Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 6: Two sample situationer

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality

2. Planlægning: Power og sample size

3. Two-sample t-test og p-værdi

4. Konfidensinterval for forskellen

5. Overlappende konfidensintervaller?

6. Det parrede setup

7. Checking the normality assumptions

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

Eksempel

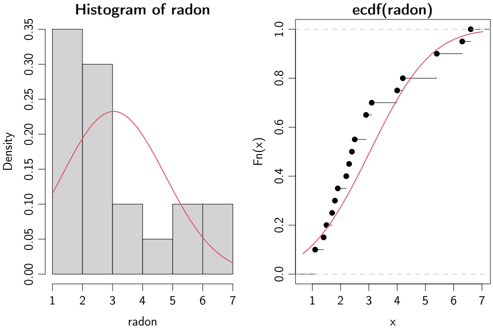
Vi har følgende observationer af radon koncentrationer

```
#################################
## Eksempel radon data
radon <- c(2.4, 4.2, 1.8, 2.5, 5.4, 2.2, 4.0, 1.1, 1.5, 5.4,
           6.3, 1.9, 1.7, 1.1, 6.6, 3.1, 2.3, 1.4, 2.9, 2.9)
```

- Find et 95% konfideninterval for det forventede radon niveau og udfør hypotese testen $H_0 : \mu \geq 3.5$, mod det ensided alternativ $H_1 : \mu < 3.5$.
- Undersøg om modellens forudsætninger er opfyldt.

Eksempel - Radon data

```
hist(radon,freq=FALSE)
curve(dnorm(x, mean=mean(radon),sd=sd(radon)),col=2,lwd=2,add=TRUE)
plot(ecdf(radon))
curve(pnorm(x, mean=mean(radon),sd=sd(radon)),col=2,lwd=2,add=TRUE)
```



Normal Q-Q plot

En bedre metode er at sammenligne de observerede fraktiler direkte med de forventede fraktiler i normal fordelingen.

Metode 3.42 - Den formelle definition

The ordered observations $x_{(1)}, \ldots, x_{(n)}$ are plotted versus a set of expected normal quantiles $z_{p_1}, \ldots, z_{p_n}$. Different definitions of $p_1, \ldots, p_n$ exist:

- In R, when $n > 10$:

$$p_i = \frac{i - 0.5}{n},\ i = 1, \ldots, n$$

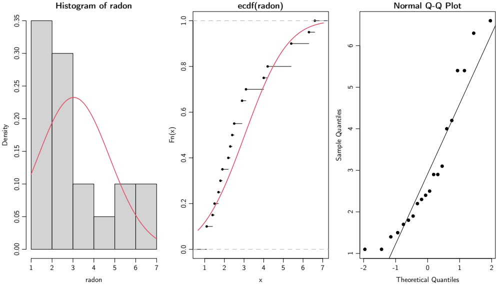
- In R, when $n \leq 10$:

$$p_i = \frac{i - 3/8}{n + 1/4},\ i = 1, \ldots, n$$

Eksempel - Radon data

Punkterne skal ligge på en ret linie

```
qqnorm(radon,pch=19)
qqline(radon)
```


Histogram of radon

ecdf(radon)

Normal Q-Q Plot

Er det for langt fra den rette linie? Wallyplot kan hjælpe...

Hvad hvis data ikke er normafordelte?

I tilfælde hvor normalfordelingen ikke holder kan man prøve at transformere data, nogle muligheder er

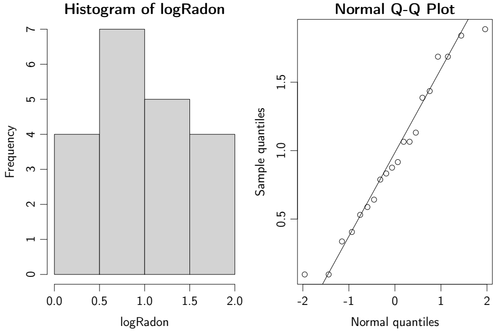
- $\log(x)$, $\sqrt{x}$, og $x^{1/3}$ (gør store værdier mindre, dvs. ved højreskæve fordelinger)
- x^2 eller x^3 (gør store værdier størrer, dvs. ved venstreskæve fordelinger)
- Udfør statistiske test på transformerede data
- OBS: Fraktiler en invariante ifht. (monotone) transformationer, det er middelværdier ikke!

$\log$-transformationen er den klart mest anvendte.

