Fachhochschule Dortmund
University of Applied Sciences and Arts

Data Science

10: Confidence intervals & tests

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Recap

Situation: The form of the density of a distribution is known up to one parameter θ. *theta* can take a value given in the parameter space Θ.

A **point estimator** $\hat{\theta}$ is a function of an independent and equally distributed random sample $X_1, \dots, X_n$ to estimate the value of θ.

- $\hat{\theta}(X_1, \dots, X_n)$ depends on the random variables $X_1, \dots, X_n$ and is also random.
- $\hat{\theta}(x_1, \dots, x_n)$ is computed from an observed sample and is called **estimated value** or **estimation**.

Example

If X is normally distributed, then the arithmetic mean $\hat{\mu} = \bar{x}$ is an estimation for $E(X) = \mu$.

Data Science

Recap

Let $g_l(X_1, \ldots, X_n)$ and $g_u(X_1, \ldots, X_n)$ be two functions of a random sample with $g_l \leq g_u$ such that

$$P(g_l \leq \theta \leq g_u) = 1 - \alpha.$$

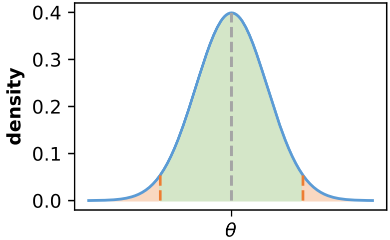
Then we call the interval $[g_l, g_u]$ **confidence interval** for θ with **confidence level** $1 - \alpha$.

- The boundaries g_l and g_u are called **lower** and **upper confidence bound**.
- If $g_l \neq -\infty$ and $g_u \neq \infty$ then we call the confidence interval **two-sided**.

Data Science

Recap

Assuming that θ is a random variable. Then is $\hat{\theta}$ a sampled value from a distribution. Thus we can compute the probability that $\hat{\theta}$ was chosen.



- Green area: Probability that $\hat{\theta}$ lies in this area
- Red area: Probability that $\hat{\theta}$ lies in this area

Green and red area define interval sizes - moving these to $\hat{\theta}$ in center gives interval where θ lies compared to $\hat{\theta}$ with corresponding probabilities.

Data Science

Recap

$(1-\alpha)\%$-confidence interval for expected value $E(X)$

σ	n	distribution	confidence interval
known	arbitrary	$X \sim \mathcal{N}(\mu, \sigma^2)$	$\left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$
known	large	arbitrary	
unknown	arbitrary	$X \sim \mathcal{N}(\mu, \sigma^2)$	$\left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$
unknown	large	arbitrary	$\left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$

$X_1, \ldots, X_n$ random sample, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2, \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \hat{p} = \overline{X}.$

Data Science

Today

Statistical tests

Statistical tests

Statistical tests

Statistical tests (z- and t-test)

Confidence intervals give a range, in which the desired value is probably located - but often one has an assumption which value the estimator estimates. Probability to verify that this assumption is correct - or maybe probability to reject this assumption?

Example

- The average grade in math tests is 3
- The average height of males in Germany is $1.8m$
- The average speed on the highway is $120km/h$

Example

A machine packs chocolates in boxes with a target weight of 15g

Question: Does the machine need an adjustment?

- Weight of the boxes is a random variable X with parameter $E(X)$
- $\overline{X}$: arithmetic mean of $n = 10$ randomly chosen boxes
  - If $\overline{x}$ around 10 or 20 –› Probably $E(X)$ is more than or less than 15!
  - If $\overline{x}$ around 15 –› Probably $E(X)$ is also close to 15!
- **Attention:**
  - $\overline{x}$ around 10 or 20 is also possible even if $E(X)$ is close to 15 (first degree error)
  - $\overline{x}$ around 15 is also possible even in $E(X)$ more than or less than 15 (second degree error)

Ideal: Formal rule and statement on error probability.

Testproblem

Formulation of a **null hypothesis** H_0 and formulation of an **alternative hypothesis** H_1 which are mutually exclusive.

