Kursus 02323: Introducerende Statistik

Forelæsning 6: Sammenligning af to populationer

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2021

Kapitel 3: Statistik for to populationer (2 stikprøver)

Specifikke metoder, to populationer:

- Konfidensinterval for forskel i middelværdi
- Test for forskel i middelværdi (t-test)
- To PARREDE grupper: "Tag differencen" $\Rightarrow$ "Én gruppe"

Chapter 3: Two Samples

Specific methods, two samples:

- Confidence interval for the mean difference
- Test for the mean difference (t-test)
- Two PAIRED samples: "Take difference" $\Rightarrow$ "One sample"

Oversigt

1. Motiverende eksempel - energiforbrug
2. Hypotesetest (Repetition)
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions
8. The pooled t-test - a possible alternative

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:

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

Sample means og standard deviations:

$$\hat{\mu}_1 = \bar{x}_1 = 8.293, \ (s_1 = 1.428)$$
$$\hat{\mu}_2 = \bar{x}_2 = 10.298, \ (s_2 = 1.398)$$

Er data i overenstemmelse med nulhyposen H_0?

$$\text{Data: } \bar{x}_2 - \bar{x}_1 = 2.005$$

$$\text{Nulhypotese: } H_0: \ \mu_2 - \mu_1 = 0$$

NYT: p-**værdi for forskel:**

$$p\text{-værdi} = 0.0083$$

(Beregnet under det scenarie, at H_0 er sand)

NYT:**Konfidensinterval for forskel**:

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

Steps ved hypotesetests - et overblik (repetition)

Helt generelt består et hypotesetest af følgende trin:

1. Formuler hypoteserne (H_0 og H_1) og vælg signifikansniveau α (choose the "risk-level")
2. Beregn med data værdien af teststatistikken
3. Beregn p-værdien med teststatistikken og den relevante fordeling, og sammenlign p-værdien med signifikansniveauet og drag en konklusion

   eller

   Lav konklusionen ved de relevante kristiske værdier

Definition og fortolkning af p-værdien (repetition)

Definition 3.22 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.

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

Metode 3.49: Two-sample t-test

Beregning af teststørrelsen

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

$$\delta = \mu_2 - \mu_1 \quad \textit{(delta er forskellen i middelværdi)}$$
$$H_0: \ \delta = \delta_0 \quad \textit{(typisk er } \delta_0 = 0\textit{)}$$

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 3.50: 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 3.51: The level α two-sample t-test

1. Compute the test statistic using Equation (3-48) and ν from Equation (3-50)

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \text{ and } \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

2. Compute the evidence against the *null hypothesis*

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

vs. the *alternative hypothesis*

$$H_1: \quad \mu_1 - \mu_2 \neq \delta_0,$$

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,
or
The rejection/acceptance conclusion can equivalently be 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

Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

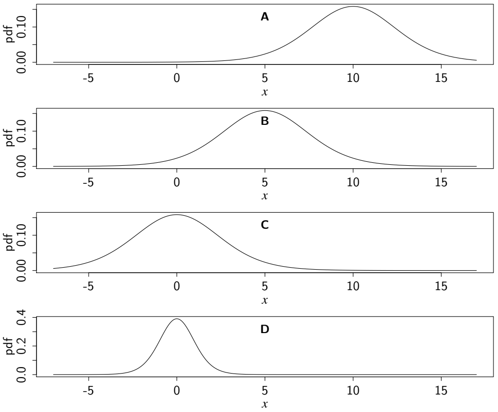
Hvilken af pdf'erne repræsenterer fordelingen af forskellen i stikprøvegennemsnit?

$$\bar{X}_2 - \bar{X}_1$$

UNDER *(dvs. antag er sand)*:

$$H_0 : \delta = 10$$

(sample sizes $n_1 = 7$ and $n_2 = 8$)
(sample std. dev. $s_1 = 18$ and $s_2 = 24$)



A B C eller D? Svar: A

Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

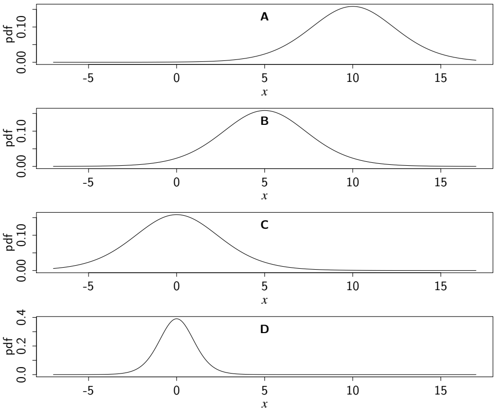
Hvilken af pdf'erne repræsenterer fordelingen af

$$\bar{X}_2 - \bar{X}_1 - \delta_0$$

under:

$$H_0 : \delta = 10$$

(sample sizes $n_1 = 7$ and $n_2 = 8$)
(sample std. dev. $s_1 = 18$ and $s_2 = 24$)



