Statistik og Sandsynlighedsregning 2

Repetition og eksamen

T-test

Overheads til forelæsninger, onsdag 7. uge

Normalfordelingen

Erfaringsmæssigt er normalfordelingen velegnet til at beskrive variationen i mange variable, blandt andet tilfældige fejl på målinger.

Fordelingens sandsynlighedsteoretiske egenskaber giver et solidt matematisk grundlag at bygge på.

Normalfordelingen er symmetrisk, har et maximum og er fuldstændigt beskrevet ved to parametre, nemlig midddelværdien og variansen (eller standardafvigelsen).

Hvis X er normalfordelt med middelværdi μ og varians σ^2 har X tæthed

$$\varphi(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Vi skriver $X \sim N(\mu, \sigma^2)$

Vi antager fremover at vi har observationer

$$x \quad = \quad (x_1, \ldots, x_n)$$

af

$$X \quad = \quad (X_1, \ldots, X_n)$$

hvor $X_i, i = 1, \ldots, n$ er normalfordelte med samme varians σ^2, men muligvis med forskellig middelværdi μ_i.

T-test benyttes når man vil teste hypoteser om middelværdien af normalfordelte variable. Vi ser på 3 forskellige slags t-test:

- *One-sample t-test* benyttes når man vil teste om uafhængige, identisk fordelte normale variable kommer fra en fordeling med en kendt middelværdi.
- *Uparret t-test* benyttes når man vil sammenligne middelværdierne i to grupper af uafhængige, identisk fordelte normale variable. Det antages at der er samme varians i de to grupper, og man ønsker at teste om middelværdierne er ens.
- *Parret t-test* benyttes når man vil teste om differencen mellem sammenhørende par af observationer af normalfordelte variable med samme varians kommer fra en normalfordeling med kendt middelværdi.

One-sample t-test

Statistisk model:

$$(\mathbf{R}^n, (N_{(\mu,\sigma^2)})_{(\mu,\sigma^2)\in\mathbf{R}\times]0,\infty[})$$

hvor $N_{(\mu,\sigma^2)}$ har tæthed

$$\varphi_{(\mu,\sigma^2)}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{s=1}^{n}(x_s-\mu)^2\right\}$$

Hypotese:

$$H \;:\; \mu = \mu_0$$

Estimatorer under den fulde model:

$$\hat{\mu} = \frac{1}{n}\sum_{s=1}^{n} x_s = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{s=1}^{n}(x_s-\bar{x})^2$$

$$\text{dog benyttes} \qquad s^2 = \frac{1}{n-1}\sum_{s=1}^{n}(x_s-\bar{x})^2$$

og

$$\hat{\mu}\sim N(\mu,\frac{\sigma^2}{n}) \quad ; \quad \text{SSD} = n\hat{\sigma}^2 = (n-1)s^2 \sim \sigma^2\chi^2_{n-1} \quad ; \quad \hat{\mu}\perp s^2$$

Estimatorer under hypotesen:

$$\tilde{\mu} = \mu_0$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{s=1}^{n}(x_s-\mu_0)^2$$

og

$$n\tilde{\sigma}^2 \sim \sigma^2\chi^2_n$$

Kvotienttesstørrelsen for test af $\mu=\mu_0$ er

$$Q(x) = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{\frac{n}{2}}$$

og testsandsynligheden er givet ved

$$\epsilon(x) = 2P\left(T_{n-1} \geq \frac{|\bar{x}-\mu_0|}{s/\sqrt{n}}\right)$$

hvor T_{n-1} er T–fordelt med $n-1$ frihedsgrader.

Bemærk: Vi beregner gennemsnittet, trækker den formodede middelværdi fra og dividerer med et estimat af standardafvigelsen. Vi har altså en teststørrelse, der under hypotesen har middelværdi 0 og varians 1.

Bemærk også at under hypotesen er $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$, dvs at $\sqrt{n}(\bar{X} - \mu_0) \sim N(0, \sigma^2)$. Desuden er $(n-1)s^2 \sim \sigma^2 \chi^2_{n-1}$ og $\bar{X} \perp s^2$.

Definitionen af en t-fordeling med f frihedsgrader er netop

$$T = \frac{U}{\sqrt{Z/f}}$$

hvor $U \sim N(0,1)$ og $Z \sim \chi^2_f$ og $U \perp Z$.

