Kursus 02402/02323 Introducerende Statistik

Forelæsning 5: Hypotesetest, power og modelkontrol - one sample

Klaus K. Andersen og Per Bruun Brockhoff

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

e-mail: klaus@cancer.dk

DTU Compute
Department of Applied Mathematics and Computer Science

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	$x =$ Beffect - Aeffect
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Eksempel - sovemedicin

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \ \mu = 0$$

Sample mean og standard deviation:

$$\bar{x} = 1.670 = \hat{\mu}$$
$$s = 1.13 = \hat{\sigma}$$

Er data i overenstemmelse med nulhyposen H_0?

Data: $\bar{x} = 1.67, \ H_0: \ \mu = 0$

NYT:p-**værdi:**

$$p - \text{værdi} = 0.00117$$

(Beregnet under det scenarie, at H_0 er sand)

NYT:**Konklusion**:

Idet data ligger usandsynligt langt væk fra H_0, så **forkaster** vi H_0 - vi har påvist en **signifikant effekt** af middel B ift. middel A.

Metode 3.22: One-sample t-test og p-værdi

Hvordan beregner man p-værdien?

For a (quantitative) one sample situation, the (non-directional) p-value is given by:

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

where T follows a t-distribution with $(n-1)$ degrees of freedom.
The observed value of the test statistics to be computed is

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where μ_0 is the value of μ under the null hypothesis:

$$H_0 : \ \mu = \mu_0$$

Definition og fortolkning af p-værdien (HELT generelt)

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 3.12 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.

Eksempel - sovemedicin

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0 : \ \mu = 0$$

Beregne test-størrelsen:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

Beregne p-værdien:

$$2P(T > 4.67) = 0.00117$$

```
2 * (1-pt(4.67, 9))
```

Fortolkning af p-værdi i lyset af Tabel 3.1:

Der er stærk evidence imod nulhypotesen.

Eksempel - sovemedicin - i R - manuelt

```
## Enter data:
x <- c(1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4)
n <- length(x)
## Compute the tobs - the observed test statistic:
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))
## Compute the p-value as a tail-probability in the t-distribution:
pvalue <- 2 * (1-pt(abs(tobs), df=n-1))
pvalue

## [1] 0.0011659
```

