MS-C1620 Statistical inference

2 Confidence intervals and hypothesis testing

Pekka Pere

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2023–2024
Period III–IV

Contents

Point estimates

With parametric models, we often want to estimate the value of some parameter using the sample $x_1, \ldots, x_n$.

- We estimate the expected value μ of a normal distribution $\mathcal{N}(\mu, \sigma^2)$ by the maximum likelihood estimate $\bar{x}$.
- We estimate the population skewness coefficient γ by the corresponding sample estimate $\hat{\gamma}$.

Such estimates are called point estimates of the parameter.

(A *parameter* is a quantity characterizing the population / generating distribution, similar to *statistic* which is property of a sample.)

A point estimate on its own rarely gives us enough information.
To gain some idea of the precision of a point estimate, they are usually accompanied with some measures of their accuracy.

Confidence interval

A confidence interval gives an estimated range of values which is likely to contain the value of an unknown population parameter.

The confidence level of a confidence interval determines the probability that the confidence interval produced (interpreted as a random interval) will contain the true parameter value.

E.g. if 95% confidence intervals for an unknown parameter are computed from 100 independent samples, approximately 95 of the these will contain the true parameter value — but we do not know which!

Note that any particular realized confidence interval either contains the true value or not; the 95% frequency concens the probability *in the sampling process*

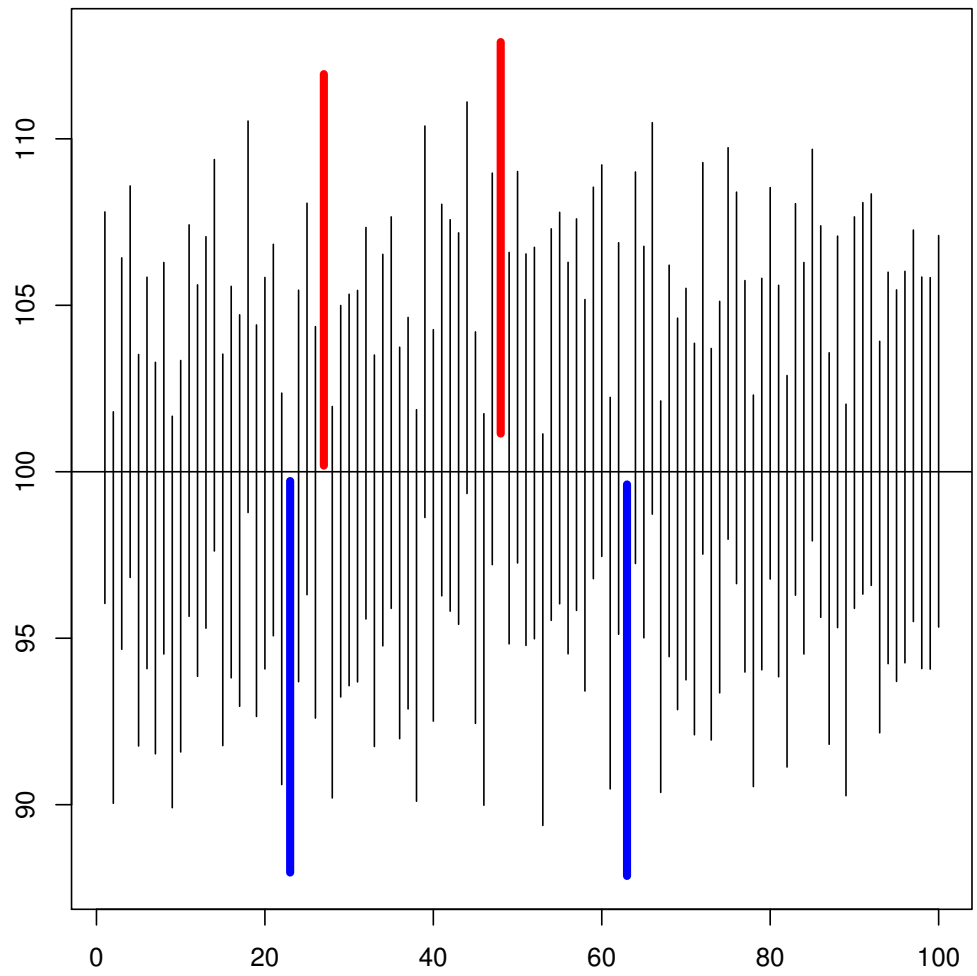


Figure: A statistical Sibelius Monument. One hundred 95 % confidence intervals for mean μ assuming a known σ^2 ($n = 36$). The observations were generated from $\mathrm{N}(100, 18^2)$.

