Statistik - Formler, lidt tekst og R-bidder

Lasse Herskind - S153746

Sidst opdateret: **14. maj 2017**

Indhold

1 Da fucking basics of R

1.1 Decleration af variabler

"c"benyttes når der er tale om en array

```
x <- 1
x <- c(1,2,3)
```

1.2 Loop

Her er et eksempel på et loop som summere listen x.

```
x <- c(1,2,3,4,5,6,7,8,9,10)
sumx <- 0
for(i in 1:10){sumx <- sumx + x[i]}
```

1.3 Median

Median er det samme som kvartil til 0.5
0.5n er et helttal

$$\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

0.5n er ikke et heltal

$$x\left(\frac{n}{2}\right)$$

Eksempel med median af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
median(x)
```

1.4 Gennemsnit

Eksempel med gennemsnit af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
mean(x)
```

HUSK:
Hvis vi gør dette over flere målinger af samme fordeling vil vi kunne gange med antallet.

1.5 Varience, varians

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Eksempel med variansen af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
var(x)
```

HUSK:
Hvis vi gør dette over flere målinger af samme fordeling vil vi kunne gange med antallet.

1.6 Standard deviation, standardafvigelse

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

Eksempel med standartafvigelsen af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
sd(x)
```

HUSK:
Hvis vi gør dette over flere målinger af samme fordeling vil vi kunne gange **Variansen** med antallet.

1.7 Covariance

Varians for 2-variable

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Eksempel med kovariansen af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
cov(x)
```

1.8 Correlation

Husk: at spredning er kvadratrod af variansen!

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x * s_y}$$

Eksempel med variansen af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
cor(x)
```

1.9 Quantile , kvartiler

Kvartile er inddelt alt efter en ordnet række. Altså vil 0,25 kvartil være det tal der er ved loftet i $n \cdot 0,25$ i rækken.
pn er et helttal

$$\frac{x_{np} + x_{np+1}}{2}$$

pn er ikke et heltal

$$x_{np}$$

Eksempel med kvartiler af talrækken x

```
x <- c(1,2,3,4,5,6,7,8,9,10)
quantile(x,type=2)
```

Tabel 1 Quantilefunktion på x

0 %	25%	50%	75%	100%
1.0	3.0	5.5	8.0	10.0

Det skal her huskes at 50% kravtil er median. 0% er minimum og 100% er maksimum.

2 Diskrete fordelinger

Til R gælder der nogle fælles egenskaber for fordelingerne nedenfor.

R	Betegnelse
binom	Binomial
hyper	Hypergeometrisk
pois	Poisson

d Tæthedsfunktion $f(x)$, probability densigty function
p Fordelingsfunktion $F(x)$, cumulative distribution function
r Tilfældige tal fra den anførte fordeling
q Fraktil, quantile, i fordeling

Binomialkoefficienten

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

2.1 Middelværdi

$$\mu = E(x) = \sum_{\text{alle x}} x \cdot f(x)$$

Eksempel med terning

$$\begin{aligned} \mu = E(X) &= 1 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} \\ &= 3.5 \end{aligned}$$

2.2 Varians

$$\sigma^2 = Var(X) = \sum_{\text{alle x}} (x-\mu)^2 \cdot f(x)$$

Eksempel med terning

$$\begin{aligned} \sigma^2 &= (1-3.5)^2 \cdot \frac{1}{6} + \cdot + (6-3.5)^2 \cdot \frac{1}{6} \\ &= 2.92 \end{aligned}$$

2.3 Binomial fordeling

Denne bruges primært ved tilfældigheder hvor det samme udfald kan ske flere gange, eksempelvis terningeslag. Tæthedsfunktion:

$$f(x;n,p) = P(X=x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

$x =$ antal ønsket succeser
$n =$ antal gentagelser
$p =$ sandsynlighed for succes

$$\mu = n \cdot p$$
$$\sigma^2 = n \cdot p \cdot (1-p)$$

dbinom(x,n,p)

Eksempel:
To brødre spiller skal, de er lige gode, hvad er chancen for at en af dem vinder tre spil i træk.

dbinom(3,3,0.5)+**dbinom**(0,3,0.5)

