MTMS.01.099 Mathematical Statistics

Lecture 10

Tõnu Kollo

Fall 2016

Reminder: A Single Random Sample. Confidence Interval for the **Mean**

Let $x_1, \ldots, x_n$ be a sample from $N(\mu, \sigma^2)$. We want to construct a confidence interval for the mean μ.
We have the following result

Theorem 3

Let $x_1, \ldots, x_n$ be a sample from $N(\mu, \sigma^2)$, where μ is unknown. Then

$$I_\mu = \overline{x} \pm \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \text{if} \quad \sigma \quad \text{known}$$

$$I_\mu = \overline{x} \pm t_{\alpha/2}(f) \frac{s}{\sqrt{n}}, \quad \text{if} \quad \sigma \quad \text{unknown},$$

where s is a sample standard deviation , $\lambda_{\alpha/2}$ and $t_{\alpha/2}(f)$ are $\alpha/2$ complement quantiles of $N(0,1)$ and $t(f)$, $f = n - 1$.

Confidence Interval for the Mean when σ is known. Remark

Remark

- Note that the **width of the interval**, given by $\lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, depends on the confidence level (which determines $\lambda_{\alpha/2}$), the standard deviation, and the sample size
- One cannot control σ, but can adjust $\lambda_{\alpha/2}$ or n
- To make the confidence interval narrower, you can either increase the sample size n or decrease the value of the quantile $\lambda_{\alpha/2}$, which amounts to decreasing the confidence level

Choice of Sample size

If we think of the width of the interval as specifying its precision or accuracy, then the confidence level (or reliability) of the interval is inversely related to its precision.

A highly reliable interval estimate may be imprecise in that the endpoints of the interval may be far apart, whereas a precise interval may entail relatively low reliability. Thus it cannot be said unambiguously that a 99% interval is to be preferred to a 95% interval; **the gain in reliability entails a loss in precision**.

An appealing strategy is to specify both the desired confidence level and interval width and then determine the necessary sample size.

Choice of Sample size (2)

The general formula **for the sample size** n **necessary to ensure an interval width** w is obtained from

$$w = \overline{x} + \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} - (\overline{x} - \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 2\lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

From this equation you can find n as

$$n = \left(2\lambda_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

- Note that, if we find that a confidence interval is too long and want an interval half as long, we must make the sample *four times* larger. Hence we must collect three times as many observations as we already have, and then compute a new confidence interval based on all the data.
- The smaller the desired width w, the larger n must be!

Choice of Sample size. Example

Example

Extensive monitoring of a computer time-sharing system has suggested that response time to a particular editing command is normally distributed with standard deviation 25 ms.
A new operating system has been installed, and we wish to estimate the true average response time μ for the new environment.
Assuming that response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10?

Choice of Sample size. Example

The sample size n must satisfy

$$10 = 2 \cdot 1.96 \cdot \left(\frac{25}{\sqrt{n}} \right)$$

Rearranging this equation gives

$$n = \left(2 \cdot 1.96 \cdot \frac{25}{10} \right)^2$$

so

$$n = 9.80^2 = 96.04$$

Since n must be an integer, a sample size of 97 is required.

A Single Sample. Confidence Interval for the **Standard Deviation**

Theorem

Let $x_1, \ldots, x_n$ be a sample from $N(\mu, \sigma^2)$. Then

$$I_\sigma = (k_1 s, k_2 s),$$

where

$$k_1 = \sqrt{\frac{f}{q_{\alpha/2}(f)}}, \quad k_2 = \sqrt{\frac{f}{q_{1-\alpha/2}(f)}}, \quad f = n - 1,$$

is a confidence interval for σ with a confidence level $1 - \alpha$ and $q_{\alpha/2}(f)$ is $\alpha/2$-complement quantile of $\chi^2(f)$.

A Single Sample. Confidence Interval for the **Standard Deviation**. Proof

Proof.
Consider s^2 to be a point estimate of σ^2, e.g. $\hat{\sigma}^2 = s^2$. We know the corresponding point estimator is

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

where $X_i \sim N(\mu, \sigma^2)$ are independent random variables. Recall from previous lecture the following result

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(f), \quad f = n-1, \quad (*)$$

and from $(*)$ we obtain $\frac{n-1}{\sigma^2}S^2 \sim \chi^2(f)$.

A Single Sample. Confidence Interval for the **Standard Deviation**. Proof (2)

It follows from the definition of $\chi^2(f)$ that the probability mass between $q_{1-\alpha/2}(f)$ and $q_{\alpha/2}(f)$ is $1-\alpha$. Hence

$$P(q_{1-\alpha/2}(f) < \frac{f}{\sigma^2}S^2 < q_{\alpha/2}(f)) = 1-\alpha.$$

So we get

$$P\left(\frac{fS^2}{q_{\alpha/2}(f)} < \sigma^2 < \frac{fS^2}{q_{1-\alpha/2}(f)}\right) = 1-\alpha.$$

Now replacing $S^2 = s^2$ (substituting a value from the sample) and taking the square roots of the limits (the probability will not change), we obtain the confidence interval for σ with confidence level 1-α.

