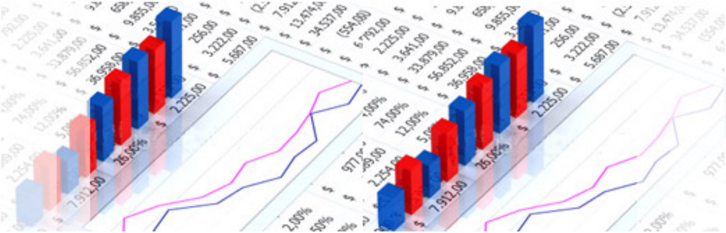
MTMS.01.099 Mathematical Statistics

Lecture 11. Interval Estimation. Hypothesis testing

Tõnu Kollo



Fall 2016

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

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

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

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

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

Choice of Sample size

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

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

- The smaller the desired width w, the larger n must be!

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

Theorem

Let $x_1, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$. Then confidence interval for σ is

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

where

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

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

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 $f = n_1 + n_2 - 2$.

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

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

Using the Normal Approximation. Confidence Interval for a nonnormal distribution

- In practice, distributions often occur that are not at all like the normal
- Nonnormal distributions can often be handled by means of the same technique as normal ones.
  - Remarkably enough, it can be shown that several of the methods described in the previous lectures are approximately true even if the distribution is nonnormal.
  - A prerequisite is, however, that **the samples are reasonably large!**
- The idea: the normal approximation is applied to the estimate one is interested in

Confidence Interval for a nonnormal distribution

Let us consider a random sample from an **unspecified distribution** (or, more generally, several samples from different distributions).

The distribution depends on an **unknown parameter** θ.

Assume that we have found a point estimate $\hat{\theta}$ (so the corresponding point estimator is $\hat{\boldsymbol{\theta}}$) which is approximately normally distributed with mean μ and standard deviation $\sigma_{\hat{\theta}} = \sqrt{Var(\hat{\theta})}$.

By Theorem 1 in Ch. 8.4 we then have

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{Var(\hat{\boldsymbol{\theta}})}} \approx N(0, 1).$$

Confidence Interval for a nonnormal distribution (2)

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sigma_{\hat{\theta}}} \approx N(0, 1)$$

- If the standard deviation $\sigma_{\hat{\theta}}$ **does not depend** on θ, we obtain the confidence interval $(\hat{\theta} \pm \lambda_{\alpha/2}\sigma_{\hat{\theta}})$ with approximate confidence level $1 - \alpha$.
- If the standard deviation $\sigma_{\hat{\theta}}$ should **depend** on θ (and perhaps on additional parameters), $\sigma_{\hat{\theta}}$ is usually replaced by some suitable standard error *d*, confidence level being still approximately equal to $1 - \alpha$. However, the approximation is often less accurate than before.

Confidence Interval for a nonnormal distribution (3)

Theorem

Assume that a point estimate $\hat{\theta}$ of θ is based on one or several random samples. Further, assume that the corresponding estimator is approximately normally distributed with mean θ and standard deviation $\sigma_{\hat{\theta}}$. Then

$$I_\theta = (\hat{\theta} - \lambda_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + \lambda_{\alpha/2}\sigma_{\hat{\theta}}) \quad \text{if } \sigma_{\hat{\theta}} \text{ does not depend on } \theta,$$

$$I_\theta = (\hat{\theta} - \lambda_{\alpha/2}d, \hat{\theta} + \lambda_{\alpha/2}d) \quad \text{if } \sigma_{\hat{\theta}} \text{ depends on } \theta,$$

are confidence intervals for θ with approximate confidence level $1 - \alpha$.

The approximation in this theorem is generally better the larger the samples.

"Precise" confidence intervals

Distribution	**Parameter**	**Intervals**	**Remarks**
$N(\mu,\sigma^2)$	μ	$\bar{x} \pm \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ $(-\infty; \bar{x} + \lambda_\alpha \cdot \frac{\sigma}{\sqrt{n}})$ $(\bar{x} - \lambda_\alpha \cdot \frac{\sigma}{\sqrt{n}}; \infty)$	σ^2 is known $\lambda_{\alpha/2}$ is $N(0,1)$ complement quantile
$N(\mu,\sigma^2)$	μ	$\bar{x} \pm t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}$ $(-\infty; \bar{x} + t_\alpha(n-1) \cdot \frac{s}{\sqrt{n}})$ $(\bar{x} - t_\alpha(n-1) \cdot \frac{s}{\sqrt{n}}; \infty)$	σ^2 is unknown $t_{\alpha/2}(n-1)$ is $t(n-1)$ complement quantile
$N(\mu,\sigma^2)$	σ^2	$\left(\frac{n-1}{q_{\alpha/2}(n-1)} s^2; \frac{n-1}{q_{1-\alpha/2}(n-1)} s^2 \right)$	$q_\alpha(n\text{-}1)$ is $\chi^2(n-1)$ complement quantile
$N(\mu_1,\sigma_1^2)$ $N(\mu_2,\sigma_2^2)$	$\mu_1 - \mu_2$	$\bar{x} - \bar{y} \pm \lambda_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $\bar{x} - \bar{y} \pm t_{\alpha/2}(n_1 + n_2 - 2) \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	σ_1^2, σ_2^2 are known $\sigma_1^2 = \sigma_2^2$ are unknown
$N(\mu_i,\sigma_1^2)$ $N(\mu_i + \Delta,\sigma_2^2)$	Δ	$\bar{z} \pm t_{\alpha/2}(n-1) \cdot \frac{s_z}{\sqrt{n}}$	

Testing Hypotheses. Introduction

A parameter can be estimated from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval).