2.4 Hypergeometrisk fordeling

Denne bruges ved tilfældigheder hvor det samme udfald ikke kan ske flere gange, eksempelvis kvalitetskontrol. Eller som i eksamenssætet hvor vi skal undersøge om mandlen er i en af de 6 tabte skåle.

$$f(x;n,a,N) = P(X = x) = \frac{\begin{pmatrix} a \\ x \end{pmatrix}\begin{pmatrix} N-a \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

$n =$ antal trækninger $\quad a =$ antal succeser $\quad N =$ samlet antal muligheder

$$\mu = n \cdot \frac{a}{N}$$

$$\sigma^2 = \frac{n \cdot a \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$$

x = antal sucesser, m = antal mulige sucesser, n = resten, k = antal udtrukket
Her er eksemplet med chancen for at ingen af de 6 tabte skåle indeholder mandlen.

```
dhyper(x=0,m=2,n=23,k=6)
```

2.5 Poisson fordeling

Denne bruges når du ikke lige kan gennemskue en øvre grænse.

$$f(x) = P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

$\lambda =$ intensiteten, typisk tidsenhed

$x =$ antal successer

$$\mu = \sigma^2 = \lambda$$

```
dpois(x,lambda)
```

Husk!:
Hvis du har en intensitet på 5/min så kan du blot gange med to for at få intensiteten på 2 minutter, altså 10/2min.

2.5.1 Eksempel

Hvis vi nu som i eksamensættet skal finde sandsynligheden for mellem 160 og 175 på en dag med en frekvens på 62540.8 om året kan det regnes således

```
lambda <- 62540.8/365
ppois(175,lambda)-ppois(159,lambda
```

Vi skal her bruge 159 da vi skal have 160 med i intervallet. Havde vi benyttet 160 ville vi få sandsynligheden for mellem 161 og 175.

3 Kontinuerte fordelinger

Til R gælder der nogle fælles egenskaber for fordelingerne nedenfor.

R	Betegnelse
unif	Uniform fordeling
norm	Normalfordeling
lnorm	Log-normalfordeling
exp	Exponentialfordelingen

d Tæthedsfunktion $f(x)$, probability densigty function
p Fordelingsfunktion $F(x)$, cumulative distribution function
r Tilfældige tal fra den anførte fordeling
q Fraktil, quantile, i fordeling

3.1 Uniform fordeling

Den uniforme bruges i tilfælde hvor der er lige sandsynlighed for alle dele. Eksempelvis med medarbejderes mødetid.

$$f(x) = \frac{1}{\beta - \alpha}$$
$$\mu = \frac{\alpha + \beta}{2}$$
$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

3.2 Normalfordeling

Skal en model opskrives kan det, og kan det antages at den er IID samt følger normalfordeling, kan den blot opskrives som en stokastisk variabel X for datamængden x.

$$\bar{X} \sim N\left(mean(x), sd(x)^2\right)$$

Normalfordelingen bruges bruges når vi har en ligelig fordelingen om middelværdien, øhm, et eller andet gøgl, ved sku ikke lige hvad jeg skal skrive her.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$\mu = \mu$$
$$\sigma^2 = \sigma^2$$

Husk: Det er standardafvigelsen vi skal bruge her! Hvis det er i interval se 2.5.1

dnorm(x , **mean**, **sd**)

Standardisering:

$$Z = \frac{X - \mu}{\sigma}$$

Tip
Husk at en kvartil i normalfordelingen vil have samme værdi i en anden normalfordeling, dette kan dermed bruges til at finde σ og μ
Et eksempel ses her:
1 % af brødet vejer mindre end 600 g
5 % af brødet vejer mere end 650 g
Hvad er middelværdien og standardafvigelsen?
Til dette skal vi benytte standardisering.

$$P(Z < z_{0.99}) = 0.01 \quad P(Z > z_{0.05}) = 0.05$$

Nu kan normalfordelingskvartilen benyttes, Vi beregner :

$$z_{0.99} = -2.326 \quad z_{0.05} = 1.645$$

Nu Kan vi så benytte os af standadisering

$$-2,326 = \frac{600 - \mu}{\sigma} \quad 1.645 = \frac{650 - \mu}{\sigma}$$
$$\downarrow \text{Der solves i maple} \downarrow$$
$$\mu = 629.3g \quad \sigma = 12.6g$$

3.3 Log-normalfordeling

Samme som normalfordeling bare lige med logaritme, den kan laves om således

$$X = \frac{ln(Y) - \alpha}{\beta}$$

3.4 Eksponential

Eksponential og poisson minder meget om hinanden, forskellen er dog at *Poisson* fortæller noget om antallet af hændelser pr. enhed, mens Eksponential fortæller noget om enhederne mellem hændelserne.

