Math 170S

Lecture Notes Section 7.2 [*,†]

Confidence intervals for two means

Instructor: Swee Hong Chan

NOTE: Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

---

[*]Version date: Friday 30th October, 2020, 11:35.

[†]This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Comparing means of X and Y: Case I

- **Object:** X and Y are **independent** random variables with **unknown mean** μ_X and μ_Y but with **known variance** σ_X^2 and σ_Y^2.

- **Input:**

  – Random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ for X and Y, respectively. Note that we have n samples for X but m samples for Y.

  – Confidence constant $1 - \alpha$

- **Output:** The value ε that allows us to say

"$\mu_X - \mu_Y$ is contained in the interval

$\left[(\overline{\mathrm{x}} - \overline{\mathrm{y}}) - \varepsilon, (\overline{\mathrm{x}} - \overline{\mathrm{y}}) + \varepsilon\right]$ with confidence 1-α."

The interval $\left[(\overline{\mathrm{x}} - \overline{\mathrm{y}}) - \varepsilon, (\overline{\mathrm{x}} - \overline{\mathrm{y}}) + \varepsilon\right]$ is the **confidence interval** for $\mu_X - \mu_Y$. This interval is centered at $\overline{\mathrm{x}} - \overline{\mathrm{y}}$, and the length of the interval is 2ε.

2 The error ε: Case I

Theorem 1. *For this case, the value ε is given by*

$$\varepsilon \;=\; z_{\alpha/2}\,\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},$$

where $z_{\alpha/2}$ is the real number such that

$$P[N(0,1) \geq z_{\alpha/2}] \;=\; \alpha/2.$$

The value $z_{\alpha/2}$ can be computed from the Table V in Appendix B. □

3 Example: Case I

Let X be the amount of winning when playing poker, and Y be the amount of winning when playing blackjack.

- The friendly instructor (FI) played poker 15 times, his average profit is 70.1 dollars, and he knows that the variance for this game is 60.

- The FI played blackjack 8 times, his average profit is 75.3 dollars, and he knows that the variance for this game is 40.

Compute the confidence interval for the profit difference between poker and blackjack, with confidence constant equal to 0.90.

4 Answer: Case I

Here we have

- $n = 15$, $\overline{\mathrm{x}} = 70.1$, $\sigma_X^2 = 60$;
- $m = 8$, $\overline{\mathrm{y}} = 75.3$, $\sigma_Y^2 = 40$.

So we have (BT)

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} = z_{0.05}\sqrt{\frac{60}{15} + \frac{40}{8}} \; = \; (1.645)(3) = 4.935.$$

Noting that $\overline{\mathrm{x}} - \overline{\mathrm{y}} = -5.2$, we conclude that the confidence interval is

$$\big[(-5.2) - 4.935, (-5.2) + 4.935\big] \; = \; \big[-10.135, -0.265\big].$$

This means that, 90 out of 100 times, μ_Y is larger than μ_X by at least 0.265. So it is much better if the FI continues playing blackjack.

Remark 2. The estimate for the confidence interval is yet another consequence of the central limit theorem. Please read the textbook for a proof.

5 One-sided confidence interval: Case I

Theorem 3. *For this case, the one-sided confidence interval is given by*

$$\big[(\overline{\mathrm{x}}-\overline{\mathrm{y}})-\varepsilon,\infty\big) \qquad \textit{and} \qquad \big(-\infty,(\overline{\mathrm{x}}-\overline{\mathrm{y}})+\varepsilon\big],$$

and ε *is given by*

$$\varepsilon = z_\alpha\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},$$

where z_α *is the real number such that*

$$P[N(0,1)\geq z_\alpha] \;=\; \alpha.$$

The value z_α *can be computed from the Table V in Appendix B.* □

6 Comparing means of X and Y: Case II

- **Object:** X and Y are **independent normal** random variables with **unknown mean** μ_X and μ_Y and **unknown variance** σ_X^2 and σ_Y^2.

- **Input:**

  - Random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

  - Confidence constant $1 - \alpha$

- **Output:** The value ε that allows us to say

  "$\mu_X - \mu_Y$ is contained in the interval $\left[(\overline{\mathrm{x}} - \overline{\mathrm{y}}) - \varepsilon, (\overline{\mathrm{x}} - \overline{\mathrm{y}}) + \varepsilon\right]$ with confidence 1-α."

