Anchored Links

(!) = skal testes om den er true

The p-value expresses *evidence* against the null hypothesis – Table 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

! = mangler eksempel

```
x <- "3.7 1.9 4.8 11.7 2.8 4.7 2.7 4.9 6.7 3.8"
lifetime <- as.numeric(strsplit(x, " ")[[1]])
```

1 sample data

CI

Method 3.9 The one sample confidence interval for μ

For a sample $x_1, \ldots, x_n$ the $100(1-\alpha)\%$ confidence interval is given by

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \tag{3-10}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile from the t-distribution with $n-1$ degrees of freedom.[a]
Most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}. \tag{3-11}$$

```
n <- 22
mu <- 0
mean <- -0.1181818
```

```r
sd <- 0.05884899

qt_scale <- qt(0.975, n-1) * sd/sqrt(n)
mean + c(-1,1)*qt_scale
```

```
## [1] -0.14427398 -0.09208962
```

pval

```r
n <- 14
mu <- 0
mean <- 367.2
sd <- 571.5
tobs <- (mean-mu)/(sd/sqrt(n))
2*(1-pt(tobs, n-1))
```

```
## [1] 0.03184036
```

CI for var/sd

Variansen:

Et $100(1-\alpha)\%$ konfidensinterval for stikprøvevariansen: $\hat{\sigma}^2$ er:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$

hvor fraktilerne kommer fra en χ^2-fordeling med $\nu = n-1$ frihedsgrader.

Standardafvigelsen:

Et $100(1-\alpha)\%$ konfidensinterval for stikprøvestandardafvigelsen $\hat{\sigma}$ er:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}\right]$$

```r
n <- 22
mu <- 0
mean <- -0.1181818
s <- 0.05884899
```

```r
chiout <- qchisq(0.975, n-1)
chiin <- qchisq(0.025, n-1)
#var
var_ci <- (n-1)*(s^2) * c(chiout^-1,chiin^-1)

#sd
sd_ci <- sqrt(var_ci)
sd_ci
```

```
## [1] 0.04527555 0.08409901
```

Which sample size to choose?

ME - simple men med shortcomings

Method 3.63 The one-sample CI sample size formula

When σ is known or guessed at some value, we can calculate the sample size n needed to achieve a given margin of error, ME, with probability $1 - \alpha$ as

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2. \tag{3-59}$$

```r
ME <- 3
x.sd <- 12.21
# norm dist
z.quantile <- qnorm(0.975)
(n <- ((z.quantile*x.sd)/ME)^2 )
```

```
## [1] 63.63338
```

```r
"so n = 64 if int"
```

```
## [1] "so n = 64 if int"
```

Power - advanced men better

Method 3.65 The one-sample sample size formula

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

Manuelt

```
z.quantb <- qnorm(0.80)
z.quanta <- qnorm(0.975)
sd <- 12.21
diff <- 4

(n = (sd* (z.quantb + z.quanta)/diff) ^2 )
```

```
## [1] 73.13395
```

med r direkte

```
#power: 80% chance for at accep den hvis den er true
# a = sig.level
# delta, den diffrence i mean vi vil kunne se forskell på. Relevant fordi vi kigger på difference
# Den her er mere accurate da den bruge t-dist
power.t.test(power=0.8, delta=4, sd=12.21, sig.level=0.05, type="one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 75.07733
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

two sample data

CI for difference

Method 3.47 The two-sample confidence interval for $\mu_1 - \mu_2$

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

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{3-45}$$

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

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}. \tag{3-46}$$

```
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)

A.mean <-  mean(xA)
B.mean <-  mean(xB)
A.len <- length(xA)
B.len <- length(xB)
A.sd <- sd(xA)
B.sd <- sd(xB)