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x > 0, \beta > 0 \\ 0 & ellers \end{cases}$$

4 Fordeling for gennemsnittet

4.1 Varians for $\bar{X}$

$$var\left(\bar{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n} var\left(X_i\right) = \frac{\sigma^2}{n}$$

4.2 Spredning af $\bar{X}$ / Standard error of the mean

Når vi benytter $\bar{x}$ som estimat for μ

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sigma_{\bar{X}-\mu} = \frac{\sigma}{\sqrt{n}}$$

4.3 Fordelingen for den standardisere fejl

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1^2)$$

$$T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t$$

5 Konfidensinterval

5.1 Brug af t-fordeling til konfidensintervallet for μ

$$\alpha = \text{konfidensen}$$
$$\bar{x} \pm t_{\left(1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}}$$

Kender vi ikke n kan vi nøjes med

$$\alpha = \text{konfidensen}$$
$$\bar{x} \pm t_{\left(1-\frac{\alpha}{2}\right)} \cdot s$$

I det lange løb vil vi finde den sande værdi i 95% af tilfældene.

$$P\left(\frac{\mid \bar{X} - \mu \mid}{\frac{S}{\sqrt{n}}} < t_{0,975}\right) = 0,95$$
$$P\left(\bar{X} - t_{0,975}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0,975}\frac{S}{\sqrt{n}}\right) = 0,95$$

I R gøres dette:

```
mean(x)+ c(-1,1)*qt(0.975,df=(n-1))*sd(x)/sqrt(n)
```

5.2 Planlægning af studie med krav til præcision

Her skal det huskes at MarginofError er en plusminus. Altså vil en ME med bredden 0.5 være lig 0.25 da den smalede bredde er lig 0.5

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

Husk her at Z i R er en qnorm.

5.3 Variansestimat

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$
$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2}$$

5.4 Konfidensintervaller på varians

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

```
(n-1)*sd(x)^2/ c(qchisq(0.975,(n-1) ) ,qchisq(0.025,(n-1)))
```

5.5 Konfidensintervaller på spredning

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

6 Hypotesetest

6.1 Hypoteser

En hypotese kan opskrives

$$H_0 : \mu = \mu_0$$

Denne kan så afkræftes ved brug af p-værdi

6.2 P-værdi

P-værdien er sandsynligheden for at få hypotetiske værdier som er mindst lige så ekstreme som de observerede tilfælde.
For tovejs værdier beregnes teststørrelsen således:

$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$
$$p = 2 \cdot P\left(T > |\, t_{obs} \,|\right)$$

For envejs hvor vi tester $>$ eller $<$ ændrer vi blot til

$$p = P\left(T > |\, t_{obs} \,|\right)$$

Fælles for begge gælder det at T følger T-fordelingen med frihedsgrad (n-1) μ_0 er middelværdien fra vores nulhypotese.

```
tobs<- (mu-mu0)/(s/sqrt(n))
pvalue<-2*(1-pt(abs(tobs),df=n-1))
```

6.2.1 Fortolkning af p-værdien

Tabel 2 Fortolkning af p-værdi

$p < 0.001$	Meget stærkt bevis mod H_0
$0.100 \leq p < 0.1$	Stærk bevis mod H_0
$0.01 \leq p < 0.05$	Bevis mod H_0
$0.05 \leq p < 0.1$	Svagt bevis mod H_0
$0.1 \leq p$	Maginalt eller intet bevis mod H_0

6.3 Hypotesetest

Vi forkaster en hypotese såfremt den beregnede p-værdi er mindre end α
Statistisk signifikans
Er p-værdien mindre end signifikanen (α) siges effekten at have statistisk signifikans.