The bootstrap

The standard formulas for confidence intervals either make heavy parametric assumptions or work only for parameters estimable by means (CLT).

The standard non-parametric procedure for estimating confidence intervals is known as the bootstrap.

Bootstrap creates pseudo-samples by drawing n observations *from the data*, *with replacement*, *repeating* the procedure for a large number of times.

If n is large enough, the pseudo-sampling approximates true sampling from the population.

Bootstrap confidence intervals

Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed (i.i.d.) sample from a distribution (parametric model) F_x.

Let θ be a parameter of the distribution F_x and assume that $\hat{\theta}$ is a point estimate of θ.

An approximate confidence interval for θ can now be obtained by bootstrap resampling as follows:

Bootstrap confidence intervals

1. Select n data points randomly with replacement from the original sample $\{x_1, x_2, ..., x_n\}$. Each data point can be selected once, multiple times, or not at all. (Note that the sample size of the new sample is the same as the sample size of the original sample.)
2. Use this new sample to calculate a new estimate for the parameter θ.
3. Repeat the previous steps B times.
4. After the replications, order the B estimates from the smallest to the largest.
5. A $100(1-\alpha)\%$ confidence interval is now obtained by choosing the $\lfloor B \times (\alpha/2) \rfloor$ ordered estimate as the lower endpoint and the $\lfloor B \times (1-\alpha/2) \rfloor$ ordered estimate as the upper endpoint.

Exact confidence intervals

Often, when the type of the distribution is known, also exact confidence intervals can be calculated.

Bootstrap, however, while an approximation, makes no assumptions on the distribution of the data.

Exact confidence intervals, normal distribution

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

A $100(1-\alpha)\%$ confidence interval for μ is obtained as,

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

where $t_{1-\alpha/2}(n-1)$ is the $(1-\alpha/2)$-quantile of Student's t-distribution with $n-1$ degrees of freedom.

For large values of n, the Student's t-distribution with $n-1$ degrees of freedom approaches the standard normal distribution and its corresponding quantile can be substituted in place of $t_{1-\alpha/2}(n-1)$.

Exact confidence intervals, normal distribution

A $100(1-\alpha)\%$ confidence interval for σ^2 is obtained as,

$$\left(\frac{(n-1) \times s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1) \times s^2}{\chi^2_{\alpha/2}(n-1)} \right),$$

where $\chi^2_{\alpha/2}(n-1)$ is the $\alpha/2$-quantile and $\chi^2_{1-\alpha/2}(n-1)$ is the $(1-\alpha/2)$-quantile of the $\chi^2(n-1)$-distribution (χ^2-distribution with $n-1$ degrees of freedom).

The validity of the confidence interval hinges strongly on the Normality assumption. The interval is not robust with respect to the distribution of the observations.

Contents

Hypothesis testing

Statistical tests are applied extensively in various fields of science.

Examples of statistical testing situations

- Testing whether psychic can predict the winner of a sports match.
- Testing whether a treatment works better than the old one.

Hypothesis testing

Statistical hypothesis testing is based on

1. Selecting a statistical model/assumptions and
2. Setting a null hypothesis, and often also an alternative hypothesis,

p-value of the test statistic is a basis for conclusions.

Statistical model

- Statistical model/assumptions casts the problem in a mathematical context and defines the rules of probability governing it.
- Statistical models are usually of the form:

  "Let $x_1, \ldots, x_n$ be an i.i.d. sample from the distribution F with the unknown parameter θ".
- The validity of the model can, and in general should, be tested separately.

Examples of statistical models/assumptions

- A psychic guesses the winner of each of the n sports matches correctly with the probability p, independent of his previous guesses.
- The treatment group responses $x_1, \ldots, x_n$ are an i.i.d. sample from $\mathcal{N}(\mu_1, \sigma^2)$ and the control group responses $y_1, \ldots, y_m$ are an i.i.d. sample from $\mathcal{N}(\mu_2, \sigma^2)$.