Test-statistic

Function of a random sample $X_1, \ldots, X_n$, which allows assessing if H_0 or H_1 is more like to be valid.

Rejection area

Values of the test-statistic, for which H_0 is rejected - also named critical area.

Data Science

Statistical tests (z- and t-test)

- **First degree error**: Reject H_0, in the case that H_0 is true
- **Second degree error**: Keep H_0, in the case that H_1 is true

	H_0 not rejected	reject H_0
H_0 correct	right decision	first degree error
H_0 false	second degree error	right decision

- **Significance level:** $P(\text{reject}H_0 | H_0 \text{ correct}) \leq \alpha$
- **Test quality:** $1 - \beta = P(\text{not reject}H_0 | H_0 \text{ false}) \leq \alpha$

The value α is set before performing the test. Usually: $\alpha = 0.01, 0.05, 0.1$

Data Science

Statistical tests (z- and t-test)

Example

- Assumption: X is normally distributed, $X \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma = 1.7$ known
- Rejection area: $\{\overline{x} | \overline{x} \leq 14 \text{ or } \overline{x} \geq 16\}$
- Then, the first degree error

$$P(\overline{X} \leq 14 | \mu = 15) + P(\overline{X} \geq 16 | \mu = 15)$$
$$= P\left(Z \leq \frac{14-15}{1.7/\sqrt{10}}\right) + P\left(Z \geq \frac{16-15}{1.7/\sqrt{10}}\right) \quad = P\left(Z \leq -1.86\right) + P(Z \geq 1.86) = 0.062$$

- A small area of rejection leads to a small first degree error, because the probability of rejections becomes less, i.e. $\{\overline{x} | \overline{x} \leq 13.5 \text{ or } \overline{x} \geq 16.5\}$ gives 0.005

Example

Area of rejection: $\{\bar{x} | \bar{x} \leq 13.5 \text{ or } \bar{x} \geq 16.5\}$

- Probability β for a second degree error for $\mu = 13$ is

$$\begin{aligned}\beta &= P(13.5 < \bar{X} < 16.5 | \mu = 13) \\ &= P\left(\frac{13.5 - 13}{1.7/\sqrt{10}} < Z < \frac{16.5 - 13}{1.7/\sqrt{10}}\right) \\ &= P(0.930 < Z < 6.510) = 0.176\end{aligned}$$

The test quality for $\mu = 13$ is given by $1 - \beta = 0.824$.

A larger sample size n for a fixed α reduces β.

Data Science

Statistical tests (z- and t-test)

For confidence intervals we differ between three settings: The random sample consists on independent and indentically distributed random variables, where the distribution is ...

- a normal distribution with known variance
- a normal distribution with unknown variance
- a arbitrary distribution

Depending on the situation we chose a different distribution to work with.

In the following we do the same for the tests. In detail, the chosen test-statistic depend on the distribution of the random sample.

In the following we consider the general test procedure.

$X_1, \ldots, X_n$ independent and identically distributed random variables.

1. Formulation of the **test-problem**:
   - $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu \neq \mu_0$ (two-sided)
   - $H_0\colon \mu \leq \mu_0$ vs. $H_1\colon \mu > \mu_0$ (right-sided)
   - $H_0\colon \mu \geq \mu_0$ vs. $H_1\colon \mu < \mu_0$ (left-sided)

   The rejection of H_0 is a hard conclusion for which the probability of a wrong decision is limited by α. Therefore, the **important statement to be verified is placed in the alternative**.

2. Chose a proper **significance level** α
3. **Test-statistic**: TS: Choice depends on the distribution of the random variable.