Eksempel - sovemedicin - i R - med indbygget funktion

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## 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
```


Definition af hypotesetest og signifikans (HELT generelt)

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 $\alpha = 0.05$)

Eksempel - sovemedicin

Med $\alpha = 0.05$ kan vi konkludere:

Idet p-værdien er mindre end α så **forkaster** vi nulhypotesen.

Og dermed:

Vi har påvist en **signifikant effekt** af middel B ift. middel A. (Og dermed at B virker bedre end A)

Kritisk værdi

Definition 3.30 - de kritiske værdier for t-testet:

The $(1-\alpha)100\%$ critical values for the (non-directional) one-sample t-test are the $(\alpha/2)100\%$ and $(1-\alpha/2)100\%$ quantiles of the t-distribution with $n-1$ degrees of freedom:

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}$$

Metode 3.31: One-sample t-test vha. kritisk værdi:

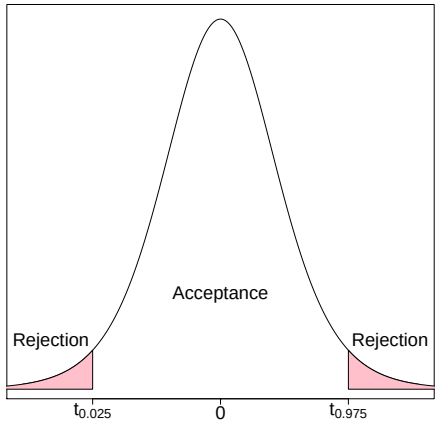
A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values:

$$\text{If } |t_{\text{obs}}| > t_{1-\alpha/2} \text{ then } \textit{reject}$$

otherwise *accept*.

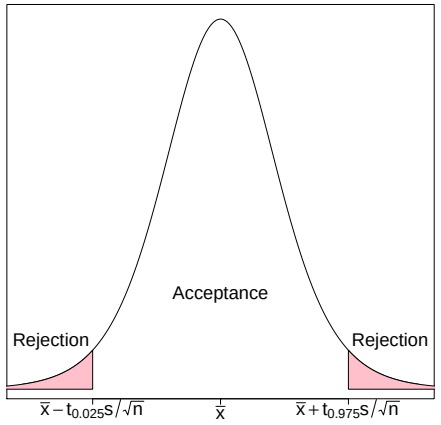
Kritisk værdi og hypotesetest

Acceptområdet er de mulige værdier for μ som ikke ligger for langt væk fra data - her på den standardiserede skala:



Kritisk værdi og hypotesetest

Acceptområdet er de mulige værdier for μ som ikke ligger for langt væk fra data - nu på den egentlige skala:



Kritisk værdi, konfidensinterval og hypotesetest

Theorem 3.32: Kritisk-værdi-metode = Konfidensinterval-metode

We consider a $(1-\alpha)\cdot 100\%$ confidence interval for μ:

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

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0: \quad \mu = \mu_0$$

(Ny) fortolkning af konfidensintervallet:

De (hypotetiske) værdier for μ, som vi accepterer ved det tilsvarende hypotesetest.

Bevis:

Remark 3.33

A μ_0 inside the confidence interval will fullfill that

$$|\bar{x} - \mu_0| < t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

which is equivalent to

$$\frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} < t_{1-\alpha/2}$$

and again to

$$|t_{\text{obs}}| < t_{1-\alpha/2}$$

which then exactly states that μ_0 is accepted, since the t_{obs} is within the critical values.

Hypotese-test med alternativer

Indtil nu - underforstået: (= non-directional)

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu \neq \mu_0$

MEN der kan være andre settings, e.g. one-sided (=directional), "less":

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu < \mu_0$

Eller one-sided (=directional), "greater":

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu > \mu_0$

Eksempel - PC skærme

Produktspecifikation

En producent af pc skærme oplyser, at skærmen i gennemsnit bruger 83 W. (Og underforstået: "under 83" er "fint nok", mens "over 83" IKKE er det)

HVIS virksomheden skulle dokumentere deres påstand:

Nulhypotese: $H_0: \ \mu \geq 83$. Alternativet: $H_1: \ \mu < 83$
Med formålet at kunne afvise(=rejecte=falsify) at forbruget kan være større.

HVIS en ekstern skulle moddokumentere påstanden:

Nulhypotese: $H_0: \ \mu \leq 83$. Alternativet: $H_1: \ \mu > 83$
Med formålet at kunne afvise(=rejecte=falsify) at forbruget højst er 83.

Metode 3.36. Steps ved hypotesetests - et overblik

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))

Det tosidede (non-directional) one-sample t-test igen

Metode 3.37. Et level α test er:

1. Compute t_{obs} as before
2. Compute the evidence against the *null hypothesis* $H_0: \ \mu = \mu_0$ vs. the *alternative hypothesis* $H_1: \ \mu \neq \mu_0$ by the

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

   where the t-distribution with $n-1$ 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.

Det ensidede (directional) one-sample t-test

Metode 3.38. Et level α ensidet ("less") test er:

1. Compute t_{obs} as before
2. Compute the evidence against the *null hypothesis* $H_0: \ \mu \geq \mu_0$ vs. the *alternative hypothesis* $H_1: \ \mu < \mu_0$ by the

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

   where the t-distribution with $n-1$ 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) t_α:
   If $t_{\text{obs}} < t_\alpha$ we reject H_0, otherwise we accept H_0.

DTU Compute
Department of Applied Mathematics and Computer Science

Det ensidede (directional) one-sample t-test

Metode 3.39. Et level α ensidet ("greater") test er:

1. Compute t_{obs} as before
2. Compute the evidence against the *null hypothesis* $H_0: \ \mu \leq \mu_0$ vs. the *alternative hypothesis* $H_1: \ \mu > \mu_0$ by the

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

   where the t-distribution with $n-1$ 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) $t_{1-\alpha}$:
   If $t_{\text{obs}} > t_{1-\alpha}$ we reject H_0, otherwise we accept H_0.

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - PC skærme

Kan man modbevise producentens påstand?

En forbrugergruppe vil nu afprøve producentens påstand og udfører et antal målinger af strømforbruget for den pågældende type pc skærm:
Der udføres nu 12 målinger af forbruget:
82 86 84 84 92 83 93 80 83 84 85 86
Herfra estimeres middelforbruget til $\bar{x} = 85.17$ og $s = 3.8099$

Så, one-sided "greater" er det relevante test:

Nulhypotese: $H_0: \ \mu \leq 83$. Alternativet: $H_1: \ \mu > 83$

Beregn test-størrelse og P-værdi:

$$t_{\text{obs}} = \frac{85.17 - 83}{3.8099/\sqrt{12}} = 1.97$$

$$p\text{-value} = P(T > 1.97) = 0.0373$$

Eksempel - PC skærme

Konklusion ved brug af $\alpha = 0.05$:

Vi forkaster nulhypotesen: Vi har påvist at skærmene's middelforbrug er signifikant større end 83W.