A Single Sample. Confidence Interval for the **Standard Deviation**. Remarks

Remarks

In the book of G. Blum there is a formula (p. 235) for large n:

$$k_1 = 1 - \lambda_{\alpha/2}\frac{1}{\sqrt{2f}}, \quad k_2 = 1 + \lambda_{\alpha/2}\frac{1}{\sqrt{2f}}$$

Unfortunately, this is not correct. The normal approximation is generally good enough if n exceeds 30. It has the advantage that no special table is needed, only the usual table of quantiles of the normal distribution.

- If n is small, the confidence interval for σ is unfortunately rather long, which is inevitable. Hence the information obtained about σ is not very accurate. It is generally of no use calculating the interval unless n is at least 20.

A Single Sample. Confidence Interval for the **Standard Deviation**. Example

Example

Consider the weight of a middle-aged man to be a random variable with $N(\mu, \sigma^2)$. Based on some research it is known that $s = 14.8$kg (29 men were included in the study). Find a CI for σ at the confidence level 0.95.

A Single Sample. Confidence Interval for the **Standard Deviation**. Example

Example

Consider the weight of a middle-aged man to be a random variable with $N(\mu, \sigma^2)$. Based on some research it is known that $s = 14.8$kg (29 men were included in the study). Find a CI for σ at the confidence level 0.95.

χ^2-quantiles are $q_{0.025}(28) = 44.5$ and $q_{0.975}(28) = 15.3$, so that

$$k_1 = \sqrt{\frac{28}{44.5}} = 0.8, \quad k_2 = \sqrt{\frac{28}{15.3}} = 1.35.$$

Therefore

$$I_\sigma = (k_1 s, k_2 s) = (0.8 \times 14.8, 1.35 \times 14.8) = (11.84, 19.98)$$

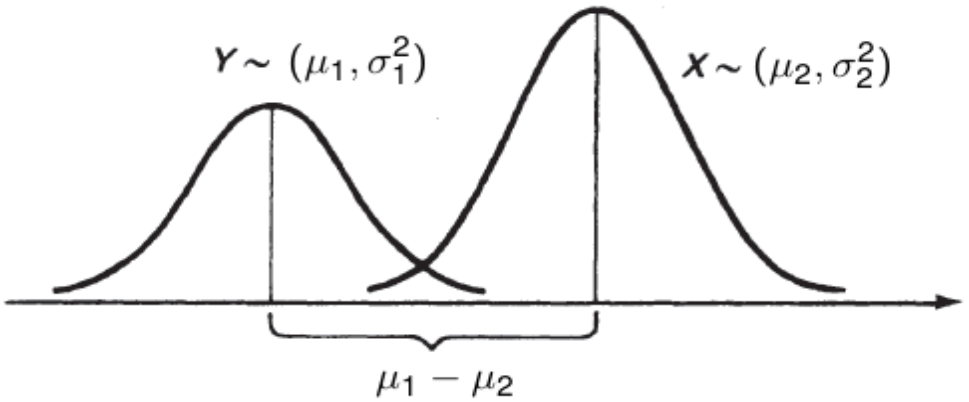
Two Samples. Confidence Interval for **Difference Between Means**

The following model is often employed in practice: Two independent samples have been collected,

$$x_1, \ldots, x_{n_1} \quad \text{from} \quad N(\mu_1, \sigma_1^2)$$

$$y_1, \ldots, y_{n_2} \quad \text{from} \quad N(\mu_2, \sigma_2^2)$$

A confidence interval for the difference $\mu_1 - \mu_2$ is required



Two Samples. Confidence Interval for **Difference Between Means**

- Is the average income in Estonia higher than the average income in Latvia?
- Is there difference in average incomes of men versus women?
- Is there difference in average age of Republicans versus Democrats?
- A producer might wish to estimate the difference in mean daily output from two machines.
- A medical researcher might wish to estimate the difference in mean response by patients who are receiving two different drugs.
- . . .

$\Rightarrow$ You need to estimate the difference between two population means.

Two Samples. Confidence Interval for **Difference Between Means**

Theorem

*Let $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ be independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. If σ_1 and σ_2 are **known**, then*

$$I_{\mu_1-\mu_2} = \overline{x} - \overline{y} \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

*is a two-sided confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$. If $\sigma_1 = \sigma_2 = \sigma$, where σ is **unknown**, then*

$$I_{\mu_1-\mu_2} = \overline{x} - \overline{y} \pm t_{\alpha/2}(f)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is a two-sided CI for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$, where $s^2 = \frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2+\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_1+n_2-2}$ and $f = n_1 + n_2 - 2$.