4. Determination of the **area of rejection** for selected α: Reject H_0, if
   - $|ts| > ts_{1-\alpha/2}$ for a two-sided test
   - $ts > ts_{1-\alpha}$ for a right-sided test
   - $ts < -ts_{1-\alpha}$ for a left-sided test
5. Compute the **value of the test statistic** for an observed sample: ts
6. **Decision:**
   - **Reject** H_0 if the value of z is in the rejection area
     Do **not reject** H_0 if the value of z is **not** in the rejection area
   - Name the used significance level
   - Formulate the significance of the test decision for the original question

Components of a statistical hypothesis test

1. Test-problem
2. Choice of significance level
3. test-statistic
4. Area of rejection
5. Value of test-statistic
6. Decision

It remains to choose the proper test-statistic!

- Example showed normal distribution with known variance.
- For a large sample size n (often $n \geq 30$) and known variance, the central limit theorem gives that $\overline{X}$ is approximately normal distributed and the Gaussian test can be used approximately.
- Is the variance unknown, then the t-distribution is for large n close the standard normal distribution and the Standard deviation can be replaced by S in the Gaussian test.

Data Science

Statistical tests (z- and t-test)

$X \sim \mathcal{N}(\mu, \sigma^2)$ or $n \geq 30$, σ known: (approximate) Gaussian test

Use standard normal distribution as test-statistic: $Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$

$X \sim \mathcal{N}(\mu, \sigma^2)$, σ unknown: t-test

Use t-distribution as test-statistic: $T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$

$n \geq 30$, σ unknown: approximate Gaussian test

Use standard normal distribution as test-statistic: $Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$

Reminder t-distribution

- For a random sample $X_1, \ldots, X_n$ of normally distributed random variables with expected value μ there holds

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ with } S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$$

  is t-distributed with parameter $n-1$ (degrees of freedom)
- Quantiles $t_{n-1,\alpha}$ can be read from a table

Fachhochschule Dortmund
University of Applied Sciences and Arts

Data Science

Statistical tests (z- and t-test)

we focus on students

(approximate) Gaussian test ($X \sim \mathcal{N}(\mu, \sigma^2)$ or $n \geq 30$, σ known)

Null hypothesis	Alternative hypothesis	Test-statistics	Rejection area
$\mu = \mu_0$	$\mu \neq \mu_0$	$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\lvert z \rvert > z_{1-\frac{\alpha}{2}}$
$\mu \geq \mu_0$	$\mu < \mu_0$		$z < -z_{1-\alpha}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$z > z_{1-\alpha}$

t-test on location ($X \sim \mathcal{N}(\mu, \sigma^2)$, σ unknown)

Null hypothesis	Alternative hypothesis	Test-statistics	Rejection area
$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$\lvert t \rvert > t_{n-1,1-\frac{\alpha}{2}}$
$\mu \geq \mu_0$	$\mu < \mu_0$		$t < -t_{n-1,1-\alpha}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$t > t_{n-1,1-\alpha}$

approximate Gaussian test ($n \geq 30$, σ unknown)

Null hypothesis	Alternative hypothesis	Test-statistics	Rejection area
$\mu = \mu_0$	$\mu \neq \mu_0$	$Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$	$\lvert z \rvert > z_{1-\frac{\alpha}{2}}$
$\mu \geq \mu_0$	$\mu < \mu_0$		$z < -z_{1-\alpha}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$z > z_{1-\alpha}$

Statistical programs often give *p*-value

- It defines the probability to observe an extreme value of the statistic in direction of the alternative, in the case that H_0 is correct.
- Is the p-value small or equal to α, H_0 is rejected

Attention: Risk of misuse due to subsequent adjustment of the significance level to the p-value. Therefore: First determine the significance level, then calculate the p-value.

The hypothesis H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$ is rejected to significance level α if

- $\bar{x}$ is in the rejection area of the test
- p-value is smaller than α
- μ_0 not in the $100(1-\alpha)\%$ confidence interval of μ.