# df
vs=c(var(xA), var(xB))
ns=c(length(xA), length(xB))

v <- ((vs[1]/ns[1]+vs[2]/ns[2])^2)/((vs[1]/ns[1])^2/(ns[1]-1)+(vs[2]/ns[2])^2/(ns[2]-1))

t.quantile <- qt(0.975, v)
change.t <- sqrt( ((A.sd)^2/A.len) + ((B.sd)^2/B.len) )
A.mean-B.mean + c(-1,1)*t.quantile*change.t
```

```
## [1] -3.4166085 -0.5922804
```

```
c(A.mean,B.mean,t.quantile,change.t)
```

```
## [1]  8.2933333 10.2977778  2.1199840  0.6661201
```

```r
# Eller direkte
t.test(xA,xB)
```

```
##
##  Welch Two Sample t-test
##
## data:  xA and xB
## t = -3.0091, df = 15.993, p-value = 0.008323
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.4166085 -0.5922804
## sample estimates:
## mean of x mean of y
##  8.293333 10.297778
```

Welch two sample t-test statistic

Method 3.49 The (Welch) two-sample t-test statistic

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

$$\begin{aligned} \delta &= \mu_2 - \mu_1, \\ H_0 &: \ \delta = \delta_0, \end{aligned} \tag{3-47}$$

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

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \tag{3-48}$$

Method 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*[a]

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

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

```
ms=c(mean(xA), mean(xB))
vs=c(var(xA), var(xB))
ns=c(length(xA), length(xB))

# Test statistic
tobs <- (ms[1]-ms[2])/sqrt(vs[1]^2/ns[1]+vs[2]^2/ns[2])

# Degrees of freedom
v=((vs[1]/ns[1]+vs[2]/ns[2])^2)/((vs[1]/ns[1])^2/(ns[1]-1)+(vs[2]/ns[2])^2/(ns[2]-1))
c(tobs,v)
```

```
## [1] -2.129038 15.992694
```

```
#Pval
(pval <- 2*(1-pt(tobs, v)) )
```

```
## [1] 1.95086
```

```
#CritVal
(critval <- qt(0.975, v))
```

```
## [1] 2.119984
```

```
# Direkte
t.test(xA, xB)
```

```
##
##  Welch Two Sample t-test
##
## data:  xA and xB
## t = -3.0091, df = 15.993, p-value = 0.008323
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.4166085 -0.5922804
## sample estimates:
## mean of x mean of y
##  8.293333 10.297778
```

Power/sample in two way

```
# Finding the sample size for detecting a group difference of 2
# with sigma=1 and power=0.9
# Her har vi ikke type = "one.sample" på
power.t.test(power=0.90, delta=2, sd=1, sig.level=0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 6.386756
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
# udfra hvad du giver den, regner den noget forskelligt eg
# 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 = 6.386756
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Pooled varance/sd

The *pooled* estimate of variance (assuming $\sigma_1^2 = \sigma_2^2$)

Method 3.52

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

```r
v1 <- 1.8^2
n1 <- 20
v2 <- 1.4^2
n2 <- 30
var <- ((n1-1)*v1+(n2-1)*v2) / (n1 + n2 -2)
sigma <- sqrt(var)
```

anova test

One-way anova

```r
D <- data.frame(strength=c(44.6, 52.8, 53.1, 51.5, 48.2, 50.5, 58.3, 50.0, 53.7, 40.8,
46.3, 55.4, 54.4, 50.5, 44.5, 48.5, 57.4, 55.3, 54.4, 43.9,
45.2, 58.1, 50.6, 47.5, 45.9, 52.3, 54.6, 53.4, 47.8, 42.5),
plastictype = factor(rep(1:5,6))
)

fit <- lm(strength ~ plastictype, data=D)
an <- anova(fit)
an
```

```
## Analysis of Variance Table
##
## Response: strength
##             Df Sum Sq Mean Sq F value    Pr(>F)
## plastictype  4 491.76 122.940  18.234 3.988e-07 ***
## Residuals   25 168.56   6.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

POST hoc pairwaise CI

Method 8.9 Post hoc pairwise confidence intervals

A single pre-planned $(1-\alpha)\cdot 100\%$ confidence interval for the difference between treatment i and j is found as

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{8-22}$$

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

If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are calculated using the formula M times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