6.4 Kritisk værdi

Den kritiske værdi er defineret ved brug af T-fordelingen. Har vi med et tovejs sampletest at gøre vil det være som følger

$$t_{\frac{\alpha}{2}} \quad \& \quad t_{1-\frac{\alpha}{2}}$$

Alternativt vil det med envejs være således:

$$t_{\alpha} \quad \& \quad t_{1-\alpha}$$

Vi beregner disse i R således

```
c(-1,1)*qt(0.975,df)
```

Husk at 0,975 benyttes da det er med 95% interval. Og df er frihedsgraden.

Ved at gøre brug af den kritiske værdi kan vi også undersøge hvorvidt nulhypotesen skal forkastes. I det tilfælde, for tovejs, at:

$$\mid t_{obs} \mid > t_{1-\frac{\alpha}{2}}$$

Eller for etvejs

$$\mid t_{obs} \mid > t_{1-\alpha}$$

vil vi forkaste nulhypotesen. Er du vågen ser du at ovenstående passer på konfidensinterval, hvis μ_0 altså ligger uden for konfidensintervallet for μ vil den forkastes.

6.5 Fejltyper ved hypotesefejl

6.5.1 Type 1, forkastelse af H_0 når sand

$$P(Type1) = \alpha$$

6.5.2 Type 2, ingen forkastelse af H_0 når falsk

$$P(Type2) = \beta$$

6.6 Styke, Power

Sandsynligheden for korrekt at forkaste H_0
Denne bestemmes med power.t.test i R

6.7 Stikprøvestørrelse

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

Til denne bruges også power.t.test. Husk hertil at angive type samt alternativ.

```
power.t.test(delta=1.363,sd=1.521,sig.level = 0.05, power=0.9,
type="one.sample", alternative = "one.sided")
```

7 Hypotesetest af to grupper

Husk!:
Hvis de to grupper er parrede kan du ofte blot opstille et en hypotesetest for en gruppe, som så er differencen hvis det er den du vil undersøge. Som regel er det nemmere bare at lave for de to sataner.

7.1 Teststørrelsen

$$\delta = \mu_2 - \mu_1$$
$$H_0 : \delta = \delta_0$$
$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

7.2 Frihedsgrad

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

7.3 P-værdi

Samme som tidligere.

7.4 Konfidensinterval for $\mu_1 - \mu_2$

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

7.5 Parrede sæt

Hvis to sæt er parret kan et tredje sæt laves som er lig $x - y$ dette kan så bearbejdes som et enkelt sæt med $n = n_x = n_y$ observationer.

8 Simulationer i R

8.1 Replicate eller ej?

Replicate bruges i de tilfælde hvor vi skal beregne konfidensintervaller af en slags, ellers vil vi i det fleste tilfælde kunne nøjes med at benytte eksempelvis rnorm. Dette skyldes at denne gentage beregninger for rnorm og lignende et antal gange

8.2 Fejlophobning / Error propagation

Ofte denne hvis intet første indskud / Er der en funktion, næsten altid denne
Denne skal ofte bruges. En typisk formulering vil ligne: "An approximate value for the variance is:"
Denne skal lige kigges på når man kommer til noget hvor de skriver **approximately**

$$Var(A) \approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2$$

8.2.1 Eksempel

$$\begin{aligned} \sigma_1^2 = 0.01^2 \quad \sigma_2^2 = 0.02^2 \quad x = 3 \quad y = 2 \qquad f(x,y) = x \cdot y \\ Var(A) &\approx y^2\sigma_1^2 + x^2\sigma_2^2 \\ &\approx 3^2 \cdot 0.01^2 + 2^2 \cdot 0.02^2 \\ &\approx 0.0025 \end{aligned}$$

8.3 Simulation i forhold til funktion

Til dette skal vi blot lave simple simulationer, og altså ikke gøre brug af replicate. Er det eksempelvis i forhold til normalfordeling ville en simulation af en størrelse kunne gøres således

```
k <- antal gentagelser
m <- middelv rdi for fordeling
s <- spredning for fordeling
X <- rnorm(k, m , s)
Y <- rnorm(k, m2, s2)
A <- X*Y
```

Fordelen ved dette vil være at man så nu kan tage middelværdi eller lignende til den simulerede mængde, og hvis dette er en funktion kunne vi nu undgå en masse beregninger ved bare at lade R simulere det. Med areal som ovenstående ville vi bare kunne

8.3.1 Sandsynlighed for bestemt *udfald*

Hvis vi vil beregne sandsynligheder for et udfald på at vores A afviger mere end eksempelvis 0.1 vil vi kunne beregne dette således

```
mean(abs(A-6)>0.1)
```

Dette vil returnere en værdi mellem 0 og 1 svarende til sandsynligheden da vi her blot får en række af booleans.