Example: Box weights

$$14.6, 15.7, 16, 13.5, 16, 16.5, 17, 15.4, 15.3, 15$$

Assumption: X is normally distributed: $X \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma = 1.7$

1. **Test-problem** $H_0 = \mu$ vs. $H_1 \neq 15$
2. **Choice of significance level** $\alpha = 0.05$
3. **test-statistic** $Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{15}}$
4. **Area of rejection** Reject H_0 if $|z| > z_{1-\alpha/2} = z_{0.975} = 1.96$
5. **Value of test-statistic** $\overline{x} = 15.5$, $\sigma = 1.7$, $n = 10$, $z = \frac{15.5 - 15}{1.7/\sqrt{10}} = 0.93$
6. **Decision** The null hypothesis is not rejected. The sample gives for the confidence level 0.05 no clue that the machine needs to be adjusted.

Example: Temperature in January

$$2.3, 4, 4.5, 1.5, 2.2, 1.7, 3.6, 6.1, 1.2, 5.3, 3.3, -0.6, 5.2, 0.2, 0.9, 2.6, 2.2, 3.4, 2.8, 2.6$$

Assumption: X is normally distributed, σ^2 unknown

1. **Test-problem** $H_0\colon \mu \leq 2$ vs. $H_1\colon \mu > 2$
2. **Choice of significance level:** $\alpha = 0.05$
3. **test-statistic:** $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$
4. **Area of rejection:** ($n = 20$) Reject H_0 if $t > t_{19,0.95} = 1.729$
5. **Value of test-statistic** $\overline{x} = 2.75$, $S^2 = 2.99$, $n = 20$, $t = 1.94$
6. **Decision:** Null hypothesis ($\mu \leq 2$) is rejected, since the value of t is in the rejection area for the given significance level.

Statistical tests

Two sample t-test for location difference

Data Science

Two sample t-test for location difference

Of interest: Test on the difference in the expected value of two distributions

Examples

- Runtime of two different algorithms
- Test-results of patients with and without therapy
- PISA-points of students different classes

Data Science

Two sample t-test for location difference

- **Question:** Measurements X and Y of the same characteristic in different situations or populations. Here, μ_X, σ_X^2 and μ_Y, σ_Y^2 are the corresponding expected value and variance. Of interest is a possible difference in the situation, i.e. between μ_X and μ_Y.
- **Assumption:**
  - $X_1, X_2, \ldots, X_n$ random sample in Situation 1 with size n
  - $Y_1, Y_2, \ldots, Y_m$ random sample in Situation 2 with size m
  - Both random samples are stochastically independent
  - For both Situations we assume a normal distribution, or we use the central limit theorem, thus

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2) \text{ and } Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Example

- Two different companies deliver chocolate bonbons in two boxes of same size
- The assumption, which should be proven, is that the weight Y of the boxes of the second company are in the mean heavier than the companies boxes of the first company.
- It is assumed, that post companies produces boxes with normally distributed weights.

Task: Perform a statistical test with $\alpha = 0.05$

Data Science

Two sample t-test for location difference

One- and twosided test problems

Null-hypothesis	Alternative hypothesis
$H_0 : \mu_X - \mu_Y = \delta_0$	$H_1 : \mu_X - \mu_Y \neq \delta_0$
$H_0 : \mu_X - \mu_Y \geq \delta_0$	$H_1 : \mu_X - \mu_Y < \delta_0$
$H_0 : \mu_X - \mu_Y \leq \delta_0$	$H_1 : \mu_X - \mu_Y > \delta_0$

Different assumptions on the variance

- σ_X^2 and σ_Y^2 are known
- σ_X^2 and σ_Y^2 are unknown but equal
- σ_X^2 and σ_Y^2 are unknown and possible unequal

These assumptions lead to different procedures - the last case is the more general, therefore this case is considered.