Frequently, however, the objective of an investigation is not to estimate a parameter but to decide which of two contradictory claims about the parameter is correct.
Methods for accomplishing this comprise the part of statistical inference called **hypothesis testing.**

Testing Hypotheses. Introduction (2)

- Are the poverty rates of Estonia and Finland different?
- Is there a difference in the mean salaries of men and women?
- Are male and female children equally probable?
- Is the Test 2 easier than Test 1 in this course?
- Is the new treatment more effective than the standard treatment for prolonging the lives of terminal cancer patients?
- ...

Example

Example.
Suppose scientists invent a new drug that supposedly will inhibit a mouse's ability to run through a maze. The scientists design an experiment in which 3 mice are randomly chosen to receive the drug and another 3 mice serve as controls by ingesting a placebo. The time each mouse takes to go through a maze is measured in seconds. Suppose the results of the experiment are as follows

Drug:	**30**	**25**	**20**
Placebo:	18	21	22

The average time for the drug group is $\overline{x}_{drug} = 25$ sec and the average time for the control group is $\overline{x}_{plac} = 20.33$ sec. The mean difference in times is 25 - 20.33 = 4.67 sec.

Is the difference caused by the drug or by the random choice of mice? Maybe accidentally there were lazier mice in the drug group?

Example (2)

Example continued.
If the drug does not really influence times, then split of the six mice into two groups was essentially random. The outcomes could just as easily be distributed

Drug:	**30**	**25**	18
Placebo:	**20**	21	22

There are $C_6^3 = 20$ ways to distribute six mice into two sets of size 3, ignoring any ordering with each set.

Example (3)

Placement	Drug			Placebo			Mean(drug)	Mean(plac)	Difference
1	20	25	30	18	21	22	25,00	20,33	**4,67**
2	20	21	22	18	25	30	21,00	24,33	-3,33
3	20	21	25	18	22	30	22,00	23,33	-1,33
4	20	21	30	18	22	25	23,67	21,67	2,00
5	20	22	25	18	21	30	22,33	23,00	-0,67
6	20	22	30	18	21	25	24,00	21,33	2,67
7	18	20	21	22	25	30	19,67	25,67	-6,00
8	18	20	22	21	25	30	20,00	25,33	-5,33
9	18	20	25	21	22	30	21,00	24,33	-3,33
10	18	20	30	21	22	25	22,67	22,67	0,00
11	18	21	22	20	25	30	20,33	25,00	-4,67
12	18	21	25	20	22	30	21,33	24,00	-2,67
13	18	21	30	20	22	25	23,00	22,33	0,67
14	18	22	25	20	21	30	21,67	23,67	-2,00
15	18	22	30	20	21	25	23,33	22,00	1,33
16	18	25	30	20	21	22	24,33	21,00	3,33
17	21	22	25	18	20	30	22,67	22,67	0,00
18	21	22	30	18	20	25	24,33	21,00	3,33
19	21	25	30	18	20	22	25,33	20,00	**5,33**
20	22	25	30	18	20	21	25,67	19,67	**6,00**

Example (4)

Example continued.

- Of the 20 possible differences in means, 3 are as large or larger than the observed 4.67
- the corresponding probability that pure chance would give a difference this large is $3/20 \approx 0.15$
- 15 % of the cases (of mice replacement/choice) caused the difference to be at least 4.67 seconds. Is this probability high enough to claim that the difference among two considered groups were caused by the random choice of mice? Or maybe the drug still has an influence?

Testing Hypotheses (3)

- The core idea of statistical significance or classical hypothesis testing – to calculate how often pure random chance would give an effect as large as that observed in the data, in the absence of any real effect.
- If that probability is small enough, we conclude that the data provide convincing evidence of a real effect.
- If the probability is not small, we do not make that conclusion.
    - **This is not the same as concluding that there is no effect**; it is only that the data available do not provide convincing evidence that there is an effect.
      In practice, **there may be just too little data to provide convincing evidence.**

Testing Hypotheses (4)

- In order to test hypotheses several statistical criteria are used (z-test, t-test, χ^2-test, sign test, Mann-Whitney test, Wilcoxon test, . . .)
- Choice of test depends on the problem considered, type of variable
- **The idea of test** is to compare the empirical value obtained from observed data (empirical data) with the critical value (from table or from some statistical program)

Roughly speaking, hypothesis testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.