8.4 Bootstrapping

Fælles for alle underdele af bootstrapping er at vi starter med at simulere noget data. Derefter beregne vi en forskel, eksempelvis på median, og til slut beregne vi en quantile.

8.4.1 Parameter

Når vi har med parametre at gøre i bootstrapping betyder dette blot at vi her antager en fordeling og benytter denne til simuleringen.
Herunder vil der være brugt rnorm.

8.4.1.1 Onesample

```
k <- 10000
x <- data
n <- lenght(data)
Median <- function(x){x, median}
xSamples <- replicate(k, rnorm( n , mean(x), sd(x) ) )
xMedian <- apply(xSamples,2,median)
quantile(xMedian, c(0.005,0.995))
```

8.4.1.2 Twosample

```
k <- 10000
x <- data
nx <- lenght(x)
y <- data
ny <- lenght(y)
Median <- function(x){x, median}
xSamples <- replicate(k, rnorm( nx , mean(x), sd(x) ) )
ySamples <- replicate(k, rnorm( ny , mean(y), sd(y) ) )
difMedian <- apply(xSamples,2,median)-apply(ySamples,2,median)
quantile(difMedian, c(0.005,0.995))
```

8.4.2 Ikkeparameter

Når vi har med ikkeparametre at gøre i bootstrapping betyder dette blot at vi ikke antager en fordeling, men istedet for blot benytte udtagninger af de data vi har fået opgivet.

8.4.2.1 Onesample

```
k <- 10000
x <- data
Median <- function(x){x, median}
xSamples <- replicate(k, sample(x, replace = TRUE))
xMedian <- apply(xSamples,2,median)
quantile(xMedian, c(0.005,0.995))
```

8.4.2.2 Twosample

Et lille fedt eksempel hvor vi her vil undersøge 99 % konfidensintervallet for medianen, hvis vi ville finde for noget andet kunne vi bare ændre på median. Vi kunne også laves vores egen funktion

```
k <- 10000
x <- data
y <- data
Median <- function(x){x, median}
xSamples <- replicate(k, sample(x, replace = TRUE))
ySamples <- replicate(k, sample(y, replace = TRUE))
difmedian <- apply(xSamples,2,median) - apply(ySamples,2,median)
quantile(difmedian, c(0.005,0.995))
```

9 Lineær regression

9.1 Opstil en lineær model

$$y_i = \beta_0 + \beta_1 \cdot x_i$$

9.2 Opstil den lineære regressionsmodel

$$Y_i = \beta_0 + \beta_1 \cdot x_i$$

Y_i er den afhængige variabel

x_i er den forklarende

ϵ_i er afvigelsen (error), antages IID og $N(0, \sigma^2)$

9.3 Mindste kvadraters metode

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2$$

$$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)(x_i - \bar{x})}{Sxx}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{x}$$

9.4 Estimat af standardafvigelse

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

$$\hat{\sigma_{\beta_0}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}$$

$$\hat{\sigma_{\beta_1}} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

9.5 Hypotesetest

$$T_{\beta_0} = \frac{\hat{\beta}0 - \beta_{0,0}}{\sigma_{\beta 0}}$$

$$T_{\beta_0} = \frac{\hat{\beta}1 - \beta_{0,1}}{\sigma_{\beta 1}}$$

$$df = n - 2$$

I R

```
fit <- lm(y~x)
summary(fit, level=0.95)
```

9.6 Konfidensintervaller for parametre

$$\hat{\beta}_0 \pm t_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\beta 0}$$
$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\beta 1}$$

I R

```
fit <- lm(y ~ x)
confint(fit,level=0.95)
```

9.7 Konfidensinterval for linjen

$$\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_0\right) \pm t_{(1-\frac{\alpha}{2})} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$$

9.8 Prædiktionsinterval for linjen med individuel varians

$$\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_0\right) \pm t_{(1-\frac{\alpha}{2})} \cdot \sigma \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$$

