Confidence interval, chapter 4.2

Grethe Hystad

October 15, 2012

Confidence interval for the mean

- Assume X_1, X_2,..., X_n are i.i.d. $N(\mu, \sigma)$, where both μ and σ are unknown.
- We know from chapter 4.1 that

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

has a t-distribution with $r = n - 1$ degrees of freedom.

Confidence interval for the mean

Select $t_{\alpha/2}(n-1)$ so that $P[T \geq t_{\alpha/2}(n-1)] = \frac{\alpha}{2}$. Then

$$1-\alpha = P\left[-t_{\alpha/2}(n-1) \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right]$$

and hence

$$1-\alpha = P\left[\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right].$$

Then if x_1,...,x_n are the observed values,

$$\left[\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \quad \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

Confidence interval for the mean

$$m = t_{\alpha/2}(n-1) \times \frac{s}{\sqrt{n}}$$

is the **margin of error**.

Example

The Trial Urban District Assessment (TUDA) is a study sponsored by the government of student achievement in large urban school district. The math test-score is on a scale from 0 to 500. A "basic" math level is a score of 262, a "proficient" level is a score of 299 and a "advanced" level is a score of 333. In 2007, a random sample of 2000 *eighth*-graders from Los Angeles had an average math scale score of $\overline{x} = 257$ with a standard deviation of 49.19. (The study reports the standard error of the mean instead of the standard deviation.) Source: TUDA results for 2007 from the National Center for Education Statistics, at nces.ed.gov/nationsreportcard

(A) Give a 95% confidence interval for the mean score of all LA eighth graders.
(B) Give a 99% confidence interval for the mean score of all LA eighth graders

Solution

Since the range of scores is between 0 *and* 500, *it is limited how skewed the distribution can be. The sample size is also large, so we can use the student t-distribution.*

Recall the $100(1-\alpha)\%$ *confidence interval for the mean:*

$\left[\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \quad \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right]$. *(A) R-code:*

$qt(0.975, 1999)$

[1] 1.961151

so

$t_{0.025}(1999) = 1.9612$.

We have $\frac{s}{\sqrt{n}} = \frac{49.19}{\sqrt{2000}} = 1.1$

A 95% *confidence interval for the mean is:*

$\overline{x} \pm \frac{s}{\sqrt{n}} t_{0.025}(1999) = 257 \pm 2.15732$ *which is* $(254.84, \ 259.16)$.

Solution continue

Solution

(B) We have $t_{0.005}(1999) = 2.5783$.

A 99% confidence interval for the mean is:
$\overline{x} \pm \frac{s}{\sqrt{n}} t_{0.025}(1999) = 257 \pm 2.83613$ *which is* $(254.16,\ \ 259.84)$.

Confidence interval for the difference of the mean

Let X_1, X_2,..., X_n and Y_1, Y_2,..., Y_m be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively, where μ_X and μ_Y are unknown. We will consider three situations:

- Case 1. σ_X^2, σ_Y^2 are unknown with $\sigma_x^2 = \sigma_y^2$ and with n and m small.
- Case 2. σ_X^2, σ_Y^2 are known or the sample sizes are large.
- Case 3. σ_X^2, σ_Y^2 are unknown and σ_x^2 is very different from σ_y^2 and the sample sizes, n, m, are small and different.

Confidence interval for the difference of the mean, Case 1

Case 1.

Let X_1, X_2,..., X_n and Y_1, Y_2,..., Y_m be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively. Assume σ_X^2 and σ_Y^2 are **unknown** with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Let $W = \overline{X} - \overline{Y}$ which is also a normal random variable.

Confidence interval for the difference of the mean, Case 1

Case 1.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively. Assume σ_X^2 and σ_Y^2 are **unknown** with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Let $W = \overline{X} - \overline{Y}$ which is also a normal random variable.
- Then $E(W) = \mu_X - \mu_Y$ and $Var(W) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$ so

Confidence interval for the difference of the mean, Case 1

Case 1.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively. Assume σ_X^2 and σ_Y^2 are **unknown** with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Let $W = \overline{X} - \overline{Y}$ which is also a normal random variable.
- Then $E(W) = \mu_X - \mu_Y$ and $Var(W) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$ so
- $Z = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$ is $N(0, 1)$.

Confidence interval for the difference of the mean, Case 1

Case 1.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively. Assume σ_X^2 and σ_Y^2 are **unknown** with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Let $W = \overline{X} - \overline{Y}$ which is also a normal random variable.
- Then $E(W) = \mu_X - \mu_Y$ and $Var(W) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$ so
- $Z = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$ is $N(0, 1)$.
- Since the two samples are independent, $U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}$ is $\chi^2(n + m - 2)$.

