 100 people and show them both "going chestnut" and "daddy look". Ask the participants to score both hair styles being shown to them. (Paired t-interval)

CI for Proportions

Example

In a political campaign survey, we want to know the supporting rate of two candidates. In the survey, we collect a sample of $n = 351$ voters. $y = 185$ out of them favor candidate A, and the remaining favor candidate B. Let p be the supporting rate in the population. How to build a CI for the supporting rate?

- Let X_i represent the result of voter i. $X_i = 1$ if he/she favors candidate A, and $X_i = 0$ if candidate B.
- Given the supporting rate p, $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p).
- Denote $Y = \sum_{i=1}^{n} X_i$, then $E(Y) = np$ and $\text{Var}(Y) = np(1-p)$.
- if n is large enough, then by central limit theorem (CLT):

$$\frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

CI for Proportions

- Recall that the MLE of p is:

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- Because $\hat{p} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$ for large n:

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

$$\Pr\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha.$$

CI for Proportions

- Since p is unknown, we have

$$\hat{p} \xrightarrow{p} p$$

  for large n, by the law of large number.
- Thus, we use $\hat{p}$ to approximate p when n is large.
- Finally, the approximate $100(1-\alpha)\%$ confidence interval is

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

CI for Proportions

Go back to the example

Example

In a political campaign survey, we want to know the supporting rate of two candidates. In the survey, we collect a sample of $n = 351$ voters. $y = 185$ out of them favor candidate A, and the remaining favor candidate B. Let p be the supporting rate in the population. How to build a CI for the supporting rate?

Solution:

$$\hat{p} = \frac{185}{351} \approx 0.527$$

An approximate 95% CI for p is:

$$0.527 \pm 1.96\sqrt{\frac{0.527(1-0.527)}{351}} = [0.475, 0.579].$$

CI for Proportions

Now consider the difference in proportions between two independent populations instead of one.

- p_i: proportion in population i with a characteristic
- Y_i: number of successes in sample from population i
- n_i: sample size in population i

Since

$$\mathbb{E}(\frac{Y_i}{n_i}) = p_i, \qquad \text{Var}(\frac{Y_i}{n_i}) = \frac{p_i(1-p_i)}{n}$$

Then, we have

$$\mathbb{E}\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) = p_1 - p_2$$

$$\text{Var}\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

For large n_1 and n_2, CLT tells:

$$\hat{p}_1 - \hat{p}_2 \overset{\text{approx}}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

CI for Proportions

- An approximate $100(1-\alpha)\%$ CI for $p_1 - p_2$ is derived from:

$$P\left(-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

- Approximate p_i by $\hat{p}_i$, so the CI is:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

CI for Proportions

Example

Let p_1 be the proportion of male students into gaming, and p_2 be the proportion of female students into gaming. Suppose that $\hat{p}_1 = 0.6$, $\hat{p}_2 = 0.4$. When $n_1 = 2000$, $n_2 = 2200$, please construct an approximate 95% CI for $p_1 - p_2$.

The approximated 95% CI for $p_1 - p_2$ is given by

$$0.6 - 0.4 \pm 1.96\sqrt{\frac{0.6(1-0.6)}{2000} + \frac{0.4(1-0.4)}{2200}} = [0.17, 0.23]$$

Sample Size Determination

- Trade-off in study design
  - Increase sample size: higher cost
  - Decrease sample size: lower precision in the estimation
- To meet a given level of precision, what is the minimum sample size needed?
- Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, σ^2 known. The $100(1-\alpha)\%$ CI for μ is

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

- Half-width $\frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ of a CI is also called **margin of error** or **maximum error of the estimate**, which reflects the precision of the estimate.
- For a given tolerance ϵ in precision, we expect

$$z_{\alpha/2}\sigma/\sqrt{n} \leq \epsilon \Rightarrow n \geq \frac{z_{\alpha/2}^2\sigma^2}{\epsilon^2}$$

Sample Size Determination

When σ^2 is unknown, the $100(1-\alpha)\%$ CI is:

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

Set:

$$t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \epsilon \quad \Rightarrow \quad n \geq \frac{t^2_{\alpha/2}(n-1)s^2}{\epsilon^2}$$

Two issues:

- $t_{\alpha/2}(n-1)$ depends on n: use $z_{\alpha/2}$ as an approximation
- s^2 depends on n: estimate s_p^2 from prior knowledge or pilot study

Finally,

$$n \geq \frac{z^2_{\alpha/2}(s_p)^2}{\epsilon^2}$$

Sample Size Determination

Example

A psychology study uses a standardized memory test with known standard deviation $\sigma = 12$. Find the minimum sample size n needed for a 95% confidence interval with a margin of error at most 2.

We want the margin of error (half-length) of the 95% confidence interval to be at most $E = 2$. For a population mean μ with known standard deviation σ, the margin of error is given by:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2} = 1.96$ for 95% confidence. Given $\sigma = 12$, we set:

$$1.96 \cdot \frac{12}{\sqrt{n}} \leq 2 \Rightarrow n \geq 138.2976$$

Finally, we set a minimum sample size of $n = 139$ students.

Sample Size Determination

- For proportion p, the approximate CI is:

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

- Set the required precision ϵ:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \epsilon \quad \Rightarrow \quad n \geq \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{\epsilon^2}$$

- However, $\hat{p}$ depends on n.

Sample Size Determination

Since $\hat{p}$ is unknown before sampling, two approaches:

1. Use prior estimate p^*:

$$n \geq \frac{z_{\alpha/2}^2 p^*(1-p^*)}{\epsilon^2}$$

2. Use the maximum variance: for $p \in [0,1]$, $p(1-p) \leq \frac{1}{4}$. Thus,

$$n \geq \frac{z_{\alpha/2}^2}{4\epsilon^2}$$

This gives the most conservative (largest) sample size.

Sample Size Determination

Example

A particular area contains 8000 apartment units. In a preliminary survey, 12% of the respondents said they planned to sell their apartments within the next year. We want to estimate the interval for the total number of owners planning to sell. Suppose that a 95% confidence interval with the half-width less than 200 is desired. How many owners do we need to survey?

Sample Size Determination

Solution: The approximate CI of the proportions of owners planning to sell is

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

Because of N units in total, the approximate CI of the number of owners planning to sell is

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

Therefore, the margin of error is

$$Nz_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 200$$

Therefore,

$$1.96N\sqrt{\hat{p}(1-\hat{p})/n} \leq 200$$

Plug the prior estimate $p^* = 0.12$, we have $n \geq 649.1$
Therefore, we need to survey at least 650 apartment owners.