9.9 Forklarede varians

$$r^2 = cor^2$$

Den uforklarede vil derfor blot være 1- den forklarede
I R

```
cor(x,y)^2
```

Eller

```
fit <- lm(y~x)
summary(fit)
```

9.10 Beregning er residualier for specifikt datapunkt

For at beregne dette skal vi blot indsætte det spicifikke punkt i vores funktion. For at finde residualierne kan vi så trække den beregnede fra den observerede og vi har nu residualierne. *Eksempel fra eksamenopgave*

$$y = 413 \qquad x_1 = 313,8 \qquad x_2 = 54,4$$
$$\beta_0 = -81,8330 \qquad \beta_1 = 0.7893 \qquad \beta_2 = 4,7761$$
$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_{1,i} \cdot x_{2,i}$$
$$y_1 = 426,1468$$
$$\text{residualier} = y - y_1 = -13,1468$$

9.11 Multiplativ lineær regression

Ligner meget det tidligere, du har nu blot flere variable.

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

I R kan denne "modelleres"

```
fit <- lm(y~x1+x2)
```

9.11.1 Kurvelineær

Hvis vi ønsker at esimere en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

kan vi benytte den multiple lineære regression i modellen

$$x_{i,1} = x_i \qquad x_{i,2} = x_i^2$$
$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

Vi kan så udvide en model vd at indskrive disse i R

```
x3 <- x2^2
fit <- lm(y~x1+x2+x3)
```

9.11.2 Modelkontrol

Her kan du benytte qqnorm og qqline og så se om disse stemmer overens med fordelingen. Derudover kan du benytte et plot, i det tilfælde at det ligner en trumpet er det bitter røv, ommer

```
qqnorm(fit$residuals)
qqline(fit$residuals)

plot(fit$fitted.values, fit$residuals)
```

9.11.3 Konfidens og prædiktionsintervaller

For at gøre dette i en flervariabel model skal alle undtagen en variabel holdes konstant. På denne måde ser vi hvordan denne ene variabel opfører sig. Dette vil tegne sammenhængen + konfidensinterval og prædiktionsinterval.

```
fit <- lm(y ~ x1+x2+x3)
newx1<-seq(0,1,by=0.01)
ForPred <- data.frame(x1=newx1, x2=konstant, x3=konstant)
CI <- predict(fit, newdata=ForPred, interval="confidence", level=0.95)
PI <- predict(fit, newdata=ForPred, interval="prediction", level=0.95)
## Plot them
plot(x1, y, ylim=range(CI,PI,y), xlab="", ylab="")
title(xlab="x", ylab="y")
lines(newx1, CI[,"fit"])
lines(newx1, CI[,"lwr"], lty=2, col=2)
lines(newx1, CI[,"upr"], lty=2, col=2)
lines(newx1, PI[,"lwr"], lty=2, col=3)
lines(newx1, PI[,"upr"], lty=2, col=3)
    legend("topleft", c("Pr diktion","0.95 konfidensb nd","0.95
    pr diktionsb nd"),lty=c(1,2,2), col=1:3)
```

9.11.4 Kollinearitet

Læs om det i slides for uge 9

10 ANOVA - Envejs variansanalyse

Generel brug
Generelt bruges ANOVA når vi har flere data tilknyttet det samme element. Eller hvis vi eksempelvis scorer et produkt.

10.1 Opstil model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$
$$\mu = \text{ samlet middelværdi}$$

10.2 Hypotese

Dette fungere meget lig tidligere. Dog bare med flere variable

10.3 ANOVA tabellen

$$k = \text{Antal behandlinger} \qquad n = \text{Antal elementer}$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SS(Tr) = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

$$F = \frac{\frac{SS(Tr)}{k-1}}{\frac{SSE}{n-k}} = \frac{MS(Tr)}{MSE}$$

Tabel 3 ANOVA-tabel envejs

Kilde	DF	Kvadratafvigelse sum	GNS. kvadrat afvigelse	Test-størrelse	p-værdi
Behandling	k-1	SS(Tr)	$MS(Tr) = \frac{SS(Tr}{k-1}$	$F_{obs} = \frac{MS(Tr)}{MSE}$	$P(F > F_{obs})$
Residual	n-k	SSE	$MSE = \frac{SSE}{n-k}$		
Total	n-1	SST = SSE + SS(Tr)			

```
y <-c(5.01, 5.59, 3.02, 6.23, 5.13, 4.76, 5.98, 5.33, 3.46, 5.31, 4.65, 4.12)
treatm <- factor(rep(c("1998", "2003", "2011"), 6))
anova(lm(y~treatm))
```