Vi kan altså direkte se at vores teststørrelse

$$T = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{n}(\bar{x} - \mu_0)/\sigma}{\sqrt{((n-1)s^2/\sigma^2)/(n-1)}}$$

er t-fordelt med $n-1$ frihedsgrader.

VIGTIGT:

Testsandsynligheden (p-værdien) angiver sandsynligheden for at man under et lignende eksperiment observerer den samme eller en større afstand mellem gennemsnittet og den formodede middelværdi som den man har observeret i det konkrete eksperiment.

Hvis denne sandsynlighed er stor kan vi godt tro på at den observerede forskel blot skyldes tilfældig variation. Hvis sandsynligheden er lille vil vi være tilbøjelige til ikke at tro på at det udelukkende skyldes tilfældigheder, men snarere at data ikke stammer fra en fordeling med den formodede middelværdi.

Hvis testsandsynligheden er mindre end 0.05 siger vi at middelværdien er signifikant forskellig fra μ_0 på 5% niveau.

Eksempel

I de første 12 dage i marts var dagens maksimumstemperatur observeret til 8, 8, 7, 6, 7, 9, 11, 9, 11, 8, 11 og 11 °C. Middelværdien af marts måneds maximumstemperatur plejer at ligge på 7 °C. Vi vil gerne teste om de observerede temperaturer kommer fra en fordeling med middelværdi 7. Vi antager at data er normalfordelt.

Vores teststørrelse bliver

$$\begin{aligned} T &= \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \\ &= \frac{\sqrt{12}(8.8333 - 7)}{1.8007} \\ &= 3.5269 \end{aligned}$$

For at finde testsandsynligheden skal teststørrelsen vurderes i en t-fordeling med 11 frihedsgrader. Vi kan slå op, f.eks i R, at

$$P(T_{11} \geq |3.5269|) = 0.00474$$

Fortolkning?

Hvis maksimumstemperaturen i marts måned 2008 er normalfordelt med middelværdi 7 °C, da vil en tilfældig stikprøve bestående af 12 observationer med 0.474% sandsynlighed have et gennemsnit på 8.333 °C eller over eller på 5.666 °C eller under. Da denne sandsynlighed er lille afviser vi at tro på at marts måned i 2008 kommer fra en fordeling med middelværdi 7 °C, men snarere fra en fordeling med en højere middelværdi. Vi estimerer middelværdien til 8.333 °C.

I R kunne vi beregne det således:

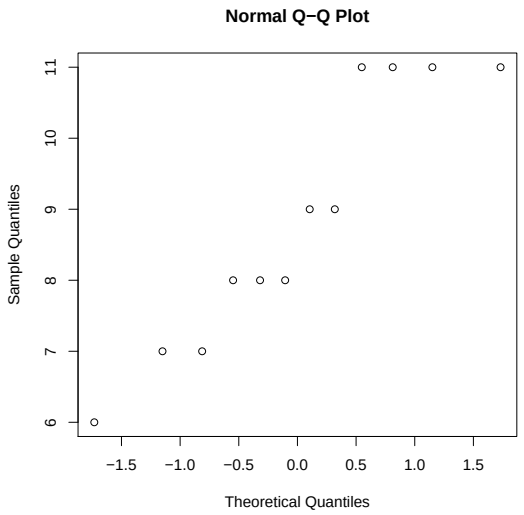
Først indlæser vi værdierne i en datavektor

```
> temp <- c(8,8,7,6,7,9,11,9,11,8,11,11)
```

Derefter ser vi om normalfordelingsantagelsen er rimelig, for eksempel ved et QQ-plot

```
> qqnorm(temp)
```

Plottet ses på næste slide. Der er ikke mange punkter, og de er afrundede til hele tal. Det er derfor svært at afgøre udelukkende fra dette data om normalfordelingsantagelsen holder. Vi vil dog acceptere den, da punkterne trods alt ligger på en nogenlunde ret linie.