```
rm(list = ls())
a.mean <- -0.5416667
b.mean <- -0.6816667

an <- 10
bn <- 10
n <- 25
k <- 4
df <- n-k

a <- 0.05/1
mse <- 0.16207

mean.dif <- an - bn

t.quant <- qt(1-(a/2), df)
inner <- mse * ((1/an) + (1/an))
scale <- sqrt(inner)

mean.dif + c(-1,1)*t.quant*scale
```

```
## [1] -0.3744114  0.3744114
```

LSD values - samme antal obs

Remark 8.13 Least Significant Difference (LSD) values

If there is the same number of observations in each treatment group $m = n_1 = \ldots = n_k$ the LSD value for a particular significance level

$$LSD_\alpha = t_{1-\alpha/2}\sqrt{2 \cdot MSE/m} \tag{8-28}$$

will have the same value for all the possible comparisons made.
The LSD value is particularly useful as a "measuring stick" with which we can go and compare all the observed means directly: the observed means with difference higher than the LSD are significantly different on the α-level. When used for all of the comparisons, as suggested, one should as level use the Bonferroni corrected version $LSD_{\alpha_{\text{Bonferroni}}}$ (see Remark 8.14 below for an elaborated explanation).

```
m <- 2  # Antal obs i hver kategory

mse <- 0.16207
ny_alpha = 0.005
n <- 10
k <- 2

LSD <- qt(1-(ny_alpha/2), n-k)*sqrt(2*mse*(1/m))
LSD
```

```
## [1] 1.542892
```

Post hoc hypothesis test

Method 8.10 Post hoc pairwise hypothesis tests

A single pre-planned level α hypothesis tests

$$H_0 : \mu_i = \mu_j, \quad H_1 : \mu_i \neq \mu_j, \tag{8-23}$$

is carried out by

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{8-24}$$

and

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|), \tag{8-25}$$

where the t-distribution with $n - k$ degrees of freedom is used.

If all $M = k(k-1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times but each time with test level $\alpha_{\text{Bonferroni}} = \alpha/M$ (see Remark 8.14 below).

```
rm(list = ls())
a.mean <- -0.5416667
b.mean <- -0.6816667

an <- 10
bn <- 10
n <- 25
k <- 4
df <- n-k

a <- 0.05/1
mse <- 0.16207
mean.dif <- a.mean - b.mean

t.obs <- mean.dif/ sqrt(mse*((1/an)+ (1/bn)))
t.obs
```

```
## [1] 0.7776098
```

```
pval <- 2* (1- pt(t.obs, df))
```

Two way ANOVA

One-way ANOVA

Source of variation	Degrees of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

Two-way ANOVA

Source of variation	Degrees of freedom	Sums of squares	Mean sums of squares	Test statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$	$P(F > F_{Tr})$
Block	$l-1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$	$P(F > F_{Bl})$
Residual	$(l-1)(k-1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	$n-1$	SST			

2. Use $(l-1)(k-1)$ instead of $n-k$ as degrees of freedom and as denominator for SSE

Theorem 8.22

Under the null hypothesis

$$H_{0,Tr}: \quad \alpha_i = 0, \quad i = 1, 2, \ldots, k, \tag{8-44}$$

the test statistic

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}, \tag{8-45}$$

follows an F-distribution with $k-1$ and $(k-1)(l-1)$ degrees of freedom. Further, under the null hypothesis

$$H_{0,Bl}: \quad \beta_j = 0, \quad j = 1, 2, \ldots, l, \tag{8-46}$$

the test statistic

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}, \tag{8-47}$$

follows an F-distribution with $l-1$ and $(k-1)(l-1)$ degrees of freedom.

testen laves som pval = 1-pf(fobs, l-1, (k-1)(l-1)

Anova test