```
x <- c(82, 86, 84, 84, 92, 83, 93, 80, 83, 84, 85, 86)
t.test(x, mu = 83, alt = "greater")

##
##  One Sample t-test
##
## data:  x
## t = 1.97, df = 11, p-value = 0.037
## alternative hypothesis: true mean is greater than 83
## 95 percent confidence interval:
##  83.192    Inf
## sample estimates:
## mean of x
##    85.167
```

Mulige fejl ved hypotesetests

Der findes to slags fejl (dog kun een af gangen!)

Type I: Rejection of H_0 when H_0 is true
Type II: Non-rejection of H_0 when H_1 is true

Risikoen for de to typer fejl kaldes sædvanligvis:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

DTU Compute
Department of Applied Mathematics and Computer Science

Retsalsanalogi

En person står stillet for en domstol:

A man is standing in a court of law accused of criminal activity.
The null- and the the alternative hypotheses are:

H_0 : The man is not guilty

H_1 : The man is guilty

At man ikke kan bevises skyldig er ikke det samme som at man er bevist uskyldig:

Absence of evidence is NOT evidence of absence!

Or differently put:
Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!

Mulige fejl ved hypotesetests

Theorem 3.43: Signifikansniveauet = Risikoen for Type I fejl

The significance level α in hypothesis testing is the overall Type I risk:

$$P(\text{Type I error}) = P(\text{Rejection of } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

To mulige sandheder vs. to mulige konklusioner:

	Reject H_0	Fail to reject H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0 (Power)	Type II error (β)

DTU Compute
Department of Applied Mathematics and Computer Science

Planlægning, Styrke (=Power)

Hvad er styrken for et kommende studie/eksperiment:

- Sandsynligheden for at opdage en (formodet) effekt
- $P(\text{Forkaste } H_0)$ 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

DTU Compute
Department of Applied Mathematics and Computer Science

Planlægning, Styrke (=Power)

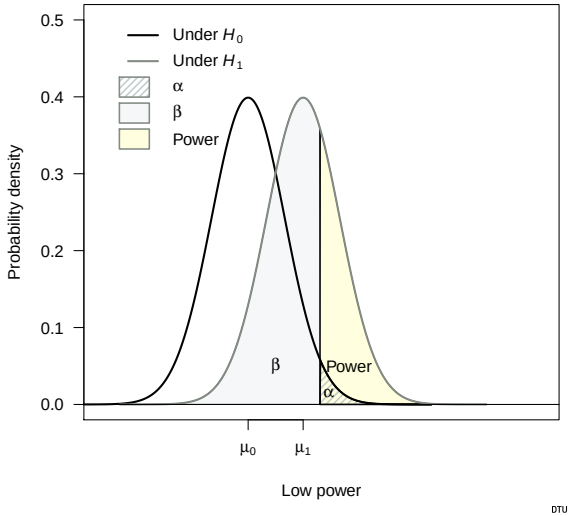
Når man har fastlagt hvilket test, der skal bruges:

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 $(1 - \beta)$.

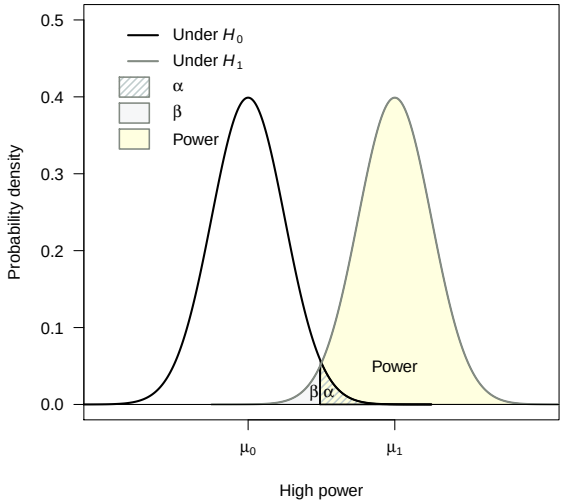
DTU Compute
Department of Applied Mathematics and Computer Science

Low power eksempel



DTU Compute
Department of Applied Mathematics and Computer Science

High power eksempel



DTU Compute
Department of Applied Mathematics and Computer Science

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 3.47: Tilnærmet svar for et en-sidet one-sample t-test:

For the one-sided, one-sample t-test for given α, β and σ:

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

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

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - The power for $n = 40$