Vi skal bruge antallet af observationer, og definerer derfor

```
> n <- length(temp)
```

Vi kan nu beregne vores teststørrelse

```
> Tstatistic <- sqrt(n)*(abs(mean(temp)-7))/sd(temp)
```

Bemærk at vi har taget den absolutte værdi af forskellen mellem gennemsnittet og den formodede middelværdi. Vi får

```
> Tstatistic
[1] 3.526932
```

Vi kan nu beregne vores testsandsynlighed

```
> 2*(1-pt(Tstatistic, df=n-1))
[1] 0.004740328
```


Det havde selvfølgelig været nemmere at bruge den prædefinerede funktion `t.test`, hvor vi blot skal angive datavektoren og værdien af hypotesen $\mu_0 = 7$:

```
> t.test(temp, mu=7)

One Sample t-test

data:  temp
t = 3.5269, df = 11, p-value = 0.00474
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 7.689240 9.977427
sample estimates:
mean of x
 8.833333
```

Outputtet skal læses som følger:

```
One Sample t-test
```

Først angives hvilken test der er blevet foretaget. R kan altså selv finde ud af om det er one-sample eller two-sample udfra formatet i argumenterne til funktionskaldet.

```
data:  temp
```

Her fortælles hvilke data der er blevet analyseret. Det kan være nyttigt hvis man f.eks har gemt testet som et objekt man senere skal se på, og derfor måske ikke kan huske hvordan funktionskaldet så ud.


```
t = 3.5269, df = 11, p-value = 0.00474
```

Her angives det vi er interesseret i, nemlig teststørrelsen t, antal frihedsgrader (`df` = n-1) og testsandsynligheden (`p-value`). Bemærk at p-værdien er mindre end 0.05, og vi vil derfor afvise vores hypotese på 5% niveau. Vi ved også at et 95% konfidensinterval ikke vil indeholde værdien 7.

```
alternative hypothesis: true mean is not equal to 7
```

Her angives hvilken værdi vi har testet middelværdien imod.


```
95 percent confidence interval:
 7.689240 9.977427
```

Vi får også et 95% konfidensinterval for den sande middelværdi. Intervallet angiver mængden af mulige middelværdier som gennemsnittet af data ikke er signifikant forskelligt fra. Bemærk at vores hypotese $\mu_0 = 7$ IKKE er indeholdt i konfidensintervallet. Derfor ved vi også at p-værdien er mindre end 0.05.

```
sample estimates:
mean of x
 8.833333
```

Til sidst får vi angivet gennemsnittet af data.

Bemærk at man kunne have fået den samme information (og mere til) ved at lave en lineær regression på et konstant led:

```
> summary(lm((temp - 7) ~ 1))

Call:
lm(formula = (temp - 7) ~ 1)
Residuals:
    Min      1Q  Median      3Q     Max
-2.8333 -1.0833 -0.3333  2.1667  2.1667
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8333     0.5198   3.527  0.00474 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Residual standard error: 1.801 on 11 degrees of freedom
```

Outputtet skal læses som følger:

```
Call:
lm(formula = (temp - 7) ~ 1)
```

Først angives funktionskaldet, og dermed hvilken analyse der er blevet foretaget. Formlen siger at `temp` søges forklaret ved et konstant led, angivet ved 1. Bemærk at vi har trukket den formodede middelværdi på 7 °C fra, således at testet for om det konstante led er 0 giver mening.


```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8333 -1.0833 -0.3333  2.1667  2.1667
```

Her gives information til at vurdere fordelingen af residualerne, nemlig minimum, 25%, 50% (medianen), 75% kvartilen og maximum. Hvis normalfordelingsantagelsen skal være rimelig, bør maximum og minimum være nogenlunde lige store i absolut værdi, det samme gælder for 25% og 75% kvartilerne. Gennemsnittet er defineret til at være 0, og medianen skal helst være tæt på 0.

Residualerne er her differencen mellem de faktiske målte temperaturer og deres gennemsnit, der jo er estimatet af det konstante led.

Residualinformationen gives ikke ved `t.test`, så her får vi ekstra information.


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8333     0.5198   3.527  0.00474 **
```

Her får vi først estimatet på det konstante led (interceptet), nemlig 1.8333. Da vi jo har trukket 7 °C fra alle målinger, kan vi se at det netop svarer til gennemsnittet fra før på 8.8333 °C. Dernæst anives et estimat af standardfejlen på middelværdiestimatet. Standardfejlen er $\hat{\sigma}/\sqrt{n}$, dvs $\hat{\sigma} = 0.5198 \cdot \sqrt{12} = 1.80064$. Dernæst angives t-teststørrelsen og p-værdien, der ses at være de samme som ved `t.test`.

Information om standardfejlen gives ikke ved `t.test`, så her får vi også ekstra information. Den kan dog beregnes udfra t-teststørrelsen og gennemsnittet hvis man kender stikprøvestørrelsen.


