Course 02403

Lecture 7: Two sample t-test

DTU Compute
Technical University of Denmark
2800 Kgs. Lyngby

Agenda

1. Assumptions and how to check them
2. Two sample t-test
   - Pooled variance set up
   - Welch set up
3. Checking assumptions
4. Two sample t-test as an LM
5. The paired t-test
6. Planning: Sample size and power

Overview

Assumption and how to check them

Using the t-test we assume that $Y_i \sim N(\mu, \sigma^2)$ and iid., i.e. we in general need check

1. The distribution assumption
2. The independence assumption
3. The indentical distribution assumption

1) check the distribution of the data against the assumed distribution (the normal), 2) should be checked if there is an ordering in data, 3) will be discussed later in the course

The distribution assumption

The distribution assumption can be checked by

- Compare the histogram with the relevant density
- Compare den empirical distribution and the expected distribution
- Compare expected and empirical quantiles (QQ-plot)

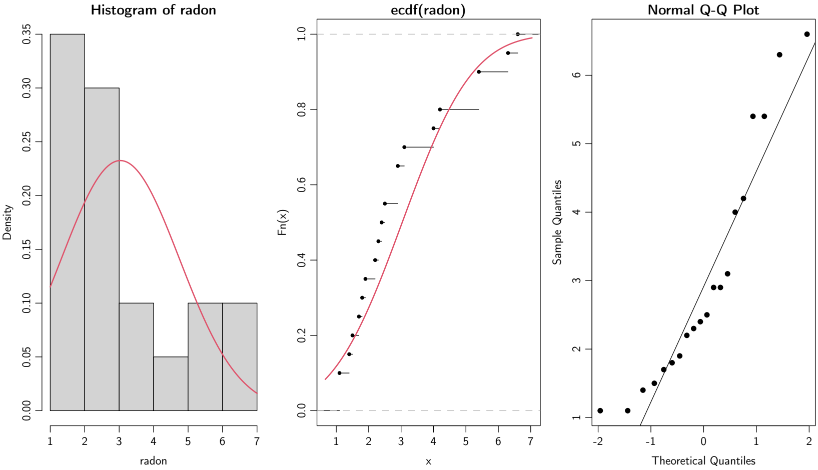
The normal qq-plot

The ordered observations $y_{(1)},...,y_{(n)}$ are plotted against a set of normal quantiles $z_{p_1},...,z_{p_n}$, different choices for p_i exist, the simplest one is

$$p_i = \frac{i - 0.5}{n}$$

points should be on a straight line.

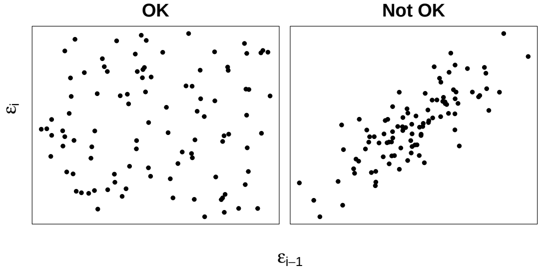
The distribution assumption: Example



The independence assumption

In general difficult to check, but if data are taken as a time series then serial correlation should be checked by

- Plot y_{t+1} against y_t (should appear random)
- Check the observed correlation between y_{t+1} and y_t, should be small (see Chapter 9 for a test).
- If there is a high temporal correlation, it should be included in the model (not covered in this course)

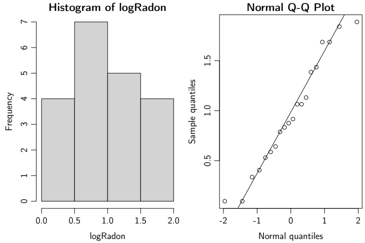


Transformation towards normality

When the normal assumption is not fulfilled, then it will sometimes be possible to transform data. I.e. the model is now

$$g(Y_i) \sim N(\mu, \sigma^2) \quad \text{and iid.}$$

the most common transformation is $g(Y) = \log(Y)$, but power transformations ($g(Y) = Y^\gamma$) are also common.