Testing Hypotheses (5)

The goal of hypothesis testing is to determine the likelihood that a population parameter, such as the mean, is likely to be true. There are four steps of hypothesis testing:

Step 1: State the hypotheses.
Step 2: Set the criteria for a decision.
Step 3: Compute the test statistic.
Step 4: Make a decision.

Hypotheses

Definition

A **statistical hypothesis**, or just hypothesis, is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution.

In any hypothesis-testing problem, there are **two contradictory hypotheses** under consideration.

Hypotheses (2)

Definition

The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis**, denoted by H_1, is the assertion that is contradictory to H_0.

H_0 – statement that corresponds to no real effect. This is the status quo, in the absence of the data providing convincing evidence to the contrary.
H_1 – statement that there is a real effect. The data may provide convincing evidence that this hypothesis is true.

Hypotheses (3)

A hypothesis should involve a statement about a population parameter or parameters, commonly referred to as θ,

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

This is called *two-sided hypothesis*. A *one-sided hypothesis* is of the form

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

or also

$$H_0 : \theta \geq \theta_0$$

$$H_1 : \theta < \theta_0$$

Hypotheses (4)

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0 , we will continue to believe in the plausibility of the null hypothesis.

The two possible conclusions from a hypothesis-testing analysis are then **reject** H_0 or **fail to reject** H_0. If H_0 is rejected, then H_1 is said to be proved. You can **never** prove H_0!!!

If H_0 specifies a single value for θ, the hypothesis is called **simple** ($H_0 : \theta = \theta_0$), otherwise it is called **composite** ($H_0 : \theta \leq \theta_0$).

Hypotheses (5)

Example with mice continued

Let μ_{plac} to be the average speed of running through the maze without using drug and μ_{drug} then denotes average speed of running through the maze using drug. Researcher is interested, whether the drug has influence on the speed, e.g.
$H_1 : \mu_{drug} > \mu_{plac}$ or, equivalently, $H_1 : \mu_{drug} - \mu_{plac} > 0$.

How to test hypotheses?

How do we test Hypothesis? Test Procedures

A test procedure is a rule, based on sample data, for deciding whether to reject H_0.

A test procedure is specified by the following:

1. A **test statistic**, a function of the theoretical sample on which the decision (reject H_0 or do not reject H_0) is to be based.
2. A **critical region** (or also, rejection region), the set of all test statistic values for which H_0 will be rejected.

The null hypothesis will then be rejected **if and only if** the observed or computed test statistic value falls in the rejection region.

Test statistic

Definition

A **test statistic** is a function of theoretical sample whose value determines the result of the test. The function itself is generally denoted $T = T(\boldsymbol{X})$, where $\boldsymbol{X}$ is the theoretical sample.

After being evaluated for the sample data $\boldsymbol{x}$, the result is called an observed test statistic and is written in lowercase, $t = T(\boldsymbol{x})$.

Example with mice

Test statistic is here the difference of sample means,
$T(\boldsymbol{X}) = \overline{X}_{drug} - \overline{X}_{plac}$.
Realisation of this is $t = \overline{x}_{drug} - \overline{x}_{plac}$= 25 - 20.33 = 4,67.

Definition

The set of all test statistic values for which H_0 will be rejected, is said to be **critical region**, denoted by C.

Definition

Test of significance is the control rule, which leads to rejection of H_0 if the test statistic value falls in the critical region C:

$$T(\boldsymbol{X}) \in C \xrightarrow{+} \text{reject } H_0$$

Example with mice

We consider the influence of drug to be proved, if the difference of means exceeds 5 sec. In that case the critical region is $C = (5, \infty)$ and test of significance is

$$\overline{X}_{drug} - \overline{X}_{plac} > 5 \xrightarrow{+} \text{reject } H_0$$

When we decide to reject the null hypothesis, we can be correct or incorrect. The incorrect decision is to reject a true null hypothesis. This decision is an example of a **Type I error**.
A **Type II error** is the probability of retaining a null hypothesis that is actually false.

Since we assume the null hypothesis is true, we control for Type I error by stating a *level of significance*.

Definition

The largest probability of committing a Type I error is called the **level of significance** and is denoted by α:

$$\alpha = \begin{cases} P(T(\boldsymbol{X}) \in C | H_0 \text{ correct}), \text{if } H_0 \text{ is simple} \\ sup_{\theta \in H_0} P(T(\boldsymbol{X}) \in C | H_0 \text{ correct}), \text{if } H_0 \text{ is composite} \end{cases}$$

Usual level of significance: $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$.

There are four outcomes for making a decision. The decision can be either correct (correctly reject or retain null) or wrong (incorrectly reject or retain null).

		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT

Why is Type I error so important?

In a jury trial the hypotheses are:

- H_0: defendant is innocent;
- H_1: defendant is guilty.

H_0 (innocent) is rejected if H_1 (guilty) is supported by evidence beyond reasonable doubt. Failure to reject H_0 (prove guilty) does not imply innocence, only that the evidence is insufficient to reject it.

Which error is more serious, convicting an innocent person (Type I) or freeing a guilty person (Type II)?