```
power.t.test(n = 40, delta = 4, sd = 12.21,
      type = "one.sample", alternative= "one.sided")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.65207
##     alternative = one.sided
```

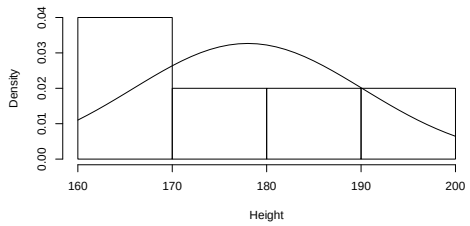

Eksempel - The sample size for power$= 0.80$

```
power.t.test(power = .80, delta = 4, sd = 12.21,
      type = "one.sample", alternative= "one.sided")

##
##      One-sample t test power calculation
##
##               n = 58.984
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
```

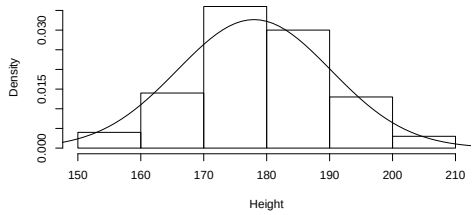

Eksempel - højde af studerende - er de normalfordelt?

```
x <- c(168,161,167,179,184,166,198,187,191,179)
hist(x, xlab="Height", main="", freq = FALSE)
lines(seq(160, 200, 1), dnorm(seq(160, 200, 1), mean(x), sd(x)))
```



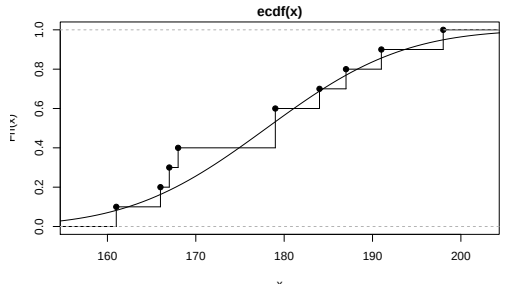
Eksempel - 100 observation fra en normal fordeling:

```
xr <- rnorm(100, mean(x), sd(x))
hist(xr, xlab="Height", main="", freq = FALSE)
lines(seq(130, 230, 1), dnorm(seq(130, 230, 1), mean(x), sd(x)))
```



Eksempel - højde af studerende - ecdf

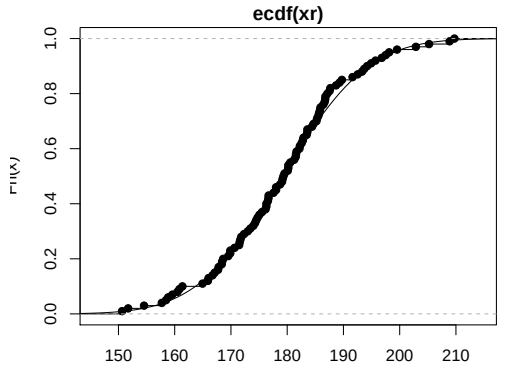
```
plot(ecdf(x), verticals = TRUE)
xp <- seq(0.9*min(x), 1.1*max(x), length.out = 100)
lines(xp, pnorm(xp, mean(x), sd(x)))
```



DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - 100 observation fra en normal fordeling, ecdf:

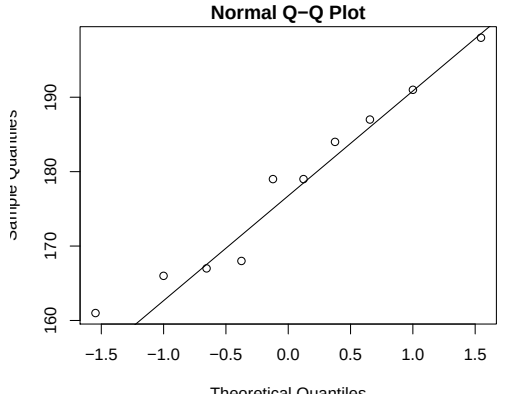
```
xr <- rnorm(100,  mean(x), sd(x))
plot(ecdf(xr), verticals = TRUE)
xp <- seq(0.9*min(xr), 1.1*max(xr), length.out = 100)
lines(xp, pnorm(xp, mean(xr), sd(xr)))
```



DTU Compute
Department of Applied Mathematics and Computer Science

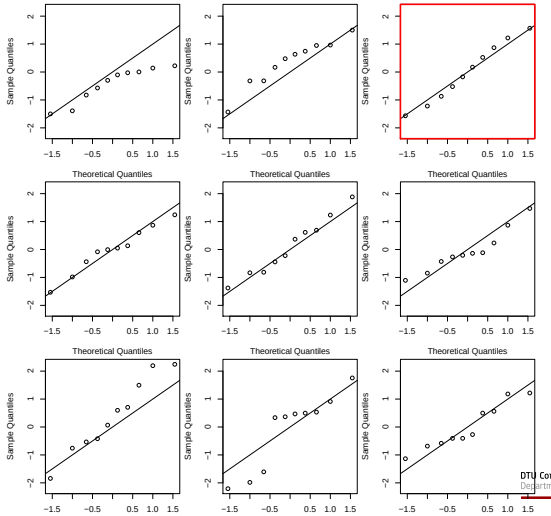
Eksempel - højde af studerende - Normal Q-Q plot