Eksempel - Radon data - log-transformerede data er tættere på normal fordelingen

```
##TRANSFORM USING NATURAL LOGARITHM
logRadon<-log(radon)

hist(logRadon)
qqnorm(logRadon,ylab = 'Sample quantiles',xlab = "Normal quantiles")
qqline(logRadon)
```



Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

Planlægning, KI - formulering

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

Hvis σ er kendt får vi

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

For kendt σ og ønsket ME kan vi løse for n

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

Planlægning, Styrke (=Power)

Hvad er styrken for et kommende studie/eksperiment:

- Sandsynligheden for at opdage en (formodet) effekt
- $P($Forkaste H_0 når H_1 er sand$)$
- Probability of correct rejection of H_0
- Udfordring: Nulhypotesen kan være forkert på mange måder!
- I praksis: Scenarie-baseret approach
    - E.g. "Hvad nu hvis $\mu = 86$, hvor godt vil mit studie være til at opdage dette? "
    - E.g. "Hvad nu hvis $\mu = 84$, hvor godt vil mit studie være til at opdage dette? "
    - etc

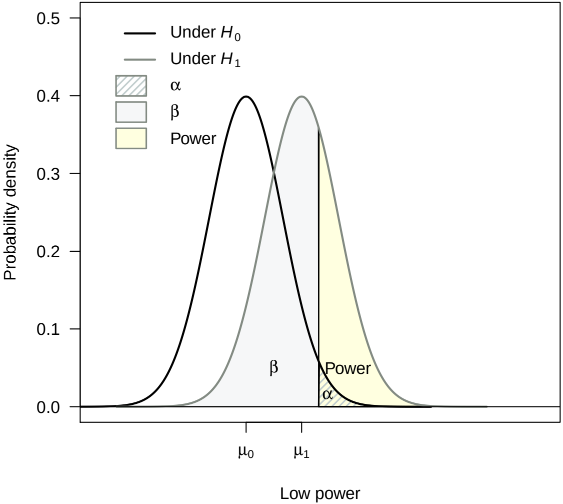
Planlægning, Styrke (=Power)

Når man har fastlagt hvilket test, der skal bruges:

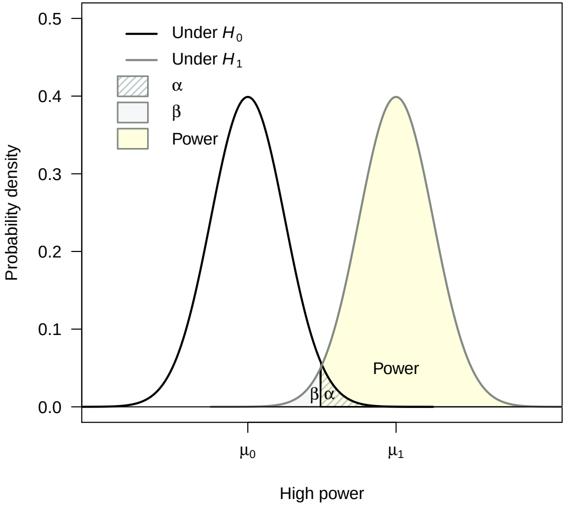
Kender man (eller fastlægger/gætter på) fire ud af følgende fem oplysninger, kan man regne sig frem til den femte:

- Stikprøvestørrelse (sample size) n
- Significance level α of the test.
- A change in mean that you would want to detect (effect size) $\mu_0 - \mu_1$.
- The population standard deviation, σ.
- The power $(1 - \beta)$=P(Forkaste H_0 når H_1 er sand).

Low power eksempel



High power eksempel



Planlægning, Sample size n

Det store spørgsmål i praksis: HVAD skal n være?

Forsøget skal være stort nok til at kunne opdage en relevant effekt med stor power (som regel mindst 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.

Eksempel - The sample size for power$= 0.80$

```
## Stikprøvestørrelse for t-test
power.t.test(power = .80, delta = 4, sd = 12.21,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75.077
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

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. I studiet er energyforbruget for 9 sygeplejersker fra hospital A målt og energyforbruget for 9 (andre) sygeplejersker fra hospital B målt. Målingerne er givet i 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:

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

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**:

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

Sammenvægtet (Pooled) variance set up

Assume that $X_{1,i} \sim N(\mu_1, \sigma^2)$ and $X_{2,j} \sim N(\mu_2, \sigma^2)$. Then the pooled two-sample statistic seen as a random variable (Theorem 3.54, Example 2.85 og Exercise 2.16):

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

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.

The pooled two-sample t-test statistic

Beregning af den poolede teststørrelse (Metode 3.52 og 3.53)

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

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

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

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 3.49: Two-sample t-test

Beregning af teststørrelsen

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

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

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

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

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

Metode 3.51: 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 $\delta = 0$

DTU Compute
Department of Applied Mathematics and Computer Science

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:

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