Data Science

Two sample t-test for location difference

The test statistic with sample variance S_X^2 and S_Y^2

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{S_X^2/n + S_Y^2/m}}$$

is t-distributed with degrees of freedom

$$k = \lfloor (S_X^2/n + S_Y^2/m)^2 / (\frac{1}{n-1}(S_X^2/n)^2 + \frac{1}{m-1}(S_Y^2/m)^2) \rfloor$$

Null-hypothesis	**Alternative hypothesis**	**Rejection area**
$H_0 : \mu_X - \mu_Y = \delta_0$	$H_1 : \mu_X - \mu_Y \neq \delta_0$	$\lvert t \rvert > t_{k,1-\alpha/2}$
$H_0 : \mu_X - \mu_Y \geq \delta_0$	$H_1 : \mu_X - \mu_Y < \delta_0$	$t < -t_{k,1-\alpha}$
$H_0 : \mu_X - \mu_Y \leq \delta_0$	$H_1 : \mu_X - \mu_Y > \delta_0$	$t > t_{k,1-\alpha}$

Example

There was an investigation of 20 boxes of the first and 22 boxes of the second company.
$X_1, \ldots, X_{20} \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_{22} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

1. **Test-problem:** $H_0 : \mu_X - \mu_Y \geq 0$ vs. $H_1 : \mu_X - \mu_Y < 0$
2. **Significance level:** $\alpha = 0.05$
3. **Test-statistic:** $T = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{S_X^2/n + S_Y^2/m}}$ with $\delta = 0$.

1. **Area of rejection:** Reject H_0 if $t < -1.685$ since $-t_{k,1-0.05} = t_{39,0.095} = -1.685$ with

$$k = \left\lfloor (S_X^2/n + S_Y^2/m)^2 / (\frac{1}{n-1}(S_X^2/n)^2 + \frac{1}{m-1}(S_Y^2/m)^2) \right\rfloor$$

$$= \left\lfloor (0.8/20 + 0.9/22)^2 / (\frac{1}{19}(0.8/20)^2 + \frac{1}{21}(0.9/22)^2) \right\rfloor$$

$$= \lfloor 39.940 \rfloor = 39$$

Data Science

Two sample t-test for location difference

1. **Value of the test statistic:** Results of the measure: $\overline{x} = 14.5$, $\overline{y} = 16.3$, $s_Y^2 = 0.9$

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_X^2/n + s_Y^2/m}} = \frac{14.5 - 16.3}{\sqrt{0.8/20 + 0.9/22}} = -6.328$$

2. **Decision:** The null hypothesis should be rejected, for a significance level of 5% the bonbons of the second producer are heavier than the bonbons of the first producer.

Data Science

Two sample t-test for location difference

For this test, it is important that the samples are independent - but there might be a dependent random sample, both samples are measured at the same statistical unit - this must be taken into account for the test procedure.

- **Example:**
  - Comparison of blood pressure of a group of patients before and after a treatment
  - Comparison of the sales of specific companies in two different years.
- **Possible solution:** Take the difference $D_i = X_i - Y_i$ as random sample, formulate the test problem for $E(D)(= E(X) - E(Y))$ and use the one sample test.

Data Science

Two sample t-test for location difference

t-Test for location difference

Assumption: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, σ_X, σ_Y unknown

Null-hypothesis	Alternative hypothesis	Test-statistic	Rejection area
$H_0 : \mu_X - \mu_Y = \delta_0$	$H_1 : \mu_X - \mu_Y \neq \delta_0$		$\|t\| > t_{k,1-\alpha/2}$
$H_0 : \mu_X - \mu_Y \geq \delta_0$	$H_1 : \mu_X - \mu_Y < \delta_0$	$T = \frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{S_X^2/n+S_Y^2/m}}$	$t < -t_{k,1-\alpha}$
$H_0 : \mu_X - \mu_Y \leq \delta_0$	$H_1 : \mu_X - \mu_Y > \delta_0$		$t > t_{k,1-\alpha}$

with $k = \lfloor (S_X^2/n + S_Y^2/m)^2 / (\frac{1}{n-1}(S_X^2/n)^2 + \frac{1}{m-1}(S_Y^2/m)^2) \rfloor$

Summary & Outlook

- You are able to perform statistical tests and interpret the results

Data preparation and decision trees

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)

by

Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.