Fachhochschule
Dortmund
University of Applied Sciences and Arts

# Data Science

## 10: Confidence intervals & tests

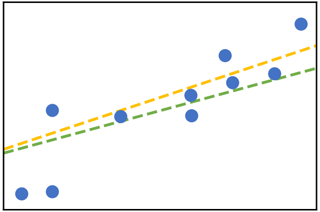
Klaus Kaiser
WiSe 2024 / 2035

# Data Science

## Recap



**Other way around:** Only a sample is given - what are $\beta_0$ and $\beta_1$?



Which line is the "best"? We need a predictor for the values $\beta_0$ and $\beta_1$!

**Simple linear regression** is a model that estimates the linear relationship between one independent and one dependent variable.

# Data Science

## Recap



The linear regression line for $Y$ given $X$ with observations $(x_i, y_i) \in \mathbb{R}^2$ for $i = 1, \ldots, n$ is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

with

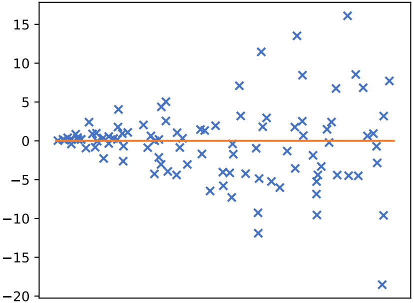
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (\overline{x} - x_i)^2} \text{ and } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

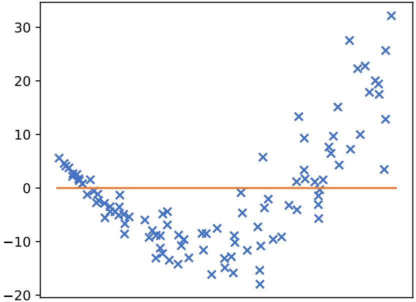
- If $\epsilon$ is normally distributed, then are also $\beta_0$ and $\beta_1$ normally distributed.

# Data Science

## Recap






Residual plot **not** fulfilling all conditions formulated before.

# Data Science

## Recap



The value

$$R^2 = \frac{RSS^*}{TSS} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

is called **coefficient of determination.**

There holds

- $0 \leq R^2 \leq 1$
- $R^2 = 1 - \frac{ESS}{TSS}$
- $R^2 = 1$ if and only if $e_i = 0$ for all $i = 1, \dots, n$: optimal fit, i.e. all observations lie on the regression line.
- $R^2 = 0$ if and only if $\hat{y}_i = \overline{y}$ for all $i = 1, \dots, n$

# Data Science

## Today



**confidence intervals and statistical tests**

# Confidence intervals

**So far:** Point estimator for linear regression coefficients!

- Can we also estimate different values for observations (e.g. expected value)
- How accurate is the estimated value?

Since the estimated values are computed due to observations, we can not expect that these values are accurate. Especially, due to statistics, the values could be far away from the exact ones. Can we measure the uncertainty?

# Confidence intervals

Point and interval estimators

Estimator: Use a sample to gain information about unknowns aspects of the distribution.

### Some properties of an distribution are similar to observed values

| **Distribution of X** | **Sample** $(x_1, \ldots, x_n)$ |
|---|---|
| Distribution function $F(x)$ | empirical distribution function $F_n(x)$ |
| density $f(x)$ | Histogram |
| expected value $\mu$ | arithmetic mean $\bar{x}$ |
| Variance $\sigma^2$ | empirical variance $\tilde{s}^2$ |
| theoretical quantile $x_p$ | empirical quantile $x_p$ |

### What is a random sample?

- Repeat identical random process $n$ times independently of each other $(X_1, \ldots, X_n)$
- Consider sample $x_1, \ldots, x_n$ and compute estimation

Two random variables $X$ and $Y$ are **independent** if for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$ there holds:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

### Random sample of independent identically distributed random variables

$X_1, \ldots, X_n$ are independent and follow the same distribution.