Note that quantiles are invariant to transformations, while the mean is not.

Example: Skive fjord

- Provide a confidence interval for the chlorophyll concentration in Skive fjord.
- Check the assumptions, and adjust the model if possible.

Overview

Motivation

It is often of interest to compare two different distributions, e.g.

- Is there a difference before and after an intervention?
- Is drug A more efficient than drug B?

The general hypothesis

Is there a difference in expected value between population A and population B?

Or more formally let $Y_{1,1},\ldots,Y_{1,n_1}$ be iid. with $Y_{1,i} \sim N(\mu_1,\sigma_1^2)$ and $Y_{2,1},\ldots,Y_{2,n_2}$ be iid. with $Y_{2,i} \sim N(\mu_2,\sigma_2^2)$, test the null hypothesis

$$H_0: \quad \mu_1 - \mu_2 = \delta_0$$

usually $\delta_0 = 0$.

Pooled variance set up

Let $Y_{1,1},\ldots,Y_{1,n_1}$ be iid. with $Y_{1,i} \sim N(\mu_1,\sigma^2)$ and $Y_{2,1},\ldots,Y_{2,n_2}$ be iid. with $Y_{2,i} \sim N(\mu_2,\sigma^2)$, test the hypothesis $\mu_1 - \mu_2 = \delta_0$. We will consider the test statistics

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta_0}{S_{\bar{Y}_1 - \bar{Y}_2}}$$

In an informal way we can write this as

$$T = \frac{\text{observation-hypothesis}}{\text{standard deviation under assumptions}}$$

We need to establish the distribution of T.

Pooled variance set up: distribution

Under the assumptions (equal variance), the best estimator for the variance is

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

and further

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

If the hypothesis is also true then

$$\bar{Y}_1 - \bar{Y}_2 - \delta_0 \sim N\left(0, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

Pooled variance set up: distribution

Assume that $Y_{1,i} \sim N(\mu_1, \sigma_1^2)$ and $Y_{2,j} \sim N(\mu_2, \sigma_2^2)$, with $\sigma_1^2 = \sigma_2^2$. Then the pooled two-sample test statistic, for the hypothesis $\mu_1 - \mu_2 = \delta_0$. seen as a random variable (Theorem 3.54, Example 2.85 og Exercise 2.16):

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$$

follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Note here that we use the assumption

$$\sigma_1^2 = \sigma_2^2$$

and the null hypothesis

$$E[Y_1] - E[Y_2] = \mu_1 - \mu_2 = \delta_0$$

.

Pooled variance, test statistics and confidence interval

Based on the distribution of T we find the test statistics

Test statistics and p-value

$$t_{\text{obs}} = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}; \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2},$$

and

$$p-value = 2P(T > |t_{obs}|) = 2(1 - P(T < |t_{obs}|); \quad T \sim t(n_1+n_2-2)$$

$(1-\alpha)$-confidence interval for δ

$$CI(\alpha) = \bar{y}_1 - \bar{y}_2 \pm t_{1-\alpha/2} s_p \sqrt{\frac{1}{n_2} + \frac{1}{n_2}}$$

where $t_{1-\alpha/2}$ is based on the t-distribution with n_1+n_2-2 degrees of freedom.

Example[1]

An aircraft manufacturer uses an expensive type of screws in the production of a certain model. To reduce production costs, the manufacturer considers replacing the expensive screws with a cheaper type of screws. Therefore, the manufacturer tests the tensile strength (MPa) of the two types of screws, and the results are shown in the below table.

Tensile strengt	Cheap	Expensive
Sample mean (MPa)	1250	1300
Sample standard deviation (MPa)	54.24	28.54
Sample size	25000	15000

Assume equal variance in the two groups.