```
y <- c(3.5, 3.0, 5.4, 7.2,
       7.7, 9.0, 7.0, 6.0,
       0.4, 1.1, 1.0, 1.8)
treatm <- as.factor(c(1, 1, 1, 1,
                      2, 2, 2, 2,
                      3, 3, 3, 3))
block <- as.factor(c(1, 2, 3, 4,
                     1, 2, 3, 4,
                     1, 2, 3, 4))

fit <- lm(y ~ treatm + block)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## treatm     2 81.380  40.690 16.4293 0.003681 **
```

```
## block      3  1.943   0.648  0.2614 0.850900
## Residuals  6 14.860   2.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Find sd

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
treatm     2 81.380  40.690 16.4293 0.003681 **
block      3  1.943   0.648  0.2614 0.850900
Residuals  6 14.860   2.477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

estimation of parameters

$$\hat{\mu} = \bar{\bar{y}},$$
$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{\bar{y}},$$
$$\hat{\beta}_j = \bar{y}_{\cdot j} - \bar{\bar{y}}.$$

```
y <- c(3.5, 3.0, 5.4, 7.2,
       7.7, 9.0, 7.0, 6.0,
       0.4, 1.1, 1.0, 1.8)
treatm <- as.factor(c(1, 1, 1, 1,
                      2, 2, 2, 2,
                      3, 3, 3, 3))
block <- as.factor(c(1, 2, 3, 4,
                     1, 2, 3, 4,
                     1, 2, 3, 4))
tapply(y, block, mean)
```

```
##        1        2        3        4
## 3.866667 4.366667 4.466667 5.000000
```

```
3.866667-mean(y)
```

```
## [1] -0.558333
```

PROPORTION analysis

1 population proportion analysis

CI

Method 7.3 Proportion estimate and confidence interval

The best estimate of the probability p of belonging to a category (the population proportion) is the sample proportion

$$\hat{p} = \frac{x}{n}, \tag{7-8}$$

where x is the number of observations in the category and n is the total number of observations.

A large sample $(1 - \alpha)100\%$ confidence interval for p is given as

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7-9}$$

small size er når:

when either $np \leq 15$ or $n(1-p) \leq 15$.

```
x <- 518
n <- 1154
p.hat <- x/n
z.quantile <- qnorm(0.975)
scale <- sqrt(((p.hat*(1-p.hat))/n))
p.hat + c(-1,1) * z.quantile * scale
```

```
## [1] 0.4201767 0.4775702
```

```
scale
```

```
## [1] 0.01464147
```

```
prop.test(518,1154,correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  518 out of 1154, null probability 0.5
## X-squared = 12.066, df = 1, p-value = 0.0005135
```

```
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4203935 0.4776927
## sample estimates:
##         p
## 0.4488735
```

hypothesis test

Method 7.11 One sample proportion hypothesis test

1. Compute the test statistic using Equation (7-16)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p = p_0, \tag{7-19}$$

vs. the *the alternative hypothesis*

$$H_1 : p \neq p_0, \tag{7-20}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-21}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the 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) $\pm z_{1-\alpha/2}$:
if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

```
x <- 518
n <- 1154
p.hat <- x/n
# Test statistic
```

```r
# is 0.5 the true proportion?
p0 <- 0.5
zobs <- (x - n*p0) / sqrt(n*p0*(1-p0))

# Husk abs a zobs
p.val <- 2 * (1-pnorm(abs(zobs)))
p.val
```

```
## [1] 0.0005135367
```

```r
prop.test(518,1154,correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  518 out of 1154, null probability 0.5
## X-squared = 12.066, df = 1, p-value = 0.0005135
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4203935 0.4776927
## sample estimates:
##         p
## 0.4488735
```

ME !

Method 7.13 Sample size formula for the CI of a proportion

Given some "guess" (scenario) of the size of the unknown p, and given some requirement to the ME-value (required expected precision) the necessary sample size is then

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-24}$$

If p is unknown, a worst case scenario with $p = 1/2$ is applied and necessary sample size is

$$n = \frac{1}{4}\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-25}$$

```r
p <- 0.04
ME <- 0.01
(n=p*(1-p)*(qnorm(0.975)/ME)^2)
```

```
## [1] 1475.12
```

Two proportions

CI og sd/var !