# Data Science

## Point and interval estimators



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

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

- $\hat{\theta}(X_1, \ldots, X_n)$ depends on the random variables $X_1, \ldots, X_n$ and is also random.
- $\hat{\theta}(x_1, \ldots, x_n)$ 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

## Point and interval estimators



### Sum of random variables

For random variables $X_1, \ldots, X_n$ with expected values $E(X_1), \ldots, E_(X_n)$ there holds:

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$$

### Sum and product of *independent* random variables

For independent random variables $X_1, \ldots, X_n$ there holds:

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i) \text{ and } Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

# Data Science

## Point and interval estimators



### Sum of independet normal distributed random variables

For independent $\mathcal{N}(\mu, \sigma^2)$ distributed random variables $X_1, \ldots, X_n$ there holds:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n}$$

is a random variable with $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ distribution.

An estimator $\hat{\theta}(X_1, \ldots, X_n)$ is called **unbiased** for $\theta$ if there holds

$$E(\hat{\theta}) = \theta$$

Otherwise, it is called **biased** and the value

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

is called bias.

# Data Science

## Point and interval estimators



- The **empirical Variance** is biased as an estimator for $\sigma^2$:

$$E(\tilde{S}^2) = E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_i)^2) = \frac{n-1}{n}\sigma^2$$

Therefore, one often uses the sample variance

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

for which there holds $E(S^2) = \sigma^2$.

- No matter which distribution $X$ follows, $\overline{X}$ and $S^2$ are unbiased estimators for $E(X)$ and $Var(X)$

Estimator $\hat{\theta}$ can compute an approximate value for $\theta$, but how accurate is this value?

**Idea: Interval estimator**

Construction of an interval around $\hat{\theta}$, which contains the real value $\theta$ with a given probability.

# Data Science

## Point and interval estimators



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 $\theta$ with **confidence level** $1 - \alpha$.

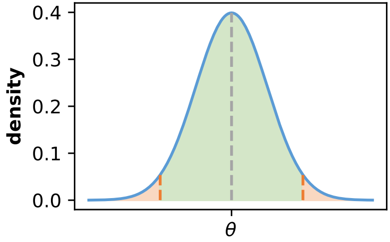
- The boundaries $g_l$ and $g_u$ are called **lower** and **upper confidence bound**.
- If $g_l \neq -\infty$ and $g_u \neq \infty$ then we call the confidence interval **two-sided**.

# Data Science

## Point and interval estimators



Assuming that $\theta$ 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 $\theta$ lies compared to $\hat{\theta}$ with corresponding probabilities.

# Data Science

## Point and interval estimators



### Confidence interval for $E(X) : X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma$ known

$X_1, \ldots, X_n$ sample of $\mathcal{N}(\mu, \sigma^2)$ distribution

- **Of interest:** How close is $\overline{X}$, unbiased estimator of the expected value, to the unknown exact mean value $\mu$.
- **Use:** Random distribution of $\overline{X}$, i.e. $\overline{X}$ is $\mathcal{N}(\mu, \frac{\sigma}{n})$ distributed
- Further assumption: $\sigma^2$ is known

# Data Science

## Point and interval estimators



- For every probability $1 - \alpha$, there are quantile $z_{1-\frac{\alpha}{2}}$ such that

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

- Since $\sigma > 0$ there holds

# Data Science

## Point and interval estimators



- For every probability $1-\alpha$, there are quantile $z_{1-\frac{\alpha}{2}}$ such that

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

- Since $\sigma > 0$ there holds

$$\begin{aligned}
& -z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}} \\
\Leftrightarrow & -z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{X}-\mu \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\
\Leftrightarrow & -\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\
\Leftrightarrow & \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

# Data Science

## Point and interval estimators



The probability that the random interval

$$\left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

contains the real value $\mu$ equals to $1 - \alpha$.