Confidence interval for the difference of the mean, Case 1

Case 1.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two samples of i.i.d. random variables from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ respectively. Assume σ_X^2 and σ_Y^2 are **unknown** with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

- Let $W = \overline{X} - \overline{Y}$ which is also a normal random variable.
- Then $E(W) = \mu_X - \mu_Y$ and $Var(W) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$ so
- $Z = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$ is $N(0, 1)$.
- Since the two samples are independent, $U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}$ is $\chi^2(n + m - 2)$.
- The independence of the sample means and sample variance implies that Z and U are independent.

Confidence interval for the difference of the mean, Case 1

We know that

$$T = \frac{Z}{\sqrt{\frac{U}{n+m-2}}}$$

has a t-distribution with $n + m - 2$ degrees of freedom. We can write

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\left[\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\right](\frac{1}{n} + \frac{1}{m})}}$$

Confidence interval for the difference of the mean, Case 1

Select $t_0 := t_{\alpha/2}(n+m-2)$ so that $P(-t_0 \leq T \leq t_0) = 1-\alpha$.
Then

$$P\left(\overline{X}-\overline{Y}-t_0 S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \leq \mu_X-\mu_Y \leq \overline{X}-\overline{Y}+t_0 S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = 1-\alpha,$$

where the pooled estimator of the common standard deviation is

$$S_p = \sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}$$

Confidence interval for the difference of the mean, Case 1

If $\overline{x}$, $\overline{y}$ and s_p are the observed values of $\overline{X}$, $\overline{Y}$ and S_p, then

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}},$$

where

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}},$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

Confidence interval for the difference of the mean, Case 2

Case 2
If σ_X^2 and σ_Y^2 are known then

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \quad \text{is} \quad N(0,1).$$

If the sample sizes are large so that $s_X^2 \approx \sigma_X^2$ and $s_Y^2 \approx \sigma_Y^2$, then

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \quad \text{is approximate} \quad N(0,1).$$

Confidence interval for the difference of the mean, Case 2

The respectively $100(1-\alpha)\%$ confidence intervals are:

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \quad \text{and}$$

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

Confidence interval for the difference of the mean, Case 3

Case 3. If the underlying distributions are close to normal but the sample sizes and the variances are different, then

$$V = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

has an **approximate** t-distribution with $\lfloor v \rfloor$ degrees of freedom, where, v, is the Welch-Satterthwaite equation,

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

$\lfloor v \rfloor$ is the greatest integer smaller than or equal to v.

Confidence interval for the difference of the mean, Case 3

A $100(1-\alpha)\%$ confidence interval of the difference in the mean is then given by

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(\lfloor v \rfloor)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

Example

Suppose a random sample of 50 students from high school A has an average SAT score of 510 with a standard deviation of 100 and a random sample of 70 students from high school B has an average SAT score of 490 with a standard deviation of 103. Assume both samples come from a Normal distribution in which their variances are approximately equal. Compute a 98% confidence interval for the difference in the mean between these two schools.

Solution

We have the following values:

Group	n	$\overline{x}$	s
A	50	510	100
B	70	490	103

Then $s_p = \sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}} = \sqrt{\frac{(50-1)100^2+(69)(103^2)}{50+70-2}} = 101.77$
The number of degrees of freedom is $50 + 70 - 2 = 118$ *and* $t_{0.01}(118) = 2.3584$. *We then have*
$\overline{x} - \overline{y} \pm t_{\alpha/2}(n + m - 2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}} =$
$(510 - 490) \pm 2.36 * 101.77 * \sqrt{\frac{1}{50} + \frac{1}{70}} = 20 \pm 44.47$. *So a* 98% *confidence interval for the difference in the mean SAT score is*
$(-24.47, \quad 64.47)$

Example

A study compared the education program for preschool children that implemented the Montessori method with other schools. The test-results of 5 year old that had been enrolled in preschool programs from the age of 3 in Milwaukee Wisconsin were compared. The students were assigned into Montessori school at age 3 by a random lottery.
Source: Evaluating Montessori Education by A. Lillard and N. Else-Quest, Science 313, 1893 (2006).

(A) The families named their school of choice in the application. Explain why comparing children whose families choose to participate in the Montessori school with children whose families choose to participate as the control group might cause a bias in the result. (All the students in the study listed the Montessori school as their first choice.)

Example continues

Example

(B) One of the test administered was testing the students ability to apply mathematics to solve problems. Here are the results:

School	sample size	mean	standard deviation
Montessori	30	19	3.11
Control	25	17	4.19

Construct a 95% confidence interval for the difference in the mean of the test score of the Montessori group and the control group. What is your conclusion?

Solution

(A) The parents of the two populations of students might have different attitudes about education.