Method 7.15

An estimate of the standard error of the estimator $\hat{p}_1 - \hat{p}_2$ is

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \tag{7-29}$$

The $(1-\alpha)100\%$ confidence interval for the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1-\hat{p}_2}. \tag{7-30}$$

This confidence interval requires independent random samples for the two groups and large enough sample sizes n_1 and n_2. A rule of thumb is that $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$, must be satisfied.

Remark 7.16

The standard error in Method 7.15 can be calculated by

$$\mathrm{V}(\hat{p}_1 - \hat{p}_2) = \mathrm{V}(\hat{p}_1) + \mathrm{V}(\hat{p}_2) = \hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}, \tag{7-31}$$

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\mathrm{V}(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}}. \tag{7-32}$$

Notice, that the standard errors are added (before the square root) such that the standard error of the difference is larger than the standard error for the observed proportions alone. Therefore in practice the estimate of the difference $\hat{p}_1 - \hat{p}_2$ will often be further from the true difference $p_1 - p_2$ than $\hat{p}_1$ will be from p_1 or $\hat{p}_2$ will be from p_2.

```
p1 <- 26/189
p2 <- 11/157
n1 <- 189
n2 <- 157
sigma_p1_p2 <- sqrt( (p1*(1-p1)/n1) + (p2*(1-p2)/n2) )
```

```r
p1-p2 + c(-1,1)*sigma_p1_p2*qnorm(0.975)
```

```
## [1] 0.004212527 0.130792360
```

Hypothesis test !

Method 7.18 Two sample proportions hypothesis test

The two-sample hypothesis test for comparing two proportions is given by the following procedure:

1. Compute, with $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, the test statistic

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{7-37}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p_1 = p_2, \tag{7-38}$$

vs. the *the alternative hypothesis*

$$H_1 : p_1 \neq p_2, \tag{7-39}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-40}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the 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) $\pm z_{1-\alpha/2}$:
   if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

```r
x1 <- 31+36
x2 <- 31+30
n1 <- 189
n2 <- 175
p <- (x1+x2)/(n1+n2)
p1 <- x1/n1
p2 <- x2/n2

zobs <- (p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))
```

```r
prop.test(x=c(23,35), n=c(57,167), correct=FALSE, conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(23, 35) out of c(57, 167)
## X-squared = 8.3288, df = 1, p-value = 0.003902
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.007922055 0.379933812
## sample estimates:
##    prop 1    prop 2
## 0.4035088 0.2095808
```

Method 7.20 The multi-sample proportions χ^2-test

The hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_c = p,$$

can be tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expecte
in cell (i, j).

The test statistic χ^2_{obs} should be compared with the χ^2-distribution
degrees of freedom.

The χ^2-distribution is approximately the sampling distribution of
tics under the null hypothesis. The rule of thumb is that it is vali
the computed expected values are at least 5: $e_{ij} \geq 5$.

Multi sample ! på en axis

	Birth control pill	No birth control pill	*Total*
Blood clot	$o_{11} = 23$ $e_{11} = 14.76$	$o_{12} = 35$ $e_{12} = 43.24$	$x = 58$
No blood clot	$o_{21} = 34$ $e_{21} = 42.24$	$o_{22} = 132$ $o_{22} = 123.8$	$(n - x) = 166$
Total	$n_1 = 57$	$n_2 = 167$	$n = 224$

The observed χ^2 test statistic can be calculated

$$\chi^2_{\text{obs}} = \frac{(23 - 14.76)^2}{14.76} + \frac{(35 - 43.24)^2}{43.24} + \frac{(34 - 42.24)^2}{42.24} + \frac{(132 - 123.8)^2}{123.8} = 8.33. \tag{7-49}$$