- For a concrete sample (i.e. a sample is observed), the interval boundaries can be computed. Resulting interval is called $100(1-\alpha)$**% confidence interval** for the expected value.
- $1 - \alpha$ is also called **confidence probability** or **safety probability**
- With larger sample size $n$, the interval becomes smaller

# Data Science

## Point and interval estimators



### Example

Sample for weight of bonbon packages [g]:

$$64.1, 64.7, 64.5, 64.6, 64.5, 64.6, 64.8, 64.2, 64.3$$

**Known:** Weight of packages is normally distributed with $\sigma = 1$

**Task:** $95\%$-confidence interval of the package weight. $n = 10$, $\bar{x} = 64.46$ (parameter estimator for $\mu = E(X)$). $\alpha = 0.05$, $z_{1-\alpha/2} = z_{0.975} = 1.96$

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

$$\Rightarrow \left[64.46 - 1.96 \frac{1}{\sqrt{10}}, 64.46 + 1.96 \frac{1}{\sqrt{10}}\right] = [63.84, 65.08]$$

# Confidence intervals

## Confidence intervals for expected values

# Data Science

## Confidence intervals for expected values



- Point estimate gives estimation for $\hat{\theta}$, which is normally not identically with the real value $\theta$
- interval estimation gives information on the accuracy of the estimated value
- An interval is estimated in such a way, that the probability that the real value $\theta$ is not contained in this interval equals to $\alpha$ (e.g. $\alpha = 0.1, 0.05$)
- Real values $\theta$ is contained in this interval with confidence-probability (confidence level) $1 - \alpha$

Next?

**Confidence interval for $E(X) : X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma$ unknown**

### Example

Sample for weight of bonbon packages [g]:

$$64.1, 64.7, 64.5, 64.6, 64.5, 64.6, 64.8, 64.2, 64.3$$

Assumption: Weight of the packages is normally distributed, $\mu$ and $\sigma^2$ unknown.

- Point estimator for unknown expected value $\mu$ and unknown variance $\sigma^2$ of the package weight:

$$?$$

- $95\%$ confidence interval of the package weight

$$?$$

# Data Science

## Confidence intervals for expected values

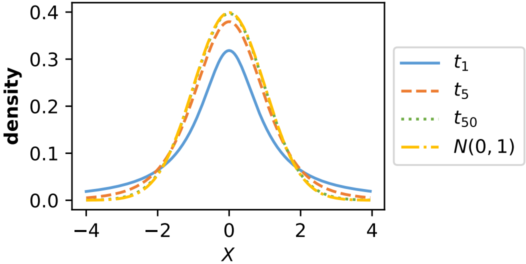


- $X_1, \ldots, X_n$ sample of a $\mathcal{N}(\mu, \sigma^2)$ distribution
- Then $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$ are also random variables
- The distribution of the random variable

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

is given by a **t-distribution** with $r = n - 1$ so-called degrees of freedom - $T \sim t_r$

- Notation: $P(T \leq t_{r,\alpha} = \alpha)$




- t-distribution is nearly $\mathcal{N}(0,1)$ distributed for $r$ large

# Data Science

## Confidence intervals for expected values



Assumption: $\mu$ and $\sigma^2$ are unknown

- $T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ is $t_{n-1}$ distributed, with $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$
- $100(1-\alpha)\%$ confidence interval for $\mu$:

$$[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}]$$

### Example

Sample for weight of bonbon packages [g]:

$$64.1, 64.7, 64.5, 64.6, 64.5, 64.6, 64.8, 64.2, 64.3$$

**Assumption:** Weight of the package is normally distributed

**Task:** $95\%$ confidence interval of the package weight?

$\bar{x} = 64.46$, $n = 10$, $t_{9,0.975} = 2.2662$ $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = 0.0515$, $s = 0.227$

$$\begin{aligned}
&[\bar{x} - t_{9,0.975}\frac{s}{\sqrt{n}}, \bar{x} + t_{9,0.975}\frac{s}{\sqrt{n}}] \\
=&[64.46 - 2.2662\frac{0.227}{\sqrt{10}}, 64.46 + 2.2662\frac{0.227}{\sqrt{10}}] \\
=&[64.297, 64.623]
\end{aligned}$$