(B) We use here the Welch-Satterthwaite equation,

$$\lfloor v \rfloor = \frac{\left(\frac{(3.11)^2}{30} + \frac{(4.19)^2}{25}\right)^2}{\frac{1}{29}\left(\frac{(3.11)^2}{30}\right)^2 + \frac{1}{24}\left(\frac{(4.19)^2}{25}\right)^2} = \lfloor 43.51 \rfloor = 43.$$

> *qt(0.975,43)*

[1] *2.02*

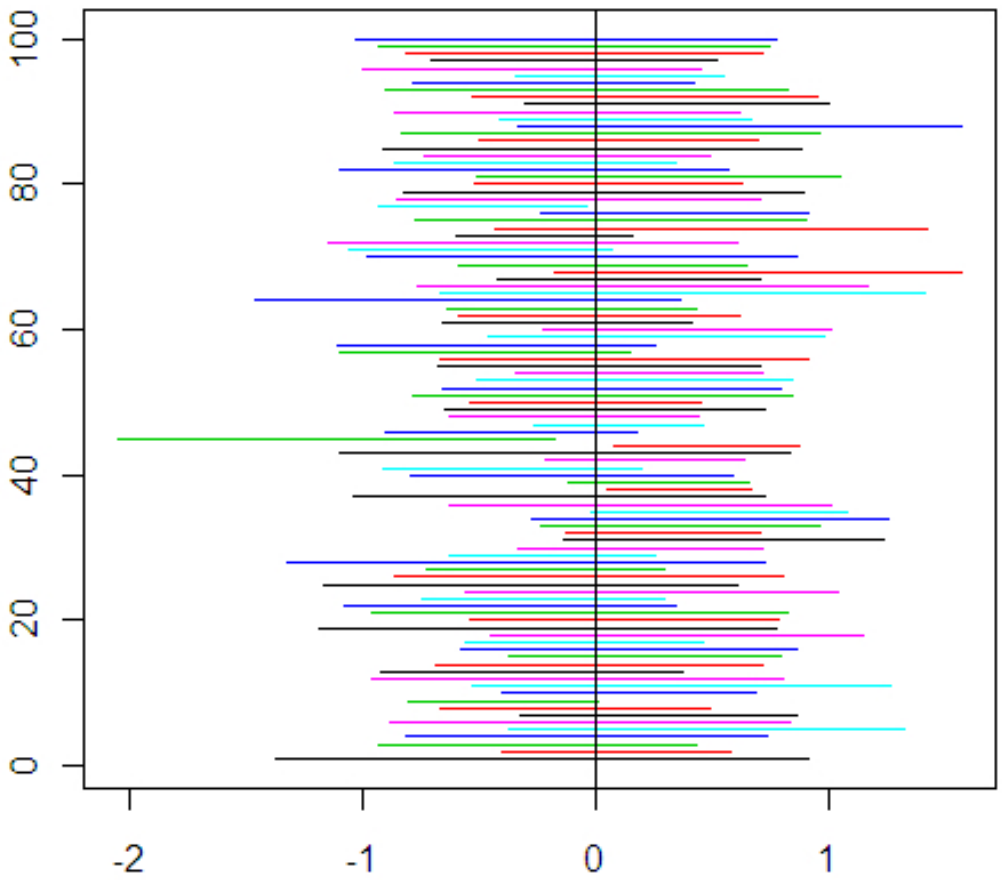
Hence $t_{0.025}(43) = 2.02$.

A 95% confidence interval for the difference in the mean is:

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(\lfloor v \rfloor)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = (19 - 17) \pm 2.02\sqrt{\frac{(3.11)^2}{30} + \frac{(4.19)^2}{25}}.$$

Thus, a 95% confidence interval is $(-0.04,\ \ 4.04)$. *Since* 0 *is not in the interval, we conclude that there is no significant difference in the mean of the test score of the Montessori group and the control group.*

10 samples from $N(0,1)$ were simulated 100 times. The 100 95%-confidence intervals for the mean are shown below: We see that 96% of the intervals contains the population mean $\mu = 0$. The margin or error is $t_{0.025}(9) * \text{standard deviation}/\sqrt{10}$



Margin of error

Definition

Margin of error, m**,** is defined as

$$m = t_{\alpha/2}(r) \times \text{ standard error},$$

where $t_{\alpha/2}(r)$ is the critical value for the t- distribution for a certain number of r degrees of freedom.

Confidence intervals are on the form,

$$\text{Point estimate } \pm \text{ margin of error.}$$

Margin of error

The margin of error decreases if

- the standard deviation decreases
- the sample size n increases
- the confidence level, $(1 - \alpha) \times 100\%$ decreases.

Margin of error

Since the margin of error decreases, when the sample size, n, increases, we will find a sample size that will give a margin of error that we want.
Lets go back to the first example: We have $m = t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$ which gives $m\sqrt{n} = t_{\alpha/2}(n-1) \cdot s$. We obtain

$$n \approx \left(\frac{t_{\alpha/2}(n-1) \cdot s}{m} \right)^2$$

We want to guarantee a margin of error of math scale score of 1.8 with 95% confidence. We set $t_{0.025}(n-1) = 2$ and $s = 50$ which are overestimates. Thus the margin of error will then be smaller or equal to 1.8. We obtain a sample size of $n = \left(\frac{2 \cdot 50}{1.8} \right)^2 \approx 3087$ to guarantee a margin of error of 1.8.