We then find the p-value, by calculating how likely it is to get 8.33 or more extreme if the null hypothesis is true, using the χ^2 distribution with $c - 1 = 2 - 1 = 1$ degrees of freedom

$$p\text{-value} = P(\chi^2 \geq 8.33) = 0.0039, \tag{7-50}$$

```
pill.study <- matrix(c(23, 35, 34, 132), ncol = 2, byrow = TRUE)
rownames(pill.study) <- c("Blood Clot", "No Clot")
colnames(pill.study) <- c("Pill", "No pill")
chi <- chisq.test(pill.study, correct = FALSE)
#X-squared er test statistic (?)
chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  pill.study
## X-squared = 8.3288, df = 1, p-value = 0.003902
```

```
chi$expected
```

```
##                Pill   No pill
## Blood Clot 14.75893  43.24107
## No Clot    42.24107 123.75893
```

Multisample på begge

Method 7.22 The $r \times c$ frequency table χ^2-test

For an $r \times c$ table the hypothesis

$$H_0 : p_{i1} = p_{i2} = \ldots = p_{ic} = p_i, \text{ for all rows } i = 1, 2, \ldots, r, \tag{7-54}$$

is tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}. \tag{7-55}$$

where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expected number in cell (i, j). This test statistic should be compared with the χ^2-distribution with $(r-1)(c-1)$ degrees of freedom and the hypothesis is rejected at significance level α if

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha}\big((r-1)(c-1)\big). \tag{7-56}$$

	4 weeks before	2 weeks before	1 week before	*Row total*
Candidate 1	79	91	93	263
Candidate 2	84	66	60	210
Undecided	37	43	47	127
Column total	200	200	200	600

$$e_{22} = \text{"2'nd column total"} \cdot \frac{\text{"2'nd row total"}}{\text{"grand total"}} = \frac{210 \cdot 200}{600} = 70.$$

```
mat <- matrix(c(24, 21, 14,
                12, 15, 22,
                15, 26, 24), ncol = 3, byrow = TRUE)
colnames(mat) <- c("Low","Medium","High")
rownames(mat) <- c("A","B","C")

mat <- as.data.frame(mat)

test <- chisq.test(mat, correct = FALSE)
test$expected
```

```
##        Low   Medium     High
```

```
## A 17.39306 21.14451 20.46243
## B 14.44509 17.56069 16.99422
## C 19.16185 23.29480 22.54335
```

Her er contribution af en

```
e <- 344*189/662
o <- 96
(e-o)^2 / e
```

```
## [1] 0.04979708
```

Simulation methods

Bootstrapping

normal LR

CI for parametre

Method 5.15 Parameter confidence intervals

$(1-\alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_0}, \tag{5-52}$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \hat{\sigma}_{\beta_1}, \tag{5-53}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-2$ degrees of freedom. Where $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ are calculated from the results in Theorem 5.8, and Equations (5-43) and (5-44).

```
B1 <- 23.25
tquant <- qt(0.975, 18)
B1.sigma <- 1.74
B1 + c(-1,1)*tquant*B1.sigma
```

```
## [1] 19.5944 26.9056
```

Method 5.18 Intervals for the line

The (1-α) **confidence interval** for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-59}$$

and the (1-α) **prediction interval** is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}, \tag{5-60}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-2$ degrees of freedom.

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
b0 <- -120
b1 <- 1.113
xnew <- 200

mean <- 178
n <- 10
df <- n-2
sd <- 3.88
sxx <- 1342

"CI:"
```

```
## [1] "CI:"
```

```
tquantile <- qt(0.975, df)
scale <- sd * sqrt(1/n + (mean - xnew)^2/sxx )
b0+b1*xnew + c(-1, 1)*tquantile*scale
```

```
## [1]  96.52732 108.67268
```

```
"Prediction Interval:"
```

```
## [1] "Prediction Interval:"
```

```
tquantile <- qt(0.975, df)
scale <- sd * sqrt(1 + 1/n + (mean - xnew)^2/sxx )
b0+b1*xnew + c(-1, 1)*tquantile*scale
```