Two Samples. Confidence Interval for Difference Between Means. Example

Example

Will new directed reading activities improve certain aspects of a child's reading ability? An educator conducted an experiment to test this. Twenty-one third graders took part in these directed reading activities for 8 weeks, while another class of 23 third graders did not. At the end of the study, all children took the Degree of Reading Power (DRP) test, a standard test that measures various aspects of reading ability.
The mean DRP score for the students who received the treatment was 51.48, while the mean for the control group was 41.52. Though the difference seems large, it may well be owing to sampling variability. Of course, in practice, we usually do not know the population variances, but in this case the standard deviations of the populations were known and assumed to be 11.01 and 17.14, respectively.

Two Samples. Confidence Interval for Difference Between Means. Example (2)

For the $n_1 = 21$ third graders who participated in directed reading activities, the mean and the standard deviation of their DRP scores are $\overline{x} = 51.48, \sigma_1 = 11.01$, and for the $n_2 = 23$ third graders in the control group, $\overline{y} = 41.52, \sigma_2 = 17.14$. Hence

$$\begin{aligned} I_{\mu_1-\mu_2} &= \overline{x} - \overline{y} \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 9.96 \pm \lambda_{\alpha/2}\sqrt{\frac{11.01^2}{21} + \frac{17.14^2}{23}} \end{aligned}$$

The 0.975 quantile of the $N(0,1)$ distribution is 1.96. Thus, the 95% confidence interval is
$9.96 \pm 1.96 \cdot 4.31 = 9.96 \pm 8.45 = (1.51, 18.41)$. With 95% confidence, third graders who participated in directed reading activities score, on average, got from 1.51 to 18.41 points higher than third graders who did not participate.

Two Samples. Confidence Interval for Difference Between Means

Remark

If $\sigma_1^2 \neq \sigma_2^2$ and both of them are unknown, we need to estimate them from the sample. Then for a two-sided confidence interval for $\mu_1 - \mu_2$ we will use

$$I_{\mu_1-\mu_2} \approx \bar{x} - \bar{y} \pm t_{\alpha/2}(f)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where

$$f = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \right\rfloor$$

and therefore the confidence level is now $\approx 1 - \alpha$.

Two Samples. Confidence Interval for Difference Between Means. Conclusions

- If $I_{\mu_1-\mu_2}$ is set on the positive side of real axis, then $\mu_1 > \mu_2$ with confidence level $1-\alpha$.
- If $I_{\mu_1-\mu_2}$ is set on the negative side of real axis, then $\mu_1 < \mu_2$ with confidence level $1-\alpha$.
- If $I_{\mu_1-\mu_2}$ contains 0, then we cannot rule out the possibility that the means might be the same, $\mu_1 - \mu_2 = 0$ or, equivalently, $\mu_1 = \mu_2$ (we cannot say which of the values μ_1 or μ_2 is larger).

Paired Samples

Dependent (paired) samples - repeated measurements on **the same** object (for example, before and after using some treatment, medicament, etc).

$$\text{Before}: \; x_1, \ldots, x_n; \; x_i \leftarrow X_i \sim N(\mu_i, \sigma_1^2);$$

$$\text{After}: \; y_1, \ldots, y_n; \; y_i \leftarrow Y_i \sim N(\mu_i + \triangle, \sigma_2^2);$$

- derive a new variable $Z_i = Y_i - X_i \sim N(\triangle, \sigma_z^2)$, where $\sigma_z^2 \neq \sqrt{\sigma_1^2 + \sigma_2^2}$
- Theorem 3 (from previous lecture) is used (where both of the parameters are unknown):

$$I_\mu = \overline{x} \pm t_{\alpha/2}(f)\frac{s}{\sqrt{n}}, \quad \text{if} \;\; \sigma \;\; \text{unknown},$$

Paired Samples

Theorem

Confidence interval for difference between means in case of dependent samples is following

$$I_{\triangle} = \overline{z} \pm t_{\alpha/2}(n-1)\frac{s_z}{\sqrt{n}},$$

where

$$s_z = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2}.$$

Paired Samples. Example

Example

There is new food supplement on the market which is meant for lowering weight. Researcher selected 10 persons to participate in experiment to find out how good the supplement actually is. The weight of participants was measured in the beginning of the experience (before taking the supplements) and in the end of experiment (after 15 weeks). This is the data:

Subject	Initial Weight	Final Weight	Difference
1	180	165	15
2	142	138	4
3	126	128	-2
4	138	136	2
5	175	170	5
6	205	197	8
7	116	115	1
8	142	128	14
9	157	144	13
10	136	130	6

Paired Samples. Example(2)

Mean difference $= \frac{66}{10} = 6.6$

Variance of difference $= \frac{\Sigma(z_i - \bar{z})^2}{n-1} = \frac{304.4}{9} = 33.82$

Degrees of freedom, $f = n - 1 = 9$

And $t_{(0.025,9)} = 2.262$

The 95% C.I. for the difference becomes:

$$6.6 \pm 2.262 \times \frac{\sqrt{33.82}}{\sqrt{10}}$$
$$= 6.6 \pm 4.2$$

That is, $2.4 < \mu_z < 10.8$