10.4 Konfidensinterval

En enkelt forudplanlagt sammenligning mellem forskelle på behandling i og j

$$\bar{y}_i - \bar{y}_j \pm t_{1-\frac{\alpha}{2}} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$
$$df = n - k$$

Hvis vi istedet vil beregne for alle $M = \frac{k(k-1)}{2}$ kombinationer skal $\alpha_{Bonferroni}$ benyttes istedet: (k=behandling)

$$\alpha_{Bonferroni} = \frac{\alpha}{M}$$

10.5 Kritiske værdier

```
qf(0.95,l-1,n-1)
```

10.6 P-værdi

$$f_{obs} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

```
1-pf(abs(fobs),l-1,n-1)
```

11 Anova - Tovejs variansanalyse

11.1 Opstil en model

$$Y_{ij} = \mu + a_i + \beta_j + \epsilon_{ij} \qquad \epsilon \sim N(0, \sigma^2)$$
$$\mu = \text{samlet middelværdi}$$

11.2 Estimater af modellens parametre

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} y_{ij}$$
$$\hat{\alpha}_i = \left(\frac{1}{l} \sum_{j=1}^{l} y_{ij} \right) - \hat{\mu}$$
$$\hat{\beta}_j = \left(\frac{1}{k} \sum_{i=1}^{k} y_{ij} \right) - \hat{\mu}$$

11.3 Anovatabellen

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{l} (y_{ij} - \hat{\mu})^2 \; SS(Tr) = l \cdot \sum_{i=1}^{k} \hat{\alpha}_i^2$$
$$SS(Bl) = k \cdot \sum_{j=1}^{l} \hat{\beta}_j^2$$
$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{l} \left(y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu} \right)^2$$
$$F_{Tr} = \frac{\frac{SS(Tr)}{k-1}}{\frac{SSE}{(k-1) \cdot (l-1)}} F_{Tr} = \frac{\frac{SS(Bl)}{l-1}}{\frac{SSE}{(k-1) \cdot (l-1)}}$$

Tabel 4 ANOVA-tabel tovejs

Kilde	DF	Kvadratafvigelse sum	GNS. kvadrat afvigelse	Test-størrelse	p-værdi
Behandling	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	(k-1)(l-1)	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	n-1	SST = SSE + SS(Tr) + SS(Bl)			

11.4 Kritiske værdier

For anova beregnes dette ved at benytte en F-fordeling. Denne skal have frihedsgraderne for den pågældende kilde og residualerne.

```
qf(0.95,l-1,(k-1)(l-1))
```

11.5 Konfidensintervaller

Samme som ved envejs variansanalyse

11.6 P-værdi

```
1-pf(abs(Fobs),l-1,(k-1)(l-1))
```

11.7 LSD - Least Significant Difference

Hvis der er lige mange observationer i hver behandling, vil denne være n. **Husk** derudover at MSE er lig $\hat{\sigma}^2$

$$LSD = t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{2 \cdot MSE}{n}} = t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{2}{n}} \cdot \sqrt{MSE}$$

Har vi forskelligt antal observationer i hver gruppe kan vi opskrive LSD som det ses ovenfor.

$$LSD = t_{1-\frac{\alpha}{2}} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Som før kan vi benytte bonferroni hvis vi ønsker at gøre dette efterfølgende.
Husk!:
For t-fordelingen benyttes samme frihedsgrad som residualierne, for tovejs variansanalyse er dette altså $(k-1)(l-1)$

12 Inferens fra andele

12.1 Estimation af andele

$$\hat{p} = \frac{x}{n}$$
$$x = \text{antal sucesser} \qquad n = \text{antal forsøg}$$

12.2 Middelværdi og varians

$$E(X) = n \cdot p \qquad Var(X) = n \cdot p \cdot (1-p)$$
$$E(5X) = 5E(X) \qquad Var(5X) = 5^2 Var(X)$$