### Central limit theorem

For $X_1, \ldots, X_n$ independent identical distributed random variables with expected value $\mu$ and variance $\sigma > 0$, then there holds for every $x \in \mathbb{R}$:

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x)$$

The central limit theorem gives a justification for the popularity of the normal distribution: The sum of many independent random variables is nearly normally distributed:

- measurement errors
- water / energy consumption in a city
- body-weight
- ...

# Data Science

## Confidence intervals for expected values



- The distribution of a sum of sufficiently many ($n$ large) independent, identically distributed random variables can be approximated, due to the central limit theorem, by a normal distribution with $E(\sum X_i) = n\mu$ and $Var(\sum X_i) = n\sigma^2$, i.e.

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

- Rule of thumbs: $n \geq 30$ is sufficiently large - but in some cases also smaller values of $n$ are sufficient, especially if the random variables are nearly symmetrically distributed.

# Data Science

## Confidence intervals for expected values



- Arbitrary distribution, $\sigma^2$ unknown, then approximately

  $$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}) \approx 1-\alpha \text{ for large } n$$

  and

  $$\left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

  approximate $100(1-\alpha)\%$ confidence interval for $\mu$.

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

| $\sigma$ | $n$ | **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, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2, \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \hat{p} = \overline{X}.$

# Statistical tests

# Statistical tests

Statistical tests and z-test

# Data Science

## Statistical tests and z-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 and z-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 $\alpha$ is set before performing the test. Usually: $\alpha = 0.01, 0.05, 0.1$

## 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

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

- 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 $\beta$ 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 $\alpha$ reduces $\beta$.

Test procedure for $\mu$ in the case of a normal distribution with known variance

$X_1, \ldots, X_n$ random sample of $X \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma^2$ known

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

   The rejection of $H_0$ is a hard conclusion for which the probability of a wrong decision is limited by $\alpha$. Therefore, the **important statement to be verified is placed in the alternative.**
2. Chose a proper **significance level** $\alpha$
3. **Test-statistic**:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

4. Determination of the rejection range for selected $\alpha$: Reject $H_0$, if
   - $|z| > z_{1-\alpha/2}$ for a two-sided test
   - $z > z_{1-\alpha}$ for a right-sided test
   - $z < -z_{1-\alpha}$ for a left-sided test
5. Compute the value of the test statistic for an observed sample: $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
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

### 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.

# Data Science

## Statistical tests and z-test



### (approximative) z-test

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

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

# Statistical tests

One sample t-test for location

# Data Science

## One sample t-test for location



- **t-Test:** Same construction idea as Gaussian test - but assuming normal distribution with unknown variance.
- Reminder t-distribution:
  - For a random sample $X_1, \ldots, X_n$ of normally distributed random variables with expected value $\mu$ 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

# Data Science

## One sample t-test for location



Test procedure for $\mu$ in the case of a normal distribution with known variance

$X_1, \ldots, X_n$ random sample of $X \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma^2$ unknown

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)
2. Chose a proper **significance level** $\alpha$
3. **Test-statistic**:

$$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

4. gives a statement on $H_0$ and distribution is known. For $\mu = \mu_0$ $T$ is t-distributed with $n - 1$ degrees of freedom, such that

$$P(T \leq t_{n-1,1-\alpha}) = P(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq t_{n-1,1-\alpha}) = 1 - \alpha$$

# Data Science

## One sample t-test for location



4. Determination of the rejection range for selected $\alpha$: Reject $H_0$, if
   - $|t| > t_{n-1,1-\alpha/2}$ for a two-sided test
   - $t > t_{n-1,1-\alpha}$ for a right-sided test
   - $t < -t_{n-1,1-\alpha}$ for a left-sided test
5. Compute the value of the test statistic for an observed sample: $t = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
6. **Decision:**
   - **Reject** $H_0$ if the value of $t$ is in the rejection area
     Do **not reject** $H_0$ if the value of $t$ is **not** in the rejection area
   - Name the used significance level
   - Formulate the significance of the test decision for the original question

### 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, $\sigma^2$ unknown

1. Test-problem $H_0$: $\mu \leq 2$ vs. $H_1$: $\mu > 2$
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 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 $\alpha$, $H_0$ 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.