Null hypothesis

- The statement of interest about a model parameter is called the null hypothesis, H_0.
- H_0 is assumed (or pretended!) to be true. It is rejected if there is strong evidence that indicates otherwise.
- In simple statistical tests the null hypothesis can often be stated as an *equality*, H_0: $\theta = \theta_0$, where θ is the parameter being tested and θ_0 is a fixed value of the parameter.
- The null hypothesis is often conceptually of the form "*equals*" or "*no difference*".

Examples of null hypotheses

- H_0: $\pi = 0.5$ (π is probability).
- H_0: $\mu_1 - \mu_2 = 0$.

Alternative hypothesis

- Null hypothesis H_0 is usually accompanied by an alternative hypothesis H_1. It is often the logical opposite of H_0.
- If H_0 is rejected then H_1 is accepted.
- The alternative hypothesis is often conceptually of the form "*differs*".

Examples of alternative hypotheses

- H_1: $\pi \neq 0.5$.
- H_1: $\mu_1 - \mu_2 \neq 0$.

Most tests in these lecture slides are for simplicity formulated using *two-sided alternative hypotheses*:

$$H_0\text{: } \theta = \theta_0 \qquad H_1\text{: } \theta \neq \theta_0.$$

(Often a one sided-alternative, such as H_1: $\pi > 0.5$, would be natural.)

Test statistic

- Test statistic measures deviation of the observed sample from the null hypothesis.
- A test statistic is a random variable. Its value depends on the random observations.
- The distribution of the test statistic under the null hypothesis must be known for assessing the compatibility of the observations with the null hypothesis.

Examples of test statistics

- The proportion of correct guesses out of the total n.
- $(\hat{\mu}_1 - \hat{\mu}_2)/\mathrm{SD}(\hat{\mu}_1 - \hat{\mu}_2)$.

p-value

- *p*-value of a test statistic is the probability of observing at least as deviating value towards H_1 as the observed value of the test statistic under the null hypothesis H_0.
- What is considered as "deviating" depends on the form of the hypotheses.
- If the *p*-value is *very small* (the observation is too strange to have happened under H_0) then we reject H_0 in favor of H_1.
- Non-rejection of H_0 does not mean that H_0 is true.

Significance level and critical values

- Significance level α is used to make a cut-off between small and large p-values.
  - If $p < \alpha$ we reject H_0.
  - If $p \geq \alpha$ we do not reject H_0.
- Commonly used significance levels are $\alpha = 0.05, 0.1, 0.01, 0.001$.
- The set of values of the test statistic for which the null hypothesis is rejected (i.e. the values that yield a p-value smaller than α) is called the critical region.
- The threshold values delimiting the regions of non-rejection and rejection for the test statistic are called the critical values.
- Neyman–Pearson theory: Set α beforehand. It may be wiser to apply the p-value more flexibly in combination with other information.

Errors in statistical hypothesis testing

The are two kinds of errors related to the rejection of the null hypothesis H_0.

- Type 1 error: True null hypothesis is rejected.
- Type 2 error: False null hypothesis is not rejected.

The type 1 error rate α is the probability of rejecting a true H_0.

The type 2 error rate is the probability of not rejecting a false H_0. Type 2 error rate is more difficult to control as it is usually a function of the possible distributions of the test statistic under H_1.

Power of a test is equal to $1 -$ "type 2 error rate". The larger the power, the better the test detects false null hypotheses.

Steps of statistical hypothesis testing

1. Select the statistical model and state the hypotheses.
2. Select a test statistic.
3. Pick a sample (for which the model holds).
4. Calculate the value of the test statistic from the data.
5. Calculate the p-value corresponding to the observed value of the test statistic.
6. Draw conclusions and reject/do not reject the null hypothesis.

Contents

One-sample t-test

One-sample t-test compares the expected value of a distribution to a given constant.

One-sample t-test, assumptions

Let $x_1, x_2, \ldots, x_n$ be and i.i.d. sample from $\mathcal{N}(\mu, \sigma^2)$.