12.3 Konfidensinterval for én andel

$$p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$
$$z \sim N(0, 1^2)$$

12.4 Margin of error

Husk her at hvis ME er opgivet som en bredde er $ME = \frac{bredde}{2}$

$$\sigma \approx \sqrt{\frac{p \cdot (1-p)}{n}}$$
$$ME = z_{1-\frac{\alpha}{2}} \cdot \sigma \qquad \sigma = \frac{ME}{z_{1-\frac{\alpha}{2}}}$$

Skal man i stedet bestemme stikprøvestørrelse kan man omskrive således:
Hvis man har et bud på p

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

Hvis ikke man har et bud på p

$$n = \frac{1}{4} \cdot \left(\frac{z_{1-\frac{\alpha}{2}}}{ME}\right)^2$$

12.5 Teststørrelse

Denne bruges når vi tester

$$H_0 : p = 0.5 \qquad H_1 : p \neq 0.5$$

$$z_{obs} = \frac{x - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

12.6 P-værdi

$$\text{for} = : 2 \cdot P(Z > |z_{obs}|)$$
$$\text{for} <: P(Z < z_{obs})$$
$$\text{for} >: P(Z > z_{obs})$$

```
2 * (1-pnorm(abs(zobs)))
pnorm(abs(zobs))
1-pnorm(abs(zobs))
```

12.7 For to andele

12.8 Estimater

Tabel 5 Estimater, uge 12

Sammenhæng mellem brug af p-piller og risikoen for hjerteinfarkt

	Infarkt	Ikke infarkt	Total
p-piller	23	34	$n_1 = 57$
Ikke p-piller	35	132	$n_2 = 167$
	$x = 58$		$n = 224$

Estimater i hver stikprøve

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \quad \hat{p}_2 = \frac{35}{167} = 0.2096$$

Fælles estimat:

$$\hat{p} = \frac{23+35}{57+167} = \frac{58}{224} = 0.2589$$

$$\hat{p}_1 = \frac{x_1}{n_1} \qquad \hat{p}_2 = \frac{x_2}{n_2}$$
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

12.8.1 Konfidensinterval

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

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

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{x_1 \cdot (n_1 - x_1)}{n_1^3} + \frac{x_2 \cdot (n_2 - x_2)}{n_2^3}}$$

12.8.2 Teststørrelse

HUSK!:
Regn estimaterne først.

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

Husk!:
n_1 og n_2 er her rækketotalerne. Hvis du er helt blank så se slides uge 12.
OBS
I sjældne tilfælde, jeg aner ikke hvorfor, sker det at vi dividere med $\frac{1}{n_1 + n_2}$

12.9 For flere andele

12.10 For flere andele

12.10.1 Gruppeestimat

Såfremt nul-hypotesen gælder, vil vi forvente at den j'te gruppe har e_{1j} succeser og e_{2j} fiaskoer

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

$$e_{2j} = n_j \cdot (1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

12.10.2 Forventet værdi i antalstabel

$$e_{ij} = \frac{(i'throwtotal) \cdot (j'thcolumtotal)}{Total}$$

12.10.3 Teststørrelse

Denne bruges når vi tester

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$
$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

12.10.4 Kritisk værdi

Husk her at χ^2-fordeling med $(r-1)(c-1)$ frihedsgrader samt alle forventede værdier $e_{ij} \geq 5$

```
qchisq(0.99,6)
```

12.10.5 P-værdi

Husk her at χ^2-fordeling med $(r-1)(c-1)$ frihedsgrader samt alle forventede værdier $e_{ij} \geq 5$

13 Plots

13.1 BoxPlot

```
x <- c(1,2,3,4,5,6,7,8,9,10)
boxplot(x)
```

13.2 Scatterplot

```
x <- c(1,2,3,4,5,6,7,8,9,10)
plot(x)
```

13.3 Plot normalfordeling

```
x <- seq(-2,2,length=100)
fx <- dnorm(x,0,1)
plot(x, fx, type="l", xlab="x value", ylab="density")
```

13.3.1 Marker område på fordeling, her normalfordeling

```
i <- x >= -1 & x <= 1
polygon(c(-1,x[i],1),c(0,fx[i],0),col="red")
```