Kursus 02402/02323 Introducerende Statistik

Forelæsning 6: Sammenligning af to grupper

Klaus K. Andersen og Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: klaus@cancer.dk

Oversigt

Oversigt

Motiverende eksempel - energiforbrug

Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge om der er en forskel i energiforbrug for forskellige typer (moderat fysisk krævende) arbejde. In the study, the energy usage of 9 nurses from hospital A and 9 (other) nurses from hospital B have been measured. The measurements are given in the following table in mega Joule (MJ):

Stikprøve fra hver hospital, $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \ \mu_1 = \mu_2$$

Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \ \mu_1 = \mu_2$$

Sample means og standard deviations:

$$\hat{\mu}_A = \bar{x}_A = 8.293, \ (s_A = 1.428)$$
$$\hat{\mu}_B = \bar{x}_B = 10.298, \ (s_B = 1.398)$$

Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0 : \ \mu_1 = \mu_2$$

Sample means og standard deviations:

$$\hat{\mu}_A = \bar{x}_A = 8.293, \ (s_A = 1.428)$$
$$\hat{\mu}_B = \bar{x}_B = 10.298, \ (s_B = 1.398)$$

Er data i overenstemmelse med nulhyposen H_0?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Nulhypotese: $H_0 : \ \mu_B - \mu_A = 0$

NYT:p-**værdi for forskel:**

$$p - \text{værdi} = 0.0083$$

(Beregnet under det scenarie, at H_0 er sand)

NYT:**Konfidensinterval for forskel**:

$$2.005 \pm 1.412 = [0.59; \ 3.42]$$

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Definition af hypotesetest og signifikans (Repetition)

Definition 3.23. Hypotesetest:

We say that we carry out a hypothesis test when we decide against a null hypothesis or not using the data.

A null hypothesis is *rejected* if the p-value, calculated after the data has been observed, is less than some α, that is if the p-value $< \alpha$, where α is some pre-specifed (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Definition 3.28. Statistisk signifikans:

An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α.
(OFTE bruges $\alpha = 0.05$)

Metode 3.36. Steps ved hypotesetests - et overblik (Repetition)

Helt generelt består et hypotesetest af følgende trin:

1. Formulate the hypotheses and choose the level of significance α (choose the "risk-level")
2. Calculate, using the data, the value of the test statistic
3. Calculate the p-value using the test statistic and the relevant sampling distribution, and compare the p-value and the significance level α and make a conclusion
4. (Alternatively, make a conclusion based on the relevant critical value(s))

Definition og fortolkning af p-værdien (Repetition)

p-værdien udtrykker *evidence* imod nulhypotesen – Tabel 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Definition 3.21 af p-værdien:

The p-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Metode 3.58: Two-sample t-test

beregning af teststørrelsen

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0: \ \delta = \delta_0$$

the (Welch) two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Theorem 3.59: Fordelingen af (Welch) t-teststørrelsen

Welch t-teststørrelsen er t-fordelt

The (Welch) two-sample statistic seen as a random variable:

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

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.

Metode 3.60: Two-sample t-test

Et level α test er

1. Compute t_{obs} and ν as given above.
2. Compute the evidence against the *null hypothesis*[a] $H_0 : \mu_1 - \mu_2 = \delta$ vs. the *alternative hypothesis* $H_1 : \mu_1 - \mu_2 \neq \delta$ by the

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

   where the t-distribution with ν degrees of freedom is used.
3. If p–value $< \alpha$: We reject H_0, otherwise we accept H_0.
4. The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s) $\pm t_{1-\alpha/2}$:
   If $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0.

[a] We are often interested in the test where $\delta = 0$

Metode 3.61: Det ensidede two-sample t-test

Et level α ensidet "less" test er

1. Compute t_{obs} and ν as given above.
2. Compute the evidence against the *null hypothesis* $H_0 : \mu_1 - \mu_2 \geq \delta$ vs. the *alternative hypothesis* $H_1 : \mu_1 - \mu_2 < \delta$ by the

$$p\text{–value} = P(T < t_{\text{obs}})$$

   where the t-distribution with ν degrees of freedom is used.
3. If p–value $< \alpha$: We reject H_0, otherwise we accept H_0.
4. The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value t_α:
   If $t_{\text{obs}} < t_\alpha$ we reject H_0, otherwise we accept H_0.

Metode 3.62: Det ensidede two-sample t-test

Et level α ensidet "greater" test er

1. Compute t_{obs} and ν as given above.
2. Compute the evidence against the *null hypothesis* $H_0 : \mu_1 - \mu_2 \leq \delta$ vs. the *alternative hypothesis* $H_1 : \mu_1 - \mu_2 > \delta$ by the

$$p\text{–value} = P(T > t_{\text{obs}})$$

where the t-distribution with ν degrees of freedom is used.
3. If p–value $< \alpha$: We <u>reject H_0</u>, otherwise we <u>accept H_0</u>.
4. The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value $t_{1-\alpha}$:
If $t_{\text{obs}} > t_{1-\alpha}$ we <u>reject H_0</u>, otherwise we <u>accept H_0</u>.

Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \ \delta = \mu_B - \mu_A = 0$$

versus the non-directional(= two-sided) alternative:

$$H_0: \ \delta = \mu_B - \mu_A \neq 0$$

Først beregninger af t_{obs} og ν:

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

and

$$\nu = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

Eksempel - energiforbrug

Dernæst findes p-værdien:

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
1 - pt(3.01, df = 15.99)

## [1] 0.0041545
```

Eksempel - energiforbrug

Dernæst findes p-værdien:

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```r
1 - pt(3.01, df = 15.99)

## [1] 0.0041545
```

Vurder evidencen (Tabel 3.1):

Der er stærk evidence imod nulhypotesen.

Eksempel - energiforbrug

Dernæst findes p-værdien:

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
1 - pt(3.01, df = 15.99)

## [1] 0.0041545
```

Vurder evidencen (Tabel 3.1):

Der er stærk evidence imod nulhypotesen.

Konkluder baseret på $\alpha = 0.05$:

Vi forkaster nulhypotesen, der er signifikant forskel på grupperne - sygeplejersker på Hospital B kan siges at have et større (middel)energiforbrug end sygeplejesker på Hospital A.

Oversigt

Metode 3.69: Konfidensinterval for $\mu_1 - \mu_2$

Konfidensintervallet for middelforskelen bliver:

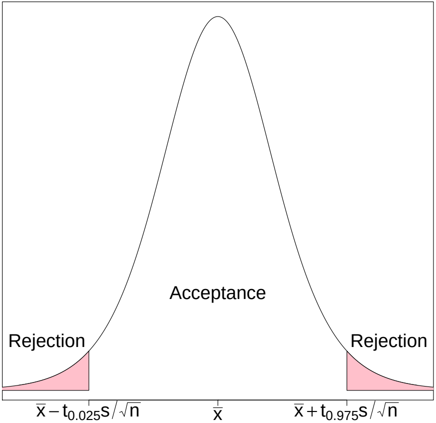
For two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$-quantile from the t-distribution with ν degrees of freedom given from equation (3.26) (as above).

Konfidensinterval og hypotesetest (Repetition)

Acceptområdet er de mulige værdier for μ som ikke ligger for langt væk fra data:



DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - energiforbrug - det hele i R:

Let us find the 95% confidence interval for $\mu_B - \mu_A$. Since the relevant t-quantile is, using $\nu = 15.99$,

$$t_{0.975} = 2.120$$

the confidence interval becomes:

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}$$

which then gives the result as also seen above:

$$[0.59;\ 3.42]$$

Eksempel - energiforbrug - det hele i R:

```
xA=c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB=c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data:  xB and xA
## t = 3.0091, df = 15.993, p-value = 0.008323
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  0.59228 3.41661
## sample estimates:
## mean of x mean of y
##   10.2978    8.2933
```

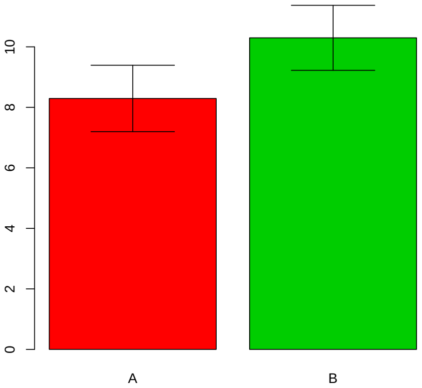
Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - energiforbrug - Præsentation af resultat

Barplot med *error bars* ses ofte

Et grupperet barplot med nogle "error bars" - herunder er 95%-konfidensintervallerne for hver gruppe vist:



Vær varsom med at bruge "overlappende konfidensintervaller"

Man bruger faktisk så ikke den rigtige variation til at vurdere forskellen:

$$\text{Stand. dev. of } (\bar{X}_A - \bar{X}_B) \neq \text{Stand. dev. of } \bar{X}_A + \text{Stand. dev. of } \bar{X}_B$$

$$\text{Var } (\bar{X}_A - \bar{X}_B) = \text{Var } (\bar{X}_A) + \text{Var } (\bar{X}_B)$$

Antag at de to standard-errors er 3 og 4: Summen er7, men $\sqrt{3^2 + 4^2} = 5$

Vær varsom med at bruge ”overlappende konfidensintervaller”

Man bruger faktisk så ikke den rigtige variation til at vurdere forskellen:

$$\text{Stand. dev. of } (\bar{X}_A - \bar{X}_B) \neq \text{Stand. dev. of } \bar{X}_A + \text{Stand. dev. of } \bar{X}_B$$

$$\text{Var } (\bar{X}_A - \bar{X}_B) = \text{Var } (\bar{X}_A) + \text{Var } (\bar{X}_B)$$

Antag at de to standard-errors er 3 og 4: Summen er7, men $\sqrt{3^2+4^2} = 5$

Det korrekte forhold mellem de to er således:

$$\text{Stand. dev. of } (\bar{X}_A - \bar{X}_B) < \text{Stand. dev. of } \bar{X}_A + \text{Stand. dev. of } \bar{X}_B$$

Vær varsom med at bruge "overlappende konfidensintervaller"

Remark 3.73. Regel for brug af "overlappende konfidensintervaller":

When two CIs do NOT overlap: The two groups are significantly different

When two CIs DO overlap: We do not know what the conclusion is

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Motiverende eksempel - sovemedicin

Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemidler A og B. For 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid (i timer) (Forskellen på effekten af de to midler er angivet):

Stikprøve, $n = 10$:

person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

Parret setup og analyse = one-sample analyse