7 The error ε: Case II-a

Theorem 4. *If $n + m - 2 \leq 30$, then do this*

1. *Compute the sample mean and the sample variance for X and Y:*

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2; \quad s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{\mathrm{y}})^2.$$

2. *Compute the pooled estimator*

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

3. *The value ε is then given by*

$$\varepsilon = \left[t_{\alpha/2}(n+m-2)\right] \, S_p \sqrt{\frac{1}{n} + \frac{1}{m}},$$

where $t_{\alpha/2}(n+m-2)$ *is the student* t*-distribution from is the real number such that*

$$P[T \geq t_{\alpha/2}(n+m-2)] \;=\; \alpha/2,$$

where T *is the student's* t *distribution with* $n+m-2$ *degrees of freedom, and can be computed from Table VI in Appendix B.* $\square$

8 The error ε: Case II-b

Theorem 5. *If* $n + m - 2 > 30$, *then*

1. *Compute the sample mean and the sample variance for* X *and* Y.

2. ε *is given by*

$$\varepsilon = z_\alpha \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}},$$

where z_α *is the real number such that*

$$P[N(0,1) \geq z_\alpha] \;=\; \alpha.$$

The value z_α *can be computed from the Table V in Appendix B.* □

9 Example: Case II

Let X be the winning made by playing poker, and Y be the winning made by playing blackjack.

- The FI played poker 5 times, his profit (in USD) is

$$21 \quad 22 \quad 23 \quad 24 \quad 25.$$

- The FI played blackjack 3 times, his profit is

$$20 \quad 24 \quad 28.$$

Suppose that both X and Y are **normal** random variables. Compute the confidence interval for the profit difference between poker and blackjack, for confidence constant 0.90.

10 Answer: Case II

We have for X that (BT)

$$\begin{aligned}\bar{\mathrm{x}} =& \frac{21+22+23+24+25}{5} \;=\; 23;\\ s_X^2 =& \frac{1}{4}\left[(21-23)^2+\ldots+(25-23)^2\right] \;=\; \frac{5}{2}.\end{aligned}$$

We have for Y that

$$\begin{aligned}\bar{\mathrm{y}} =& \frac{20+24+28}{3} \;=\; 24;\\ s_Y^2 =& \frac{1}{2}\left[(20-24)^2+(24-24)^2+(28-24)^2\right] \;=\; 16.\end{aligned}$$

The pooled estimator is equal to

$$\begin{aligned}S_p^2 =& \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2} \;=\; \frac{(5-1)(5/2)+(3-1)(16)}{5+3-2}\\ =& 7\end{aligned}$$

So we have

$$\varepsilon = \left[t_{\alpha/2}(n+m-2)\right]\, S_p \sqrt{\frac{1}{n}+\frac{1}{m}} = (1.943)\sqrt{7}\sqrt{\frac{1}{5}+\frac{1}{3}}$$
$$\approx 3.75\,.$$

Noting that $\overline{\mathrm{x}} - \overline{\mathrm{y}} = -1$, we conclude that the confidence interval is

$$\left[(-1) - 3.75, (-1) + 3.75\right] = \left[-4.75, 2.75\right].$$

Since this interval contains 0, we cannot conclude that one game is more profitable than the other game as before.

11 Comparing means of X and Y: Case III

- **Object:** X and Y are (possibly dependent) random variables such that the difference

$$D := X - Y,$$

has **unknown mean** μ_D but **known variance** σ_D^2.

- **Input:**
  - Random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Note that we have n **samples** for X and Y.
  - Confidence constant $1 - \alpha$

- **Output:** The value ε that allows us to say

"$\mu_X - \mu_Y$ is contained in the interval

$\left[(\overline{\mathrm{x}} - \overline{\mathrm{y}}) - \varepsilon, (\overline{\mathrm{x}} - \overline{\mathrm{y}}) + \varepsilon\right]$ with confidence 1-α."