One-sample t-test, hypotheses

$$\mathrm{H}_0 : \mu = \mu_0 \qquad \mathrm{H}_1 : \mu \neq \mu_0.$$

One-sample t-test

One-sample t-test, test statistic

- The t-test statistic,
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$
- follows Student's t-distribution with $n - 1$ degrees of freedom under H_0.
- The expected value of t under the null hypothesis H_0 is 0. If the value of t is **large in absolute value**, evidence against the null hypothesis H_0 is found.

If the sample size is large, then the one-sample t-test is not very sensitive to moderate deviations from normality.

Two-sample t-test

Two-sample t-test compares the expected values of two distributions.

Two-sample t-test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from $\mathcal{N}(\mu_x, \sigma_x^2)$ and let $y_1, y_2, \ldots, y_m$ be an i.i.d. sample from $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, let the two samples be independent.

Two-sample t-test, hypotheses

$$\mathrm{H}_0 : \mu_x = \mu_y \qquad \mathrm{H}_1 : \mu_x \neq \mu_y.$$

Two-sample t-test

Two-sample t-test, test statistic

- The t-test statistic,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}},$$

follows approximately the Student's t-distribution with

$$\frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)}.$$

degrees of freedom under H_0.

- The expected value of t under H_0 is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.

If the sample size is large, then the two-sample t-test is not very sensitive to moderate deviations from normality.

Paired t-test

The two-sample t-tests assumes that the samples are independent. What if this is not the case?

- A comparison of two measurement devices where both devices are used to measure the same subject under same circumstances.
- A drug study where the subjects' responses are measured both before and after the treatment.
- A comparison of health-related life style choices of matched pairs, such as spouses.

Paired data can be a great advantage: Effect of confounders tends to be smaller than with the two-sample t-test.

Paired t-test

Paired t-test, assumptions

Observations consist of an i.i.d. sample of pairs $(x_{i1}, x_{i2}), i = 1, 2, ..., n$ (the values **within** a pair need not be independent). The differences $d_i = x_{i1} - x_{i2}$ have the normal distribution $\mathcal{N}(\mu_d, \sigma_d^2)$.

Paired t-test, hypotheses

$$H_0 : \mu_d = 0 \qquad H_1 : \mu_d \neq 0.$$

The recipe is now simple: Apply the *one-sample* t-test to the differences d_i, to test whether their expected value is zero (whether there is no systematic difference between the values in a pair).
(Stop for a moment to think why this works.)

Contents

Variance test

The variance test compares the variance of a distribution to a given constant.

Variance test, assumptions

Let $x_1, x_2, \ldots, x_n$ be and i.i.d. sample from $\mathcal{N}(\mu, \sigma^2)$.

Variance test, hypotheses

The null hypothesis

$$\mathrm{H}_0 : \sigma^2 = \sigma_0^2 \qquad \mathrm{H}_1 : \sigma^2 \neq \sigma_0^2.$$

Variance test

Variance test, test statistic

- The χ^2-test statistic,

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

  follows χ^2-distribution with $n-1$ degrees of freedom under H_0.
- The expected value of the test statistic under H_0 is $n-1$ and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

The variance test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.

Variance comparison test

The variance comparison test compares the variances of two distributions.

Variance comparison test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from $\mathcal{N}(\mu_x, \sigma_x^2)$ and let $y_1, y_2, \ldots, y_m$ be an i.i.d. sample from $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, let the two samples be independent.

Variance comparison test, hypotheses

$$\mathrm{H}_0 : \sigma_x^2 = \sigma_y^2 \qquad \mathrm{H}_1 : \sigma_x^2 \neq \sigma_y^2.$$