```
## [1]  91.78651 113.41349
```

```
# Direkte med r
y <- c(8.43, 7.89, 8.28, 7.84, 9.62, 9.41, 9.40, 8.22, 9.18, 9.17,
9.25, 9.68, 8.49, 8.53, 9.30, 8.94, 9.46, 9.69, 9.37, 9.42,
9.13, 9.18)
x <- year <- 1984:2005
fit <- lm(y ~ x)

newdata <- data.frame(x = 2017)
predict(fit, newdata=newdata, interval="confidence",
level=0.95)
```

```
##        fit      lwr      upr
## 1 10.03201 9.206954 10.85707
```

```
predict(fit, newdata=newdata, interval="prediction",
level=0.95)
```

```
##        fit      lwr      upr
## 1 10.03201 8.696461 11.36756
```

	Description	Formula	R command
5.4	Least square estimators	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$	
5.8	Variance of estimators	$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}$ $V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$ $Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}$	
5.12	Tests statistics for $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$	$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}$ $T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$	

5.14	Level α t-tests for parameter	Test $H_{0,i} : \beta_i = \beta_{0,i}$ vs. $H_{1,i} : \beta_i \neq \beta_{0,i}$ with p-value $= 2 \cdot P(T > \lvert t_{\text{obs},\beta_i} \rvert)$
		where $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$.
		If p-value $< \alpha$ then *reject* H_0,
		otherwise *accept* H_0
			`D <- data.frame(`
			`  x=c(), y=c())`
			`fit <- lm(y~x, data=D)`
			`summary(fit)`

5.15	Parameter confidence intervals	$\hat{\beta}_0 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_0}$ $\hat{\beta}_1 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_1}$	`confint(fit,level=0.95)`

5.18	Confident and prediction interval	Confidence interval for the line: $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$ Interval for a new point prediction: $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}$	`predict(fit, newdata=data.frame(), interval="confidence", level=0.95)` `predict(fit, newdata=data.frame(), interval="prediction", level=0.95)`
5.23	The matrix formulation of the parameter estimators in the simple linear regression model	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ $V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ $\hat{\sigma}^2 = \frac{RSS}{n-2}$	
5.25	Coefficient of determination R^2	$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$	

	Description	**Formula**	**R command**

5.7	Model validation of assumptions	> Check the normality assumption with a q-q plot of the residuals.
		> Check the systematic behavior by plotting the residuals e_i as a function of fitted values $\hat{y}_i$
			`qqnorm(fit$residuals)`
			`qqline(fit$residuals)`
			
			`plot(fit$fitted.values, fit$residuals)`

MLR

	Description	Formula	R command

6.2	Level α t-tests for parameter	Test $H_{0,i} : \beta_i = \beta_{0,i}$ vs. $H_{1,i} : \beta_i \neq \beta_{0,i}$ with p-value $= 2 \cdot P(T > \lvert t_{\text{obs},\beta_i} \rvert)$ where $t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$. If p-value $< \alpha$ the *reject* H_0, otherwise *accept* H_0	`D<-data.frame(x1=c(), x2=c(),y=c())`
			`fit <- lm(y~x1+x2, data=D)`
			`summary(fit)`

6.5	Parameter confidence intervals	$\hat{\beta}_i \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_i}$	`confint(fit,level=0.95)`
6.9	Confident and prediction interval (in R)	Confident interval for the line $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_p x_{p,\text{new}}$ Interval for a new point prediction $\hat{\beta}_0 + \hat{\beta}_1 x_{1,\text{new}} + \cdots + \hat{\beta}_p x_{p,\text{new}} + \varepsilon_{\text{new}}$	`predict(fit, newdata=data.frame(), interval="confidence", level=0.95)` `predict(fit, newdata=data.frame(), interval="prediction", level=0.95)`
6.17	The matrix formulation of the parameter estimators in the multiple linear regression model	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ $V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$	
6.16	Model selection procedure	Backward selection: start with full model and stepwise remove insignificant terms	