- Assuming the samples were completely random, what is the 95% confidence interval for the difference in mean tensile strengths (mean of the cheap type minus mean of the expensive type) based on the test results?
- Under the null hypothesis $H_0 : \mu_{\text{cheap}} - \mu_{\text{expensive}} = -50$, what is the observed test statistic based on the test results?
- Are the assumptions reasonable?

[1] 2024 June

Welch two sample set up

Assume that $Y_{1,i} \sim N(\mu_1, \sigma_1^2)$ and $Y_{2,j} \sim N(\mu_2, \sigma_2^2)$. The (Welch) two-sample statistic seen as a random variable:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

approximately, under the null hypothesis, follows a t-distribution with ν degrees of freedom, where

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

if the two population distributions are normal or if the two sample sizes are large enough.

Welch, test statistics and confidence interval

Based on the distribution of T we find the test statistics

Test statistics and p-value

$$t_{\text{obs}} = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_1^2/n_2}};,$$

and

$$p-value = 2P(T > |t_{obs}|) = 2(1 - P(T < |t_{obs}|); \quad T \sim t(v)$$

$(1-\alpha)$-confidence interval for δ

$$CI(\alpha) = \bar{y}_1 - \bar{y}_2 \pm t_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_2}}$$

where $t_{1-\alpha/2}$ is based on the t-distribution with v degrees of freedom.

Example[2]

An aircraft manufacturer uses an expensive type of screws in the production of a certain model. To reduce production costs, the manufacturer considers replacing the expensive screws with a cheaper type of screws. Therefore, the manufacturer tests the tensile strength (MPa) of the two types of screws, and the results are shown in the below table.

Tensile strengt	Cheap	Expensive
Sample mean (MPa)	1250	1300
Sample standard deviation (MPa)	54.24	28.54
Sample size	25000	15000

- Assuming the samples were completely random, what is the 95% confidence interval for the difference in mean tensile strengths (mean of the cheap type minus mean of the expensive type) based on the test results?

[2] 2024 June

Welch vs. Pooled variance version

The Welch t-test have fewer assumption and is therefore usually preferred

- if $s_1^2 = s_2^2$ the Welch and the Pooled test statistics are the same.
- Only when the two variances become really different the two test-statistics may differ in any important way, and if this is the case, we would not tend to favour the pooled version, since the assumption of equal variances appears questionable then.
- Only for cases with a small sample sizes in at least one of the two groups the pooled approach may provide slightly higher power if you believe in the equal variance assumption. And for these cases the Welch approach is then a somewhat cautious approach.

Overlapping confidence intervals

We are given two 95% confidence intervals

$$CI_1 = [0, 4]; \quad CI_2 = [3, 7]$$

each based on $n = 100$ observations.

- Using significance level $\alpha = 0.05$ is there a difference in mean between the two groups?

Overlapping confidence intervals

We are given two 95% confidence intervals

$$CI_1 = [0,4]; \quad CI_2 = [3,7]$$

each based on $n = 100$ observations.

- Using significance level $\alpha = 0.05$ is there a difference in mean between the two groups?

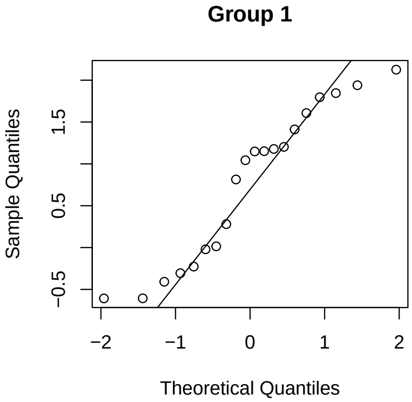
Variances are additive while standard deviations are not!