```r
2 * (1 - pt(3.01, df = 15.99))

## [1] 0.0083089
```

Vurder evidencen (Tabel 3.1):

Der er stærk evidence imod nulhypotesen.

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

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

Konfidensintervallet for middelforskelen bliver:

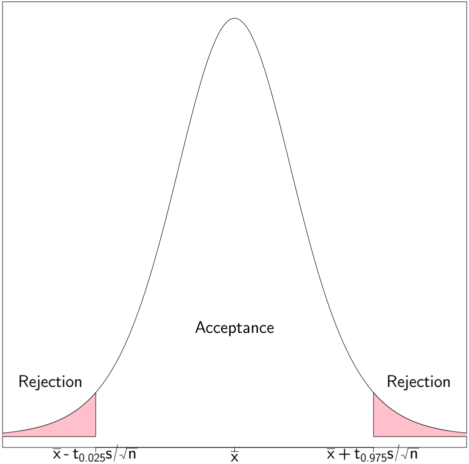
For two samples $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

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

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

Konfidensinterval og hypotesetest (Repetition)

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



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 = 3.01, df = 16, p-value = 0.0083
## 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
```

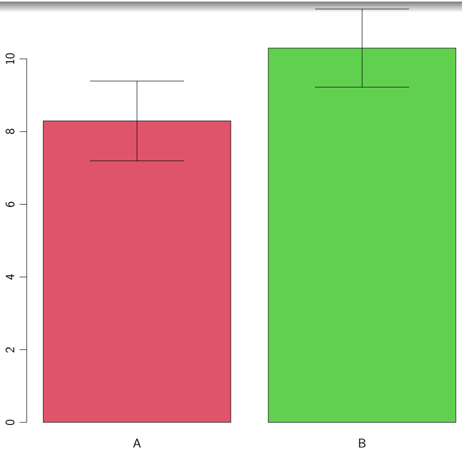
Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

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 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 er 7, men $\sqrt{3^2 + 4^2} = 5$

Det korrekte forhold mellem de to er således:

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

Vi bruger altid ''Welch' versionen

Nogenlunde 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

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

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 t-test

- Vi betragter nu en situation hvor vi vil sammenligne 2 middelværdier, men hvor data er parret
- Hypotesetestet foregår derfor ved at undersøge forskellen, D_i, mellem de parrede observationer:

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

Vi kan herefter beregne middelværdi $\bar{D}$ og varians S_D^2 for D. Test af $\bar{D}$ gøres nu som de sædvanlige test for én middelværdi

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 = 4.67, df = 9, p-value = 0.0012
## 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 = 4.67, df = 9, p-value = 0.0012
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean difference
##            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 = -1.93, df = 17.9, p-value = 0.069
## 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
```

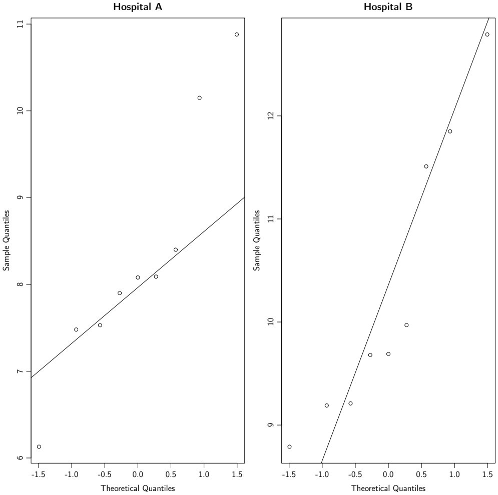
Eksempel:

Test om der er forskel på nitrat udledningen til Skive fjord i 1999 og 2006.

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions

Eksempel - Q-Q plot inden for hver stikprøve:



DTU Compute
Department of Applied Mathematics and Computer Science

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
```

Oversigt

1. Checking the normality assumption
   - The Normal QQ plot
   - Transformation towards normality
2. Planlægning: Power og sample size
3. Two-sample t-test og p-værdi
4. Konfidensinterval for forskellen
5. Overlappende konfidensintervaller?
6. Det parrede setup
7. Checking the normality assumptions