A B C eller D? Svar: C

Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

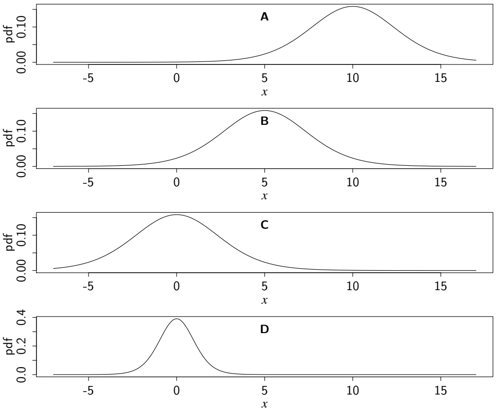
Hvilken af pdf'erne repræsenterer fordelingen af

$$T = \frac{\bar{X}_2 - \bar{X}_1 - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

under:

$$H_0 : \delta = 10$$

(sample sizes $n_1 = 7$ and $n_2 = 8$)
(sample std. dev. $s_1 = 18$ and $s_2 = 24$)



A B C eller D? Svar: D

Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt

$$H_0: \ \delta = \mu_2 - \mu_1 = 0$$

versus the alternative

$$H_1: \ \delta = \mu_2 - \mu_1 \neq 0$$

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:

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

```r
## p-værdi for nulhypotese om ingen forskel mellem sygeplejeskers energiforbrug
2 * (1 - pt(3.01, df = 15.99))

## [1] 0.0083
```

Eksempel - energiforbrug - brug funktion i R:

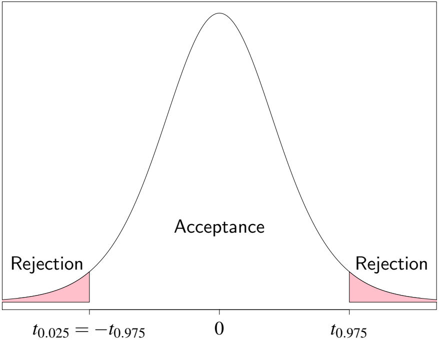
```
################################
## t-test for forskel i middelværdi på sygeplejeskers energiforbrug
xA <- c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB <- c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)
## Default i t.test() er H_0: mu_1 = mu_2 (ingen forskel i middelværdi)
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data:  xB and xA
## t = 3.009, df = 15.99, p-value = 0.00832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.59228 3.41661
## sample estimates:
## mean of x mean of y
##  10.29778   8.29333
```

I pausen: Installer *Space Frontier* (ikke 2'eren) på jeres device (Android eller iphone), men vent med at spille.

(Spring igennem menuer, så I ikke giver dem data)!

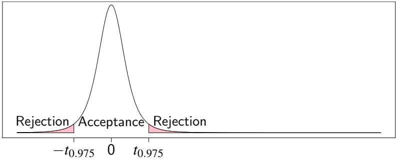
Kritiske værdier og hypotesetest

Acceptområdet er værdier for teststatistikken t_{obs} som ligger indenfor de kritiske værdier:



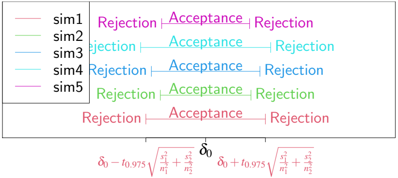
Den standardiserede skala

Hvis t_{obs} er i acceptområdet, så accepteres H_0



Den egentlige skala

Hvis $\bar{x} - \bar{y}$ er i acceptområdet, så accepteres H_0



Metode 3.47: Konfidensinterval for $\mu_1 - \mu_2$

Konfidensintervallet for middelforskellen bliver:

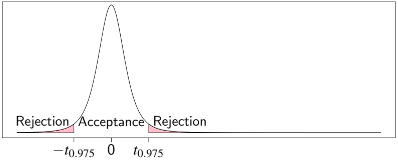
For two samples $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

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

where $t_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$-quantile from the t-distribution with ν degrees of freedom given from Equation (3-50).

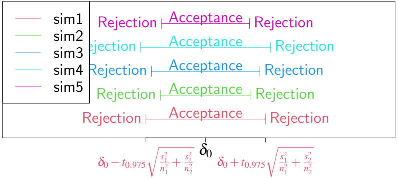
Den standardiserede skala

Hvis t_{obs} er i acceptområdet, så accepteres H_0



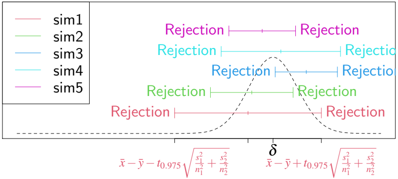
Den egentlige skala

Hvis $\bar{x} - \bar{y}$ er i acceptområdet, så accepteres H_0



Konfidensintervallet

Nulhypoteser med δ_0 udenfor konfidensintervallet ville være blevet afvist



Eksempel - energiforbrug - det hele i R:

Let us find the 95% confidence interval for $\mu_2 - \mu_1$:

Since the relevant t-quantile is, using $\nu = 15.99$,

$$t_{0.975} = 2.120$$

the confidence interval becomes:

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

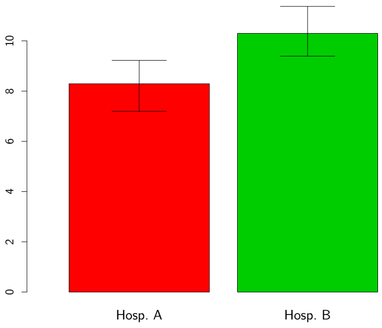
which then gives the result as also seen above:

$$[0.59;\ 3.42]$$

Eksempel - energiforbrug - 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"

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

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: Brug one-sample analyse

```r
## Det parrede setup: Tag forskellen og brug one-sample test
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 = 5, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of x
##       1.7
```

Parret setup og analyse: Brug one-sample analyse

```
## Eller angiv at testen er parret med "paired=TRUE"
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of the differences
##                     1.7
```

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). Hence: *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). Hence: *the same persons in the different groups*.

Eksempel - Sovemedicin - FORKERT analyse

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -2, df = 18, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.49  0.15
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```

Undersøgelse af computerspil

Undersøgelse om et computerspil er designet så man forbedrer sig når man spiller:

- Forsøg: Personer spiller samme bane i spillet tre gange i træk
- Nogle har spillet det før og er derfor erfarne. Alle angiver deres erfaring ved: 'nybegynder', 'mellem' og 'øvet'
- Scoren måles for hver person de tre gange de spiller banen

Der testes for forskellen mellem *nybegyndere* og *øvede personer*:

Hvilket setup skal benyttes? A: Parret B: Ikke parret C: Ved ikke

Svar) B: Ikke parret

Der testes for forskellen i score *fra første til tredje gang de spiller banen*:

Hvilket setup skal benyttes? A: Parret B: Ikke parret C: Ved ikke

Svar) A: Parret

Undersøgelse af computerspil

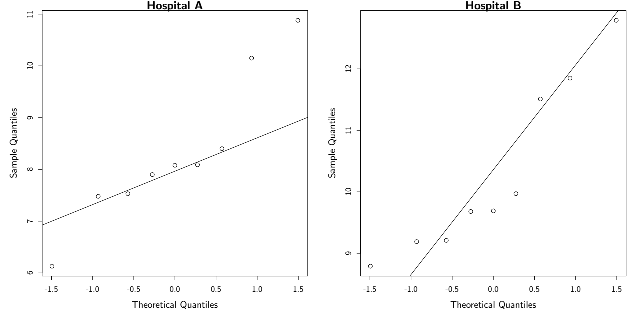
Gå ind i første level og spil i så indtil der bliver sagt stop. Noter bedste score. Dette gentager vi 3 gange ialt.

Download `"analyserGame.R"` (følg link under uge6 på "Course material"):

- Kan der påvises en signifikant forskel fra *nybegyndere* til *meget øvede* på $\alpha = 5\%$ niveau?
- Kan der påvises en signifikant forbedring mellem første og tredje gang banen spilles på $\alpha = 5\%$ niveau?

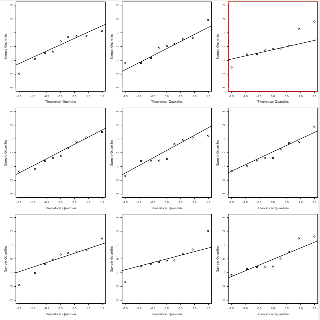
Eksempel: q-q plot inden for hver stikprøve

```
## Check af normalitetsantagelsen med q-q plots
par(mfrow=c(1,2))
qqnorm(xA, main="Hospital A")
qqline(xA)
qqnorm(xB, main="Hospital B")
qqline(xB)
```



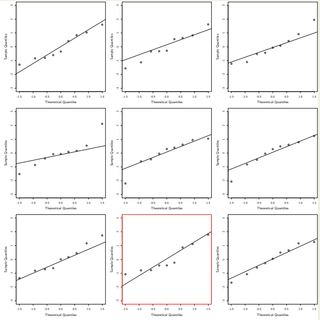
Eksempel - Sammenligning med simulerede, A

```r
## Define the plotting function
qqwrap <- function(x, y, ...){
  stdy <- (y-mean(y))/sd(y)
  qqnorm(stdy, main="", ...)
  qqline(stdy)}
## Do the Wally plot
wallyplot(xA, FUN=qqwrap, ylim=c(-3,3))
```



Eksempel - Sammenligning med simulerede, B

```
## Check af normalitetsantagelsen med q-q plots og Wally-plot
## Do the Wally plot
wallyplot(xB, FUN=qqwrap, ylim=c(-3,3))
```



Metode 3.52: The pooled two-sample estimate of variance

Det poolede variansestimat

Under the assumption that $\sigma_1^2 = \sigma_2^2$ the *pooled* estimate of variance is the weighted average of the two sample variances

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Metode 3.53: The pooled two-sample t-test statistic

Beregning af den poolede teststørrelse

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 3.54: Fordelingen af den poolede teststørrelse

Fordelingen af den poolede teststørrelse er en t-fordeling

The pooled two-sample statistic seen as a random variable

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

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 (den "ikke-poolede")

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