Overview

Checking assumptions

Welch

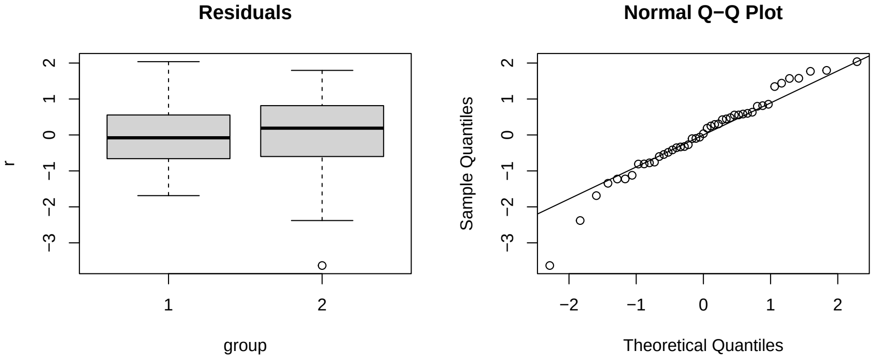
If the Welch two sample test is used then the normality assumption should be checked within each group (normal qq-plots).



Checking assumptions

Pooled variance

If the pooled variance two sample test is used then in addition the equal variance assumption should also be checked (eg. box-plots or compare observed variances). (see LM formulation for another way)



Overview

Two sample t-test as an LM

The two sample t-test, assumening equal variance in the two groups, can be written as an LM

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\sigma}^2\boldsymbol{I}).$$

where $\boldsymbol{X} \in \mathbb{R}^{(n_1+n_2)\times 2}$, while the projection matrix $(\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)$ is unique the design matrix $\boldsymbol{X}$ is not. Some parametrizations are

$$\boldsymbol{X}_1 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{bmatrix}; \quad \boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} \end{bmatrix}; \quad \boldsymbol{X}_3 = \begin{bmatrix} \mathbf{1}_{n_1} & -\frac{1}{2}\mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} & \frac{1}{2}\mathbf{1}_{n_2} \end{bmatrix}$$

- Different parmaparametrization result in different parameter interpretations, while the fitted values are unaffected.

Two sample t-test as an LM

Some questions we could ask is

- What is the parameter interpretation in each of the parametrizations?
- Are some of the parametrizations orthogonal?
- What is the projection matrix?

Two sample t-test as an LM

Let $Y_{1,1},...,Y_{1,n_1}$ be iid. with $Y_{1,i} \sim N(\mu_1, \sigma^2)$ and $Y_{2,1},...,Y_{2,n_2}$ be iid. with $Y_{2,i} \sim N(\mu_2, \sigma^2)$. There are two sensible, and nested, hypothesis

$$H_1: \quad \mu_1 = \mu_2 = \mu$$

and

$$H_0: \quad \mu_1 = \mu_2 = 0.$$

If H_0 is true then

$$\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

If H_1 is true then

$$\boldsymbol{Y} \sim N(\boldsymbol{1}\mu, \sigma^2 \boldsymbol{I})$$

Under the assumption we have

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

Separation of variation

The orthogonal partitioning of variation can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y},$$

with $\boldsymbol{H}_0 = \frac{1}{n_1+n_2}\boldsymbol{E}$, and $\boldsymbol{H}_1 = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. If H_1 is true then

$$F_1 = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_1 - \boldsymbol{H}_0)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} \sim F(1, n-2),$$

and further if H_0 is true then

$$F_0 = \frac{\boldsymbol{Y}^T\boldsymbol{H}_0\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}/(n-2)} \sim F(1, n-2).$$

Finally a cantral estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}}{n-2}$$

Residuals and fitted values

The fitted values are given by

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

The residuals are

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

Under the assumption of the model the distribution of the residuals is

$$\boldsymbol{r} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})).$$

Hence even though ε_i are iid. r_i are, in general, not!

Example:

- Under what conditions are r_i identically distributed in the two sample setup?
- Is r_i and r_j independent?