```
x1=c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2=c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)
dif=x2-x1
t.test(dif)

##
##  One Sample t-test
##
## data:  dif
## t = 4.6716, df = 9, p-value = 0.001166
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean of x
##      1.67
```

Parret setup og analyse = one-sample analyse

```
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 4.6716, df = 9, p-value = 0.001166
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean of the differences
##                    1.67
```

Parret versus independent eksperiment

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)

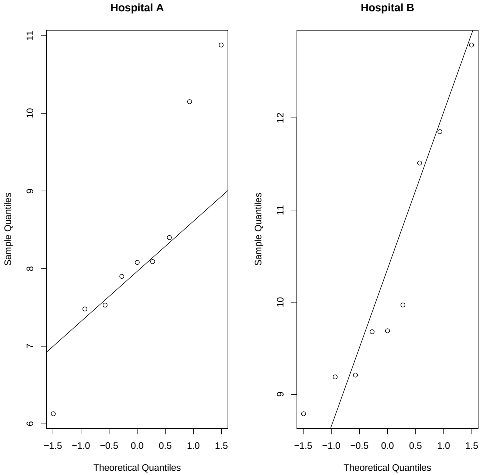
10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments to make sure that it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random).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

Eksempel - Sovemedicin - FORKERT analyse

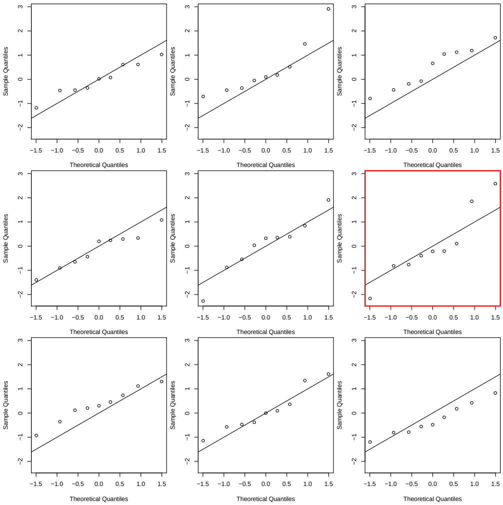
```
t.test(x1,x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -1.9334, df = 17.9, p-value = 0.06916
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -3.48539  0.14539
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```

Oversigt

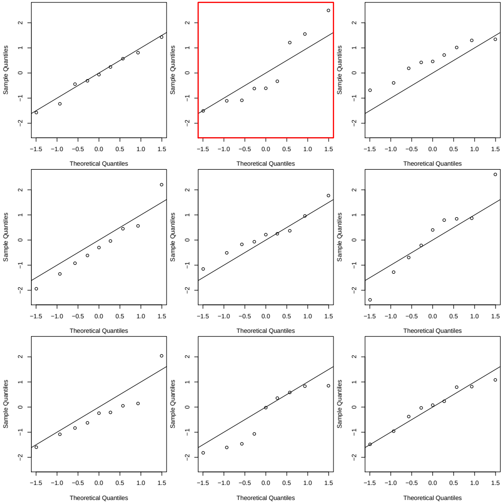
Eksempel - Q-Q plot inden for hver stikprøve:



Eksempel - Sammenligning med simulerede, A



Eksempel - Sammenligning med simulerede, B



Oversigt

Styrke og stikprøvestørrelse - two-sample

Finding the power of detecting a group difference of 2 with $\sigma = 1$ for $n = 10$:

```
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.98818
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Styrke og stikprøvestørrelse - two-sample

Finding the sample size for detecting a group difference of 2 with $\sigma = 1$ and power$= 0.9$:

```
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 6.3868
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Styrke og stikprøvestørrelse - two-sample

Finding the detectable effect size (delta) with $\sigma = 1$, $n = 10$ and power$= 0.9$:

```
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 1.5337
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Oversigt

Metode 3.64: The pooled two-sample t-test statistic

Beregning af den poolede teststørrelse (og 3.63)

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0 : \ \delta = \delta_0$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

Theorem 3.65: Fordelingen af den poolede test-størrelse

er en t-fordeling

The pooled two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}} \tag{1}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

Vi bruger altid ''Welch' versionen

Nogenlunde (idiot)sikkert at bruge Welch-versionen altid

- 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.

Oversigt