```
qqnorm(x)
qqline(x)
```



DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel - højde af studerende - Normal Q-Q plot - sammenlign med andre simulerede normalfordelte data



DTU Compute
Department of Applied Mathematics and Computer Science

Normal Q-Q plot

Metode 3.52 - 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 $n > 10$:

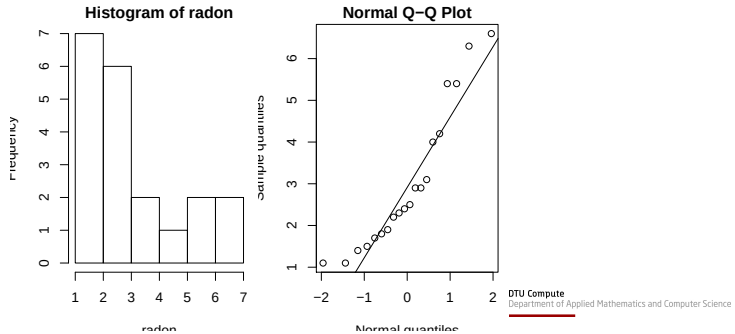
$$p_i = \frac{i - 0.5}{n + 1}, \; i = 1, \ldots, n$$

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

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

Eksempel - Radon data

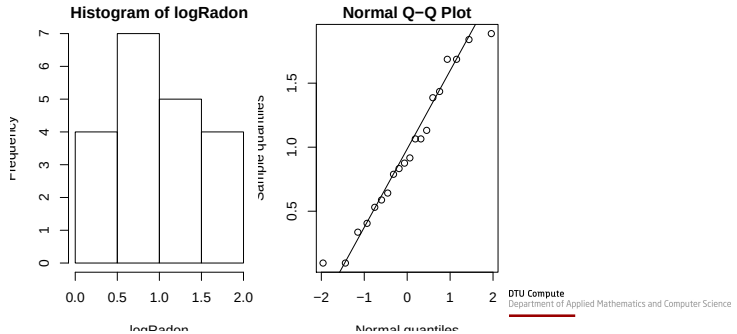
```
## READING IN THE DATA
radon<-c(2.4, 4.2, 1.8, 2.5, 5.4, 2.2, 4.0, 1.1, 1.5, 5.4, 6.3,
         1.9, 1.7, 1.1, 6.6, 3.1, 2.3, 1.4, 2.9, 2.9)
##A HISTOGRAM AND A QQ-PLOT
par(mfrow=c(1,2))
hist(radon)
qqnorm(radon,ylab = 'Sample quantiles',xlab = "Normal quantiles")
qqline(radon)
```



Eksempel - Radon data - log-transformed are closer to a normal distribution

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

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



Oversigt