Residuals analysis

We have $r_i \sim N(0, \sigma^2(1 - h_{ii}))$, and hence

$$\tilde{r}_i = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

are indentically distributed. The assumptions can be checked using the $\tilde{r}_i$'s by

- box plot per group (equal variance in groups)
- qq-plot (normal assumption)

If the equal variance assumption is violated one may use Welch, if normal assumption is violated one may use data transformation (or other type of models, not covered here).

Example: Skive Fjord

Compare the water temperature before and after 1995, include a residual analysis.

Overview

Paired t-test

- We consider situations where we want to compared two mean values, but where data is paired
- Hypothesis tests is done by investigating the difference, D_i, between the paired observations:

$$D_i = Y_{1,i} - Y_{2,i} \quad \text{for } i = 1, 2, ..., n$$

I.e. a one sample situation.

Paried versus independent experiment

Completely Randomized (independent samples)

20 patients are used and completely at random allocated to one of the two treatments (but usually making sure to have 10 patients in each group). So: different persons in the different groups.

Paired (dependent samples)

10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments to make sure that it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random). So: the same persons in the different groups.

Example: Skive fjord

Test if there is a difference between nitrate input to Skive fjord for the years 1999 and 2006.

Overview

Planning, CI - formulation

The one sample confidence interval is

$$CI = \bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}} = \bar{x} \pm ME$$

If σ is known we have

$$CI = \bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = \bar{x} \pm ME$$

For known σ and wanted ME we can solve for n

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{ME} \right)^2$$

Hence we can plan according to a wanted margin of error.

Power

The significance level (α) is about the probability of rejecting a true hypothesis. Power is the probability of detecting an effect

$$Power = 1 - \beta = P(\text{Rejecting the } H_0 \text{ when } H_0 \text{ is false})$$

Challange: The null hypotehsis can be wrong in many ways.

- It is harder to detect a small effect than a large effect.

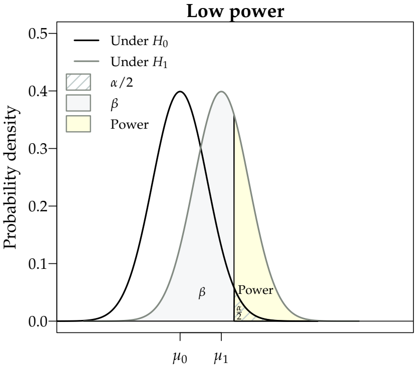
Assume that $Y_i \sim N(\mu, \sigma^2)$, we make the null hypothesis $H_0 : \mu = \mu_0$.

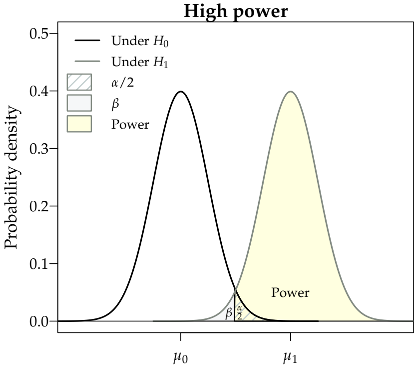
- If $\frac{\mu - \mu_0}{\sigma/\sqrt{n}}$ is large then the probability of rejecting H_0 is large
- If $\frac{\mu - \mu_0}{\sigma/\sqrt{n}}$ is small then the probability of rejecting H_0 is small

Power

Two possible truths against two possible conclusions:

	Rejecting H_0	Not rejecting H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0	Type II error (β)





Planning, Sample size n

What should n be?

The sample should be big enough to detect a relevant effect with high power (usually at least 80%):

Metode 3.65: Sample size for one-sample t-test:

One-sample t-test for given α, β and σ:

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2$$

Where $\mu_0 - \mu_1$ is the change in means that we would want to detect and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Planning and Sample size n in Python

Go to Python for finding sample size in 1 one and two sample set up.

Agenda

1. Assumptions and how to check them
2. Two sample t-test
   - Pooled variance set up
   - Welch set up
3. Checking assumptions
4. Two sample t-test as an LM
5. The paired t-test
6. Planning: Sample size and power