# Data Science

## One sample t-test for location



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

- $\bar{x}$ is in the rejection area of the test
- $p$-value is smaller than $\alpha$
- $\mu_0$ not in the $100(1-\alpha)\%$ confidence interval of $\mu$.

### Normal distribution apprixomation

- **up to now:** Assumption that a normal distribution is given
- 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

## One sample t-test for location



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

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, $\mu_X, \sigma_X^2$ and $\mu_Y, \sigma_Y^2$ are the corresponding expected value and variance. Of interest is a possible difference in the situation, i.e. between $\mu_X$ and $\mu_Y$.
- **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

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

# Data Science

## Two sample t-test for location difference



### 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 |
|---|---|
| $H_0 : \mu_X - \mu_Y = \delta_0$ | $H_1 : \mu_X - \mu_Y \neq \delta_0$ |
| $H_0 : \mu_X - \mu_Y \geq \delta_0$ | $H_1 : \mu_X - \mu_Y \ngeq \delta_0$ |
| $H_0 : \mu_X - \mu_Y \leq \delta_0$ | $H_1 : \mu_X - \mu_Y \nleq \delta_0$ |

### Different assumptions on the variance

- $\sigma_X^2$ and $\sigma_Y^2$ are known
- $\sigma_X^2$ and $\sigma_Y^2$ are unknown but equal
- $\sigma_X^2$ and $\sigma_Y^2$ are unknown and possible unequal

These assumptions lead to different procedures - the last case is the most 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 |
|---|---|---|
| $H_0 : \mu_X - \mu_Y = \delta_0$ | $H_1 : \mu_X - \mu_Y \neq \delta_0$ | $\lvert t \rvert > t_{k,1-\alpha/2}$ |
| $H_0 : \mu_X - \mu_Y \geq \delta_0$ | $H_1 : \mu_X - \mu_Y \ngeq \delta_0$ | $t < -t_{k,1-\alpha}$ |
| $H_0 : \mu_X - \mu_Y \leq \delta_0$ | $H_1 : \mu_X - \mu_Y \nleq \delta_0$ | $t > t_{k,1-\alpha}$ |

### 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 $\delta = 0$.

# Data Science

## Two sample t-test for location difference



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

$$
\begin{aligned}
k &= \left\lfloor (S_X^2/n + S_Y^2/m)^2 / (\frac{1}{n-1}(S_X^2/n)^2 + \frac{1}{m-1}(S_Y^2/m)^2) \right\rfloor \\
&= \left\lfloor (0.8/20 + 0.9/22)^2 / (\frac{1}{19}(0.8/20)^2 + \frac{1}{21}(0.9/22)^2) \right\rfloor \\
&= \lfloor 39.940 \rfloor = 39
\end{aligned}
$$

# Data Science

## Two sample t-test for location difference



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

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

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



- **up to now:** Two sample t-test for location difference with two independent random samples
- **Problem:** The might be **dependent random samples**, i.e. 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 random sample test.

### t-Test for location difference

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

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

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

# Data Science

## **Summary & Outlook:** Summary



- You understand what confidence intervals are and how they are computed
- You are able to communicate uncertenty concerning estimators
- You are able to perform statistical tests and interpret the results

# Data Science

## **Summary & Outlook:** Outlook



**Continue statistical tests**

# Data Science

## **Summary & Outlook:** Acknowledgement



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

by

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