Variance comparison test

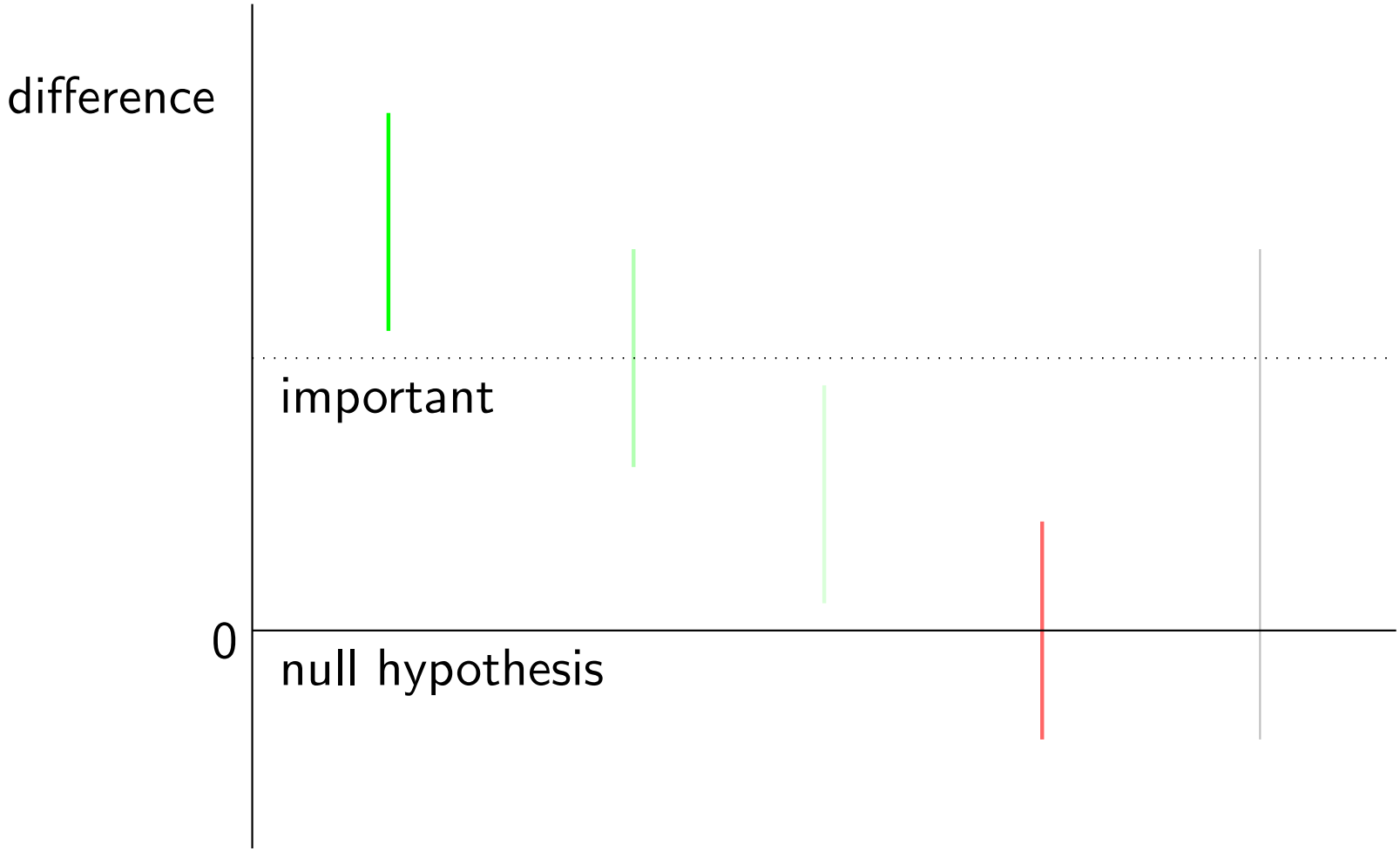
Variance comparison test, test statistic

- The F-test statistic,

$$F = \frac{s_x^2}{s_y^2},$$

  follows the F-distribution with $n-1$ and $m-1$ degrees of freedom under H_0.
- The expected value of the test statistic under H_0 is ≈ 1 and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

Also the variance comparison test is sensitive to deviations from normality and does not work, even for large samples, if the underlying distribution is skewed.



(a)	(b)	(c)	(d)	(e)
significant			*not significant*	
definitely important	possibly import.	not import.	true neg. result	inconclusive result

Figure of the previous page: Confidence intervals, statistical significance, and practical importance.

Source: P. Armitage, G. Berry, and J. N. S. Matthews (2002) *Statistical Methods in Medical Research, 4th edition*. Blackwell Science. (P. 92.)