12 The error ε: Case III

Theorem 6. *For this case, the value ε is given by*

$$\varepsilon = z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the real number such that

$$P[N(0,1) \geq z_{\alpha/2}] \;=\; \alpha/2.$$

The value $z_{\alpha/2}$ can be computed from the Table V in Appendix B. □

13 Example: Case III

We want to know if there is a paygap between husband and wife in Atlantis. Let X be the salary of the husband, and Y be the salary of the wife.

From interviewing 100 families, the average husband's salary is 15 cowrie shells, and the average wife's salary is 20 cowrie shells.

We were told that the paygap is a a random variable with **variance** 4 cowrie shells.

Compute the confidence interval for the paygap between husband and wife, with confidence constant 0.90.

14 Answer: Case III

Here $n = 100$, $\overline{\mathrm{x}} = 15$, $\overline{\mathrm{y}} = 20$. So we have

$$\varepsilon = z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}} = z_{0.05}\frac{2}{\sqrt{100}} \;=\; (1.645)(0.2) = 0.329.$$

Noting that $\overline{\mathrm{x}} - \overline{\mathrm{y}} = -5$, we conclude that the confidence interval is

$$\big[(-5) - 0.329, (-5) + 0.329\big] \;=\; \big[-5.329, -4.671\big].$$

This means that, 90 out of 100 times, μ_Y is larger than μ_X by at least 4.671, but not larger than 5.329.

15 Comparing means of X and Y: Case IV

- **Object:** X and Y are (possibly dependent) random variables such that the difference

$$D := X - Y,$$

is a **normal random variable** with **unknown mean** and **unknown variance**.

Input and output are the same as in Case III.

16 The error ε: Case IV

Theorem 7. *In this case, the value ε is computed by:*

- *Compute the sample variance for D:*

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - y_i - (\overline{\mathrm{x}} - \overline{\mathrm{y}}))^2.$$

- *The value ε is then given by*

$$\varepsilon = \begin{cases} t_{\alpha/2}(n-1)\ \frac{s_D}{\sqrt{n}} & \text{if } n-1 \leq 30; \\ z_{\alpha/2}\ \frac{s_D}{\sqrt{n}} & \text{if } n-1 > 30. \end{cases} \quad \square$$

17 Example: Case IV

Let's go back to Atlantis example. From interviewing 8 families, the salary of the husbands are

$$0.30 \quad 0.23 \quad 0.41 \quad 0.53 \quad 0.24 \quad 0.36 \quad 0.38 \quad 0.51.$$

the salary of the wives are

$$0.43 \quad 0.32 \quad 0.58 \quad 0.46 \quad 0.27 \quad 0.41 \quad 0.38 \quad 0.61.$$

The friendly instructor also knows that the paygap is a **normal random variable**.

Compute the confidence interval for the paygap between the husband and the wife, with confidence constant 0.95.

18 Answer: Case IV

Here we have $n = 8$, so the sample mean for D is (BT)

$$\begin{aligned}
\overline{\mathrm{x}} - \overline{\mathrm{y}} =& \frac{1}{8}\left[(0.30 - 0.43) + (0.23 - 0.32) + \ldots + (0.51 - 0.61)\right] \\
=& \frac{1}{8}\Big[(-0.13) + (-0.09) + (-0.17) + (0.07) \\
& + (-0.03) + (-0.05) + (0) + (-0.10)\Big] \\
=& -0.0625,
\end{aligned}$$

The variance for D is

$$\begin{aligned}
s_D^2 =& \frac{1}{7}\left[(-0.13 - (-0.0625))^2 + \ldots + (-0.10 - (-0.0625))^2\right] \\
=& (0.07675)^2.
\end{aligned}$$

So we have

$$\varepsilon = \left[t_{\alpha/2}(n-1)\right] \frac{s_D}{\sqrt{n}} \;=\; (2.365)\frac{0.07675}{\sqrt{8}} \approx 0.06417.$$

Noting that $\overline{\mathrm{x}} - \overline{\mathrm{y}} = -0.0625$, we conclude that the confidence interval is

$$\begin{aligned} &\left[(-0.0625) - 0.06417, (-0.0625) + 0.06417\right] \\ &= \left[-0.12667, 0.00167\right]. \end{aligned}$$

Since this interval contains 0, we cannot conclude who has higher pay as before.