```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Her angives symboler til hurtigt at identificere signifikansniveauet for den beregnede p-værdi. *** betyder at p-værdien er mindre end 0.001, ** betyder at p-værdien er mellem 0.001 og 0.01, osv med de øvrige symboler. De to stjerner efter forrige linie angiver altså at vi kan afvise hypotesen på 1% niveau (og derfor selvfølgelig også på 5% niveau), men ikke på 0.1% niveau.

```
Residual standard error: 1.801 on 11 degrees of freedom
```

Til sidst angives et estimat for σ og antallet af frihedsgrader. Dette er også en ekstra information vi ikke får i `t.test`, men som ovenfor nævnt kan den let beregnes. Vi kan se fra antallet af frihedsgrader at antallet af observationer er 12. Bemærk at det er s, der angives, dvs

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Bemærk at vi også med `lm` kan få konfidensintervallet for middelværdien som ved `t.test`:

```
> confint(lm( temp ~ 1 ))
                2.5 %   97.5 %
(Intercept) 7.68924 9.977427
```

Her har vi ikke trukket 7 °C fra for at få samme konfidensinterval som ved `t.test`.

Er det rimeligt at antage at forudsætningerne for analysen i dette tilfælde er opfyldt?

Normalfordelingsantagelsen?

Uafhængighed?

Konstant middelværdi?

Uparret t-test:
Sammenligning af middelværdi i to normalfordelinger

Observation

$$x = (x_{rs})_{r=1,2,s=1,\ldots n_r}$$

fra

$$X = (X_{rs})_{r=1,2,s=1,\ldots n_r}$$

uafhængige normalfordelte variable

$X_{rs} \sim N(\mu_r, \sigma^2)$ med $\mu_r \in \mathbf{R}$ og $\sigma > 0$.

Sæt $n = n_1 + n_2$. X har tæthed

$$\varphi_{\mu_1,\mu_2,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\mu_r)^2\right\}$$

Statistisk model og hypotese

Statistisk model

$$(\mathbf{R}^n, (N_{(\mu_1,\mu_2,\sigma^2)})_{(\mu_1,\mu_2,\sigma^2)\in\mathbf{R}^2\times]0,\infty[})$$

hvor $N_{(\mu_1,\mu_2,\sigma^2)}$ har tæthed $\varphi_{\mu_1,\mu_2,\sigma^2}(x)$

Hypotese:

$$H: \ \mu_1 = \mu_2 = \mu$$

Estimatorer og teststørrelse

$$\begin{aligned}
\text{MLE under } M \ &: \ \hat{\mu}_r = \bar{x}_r \\
&\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\bar{x}_r)^2 \\
\text{Dog benyttes} \ &: \ s^2 = \frac{1}{n-2}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\bar{x}_r)^2 \\
\text{MLE under } H \ &: \ \tilde{\mu} = \bar{x} \\
&\quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\bar{x})^2 \\
\text{Dog benyttes} \ &: \ s^2 = \frac{1}{n-1}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\bar{x})^2
\end{aligned}$$

Testsandsynlighed og fordeling af estimatorer

Fordeling af MLE under M:

$$\begin{aligned}
&\hat{\mu}_1 \perp \hat{\mu}_2 \perp \hat{\sigma}^2 \\
&\hat{\mu}_r \sim N(\mu_r, \tfrac{1}{n_r}\sigma^2) \\
&n\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-2}
\end{aligned}$$

Fordeling af MLE under H:

$$\begin{aligned}
&\tilde{\mu} \perp \hat{\sigma}^2 \\
&\tilde{\mu} \sim N(\mu, \tfrac{1}{n}\sigma^2) \\
&n\tilde{\sigma}^2 \sim \sigma^2\chi^2_{n-1}
\end{aligned}$$

Kvotienteststørrelse

$$Q(x) = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{\frac{n}{2}}$$

Testsandsynlighed

$$\epsilon(x) = 2P\left(T_{n-2} \geq \frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right),$$

hvor $s^2 = \frac{1}{n-2}\sum_{r=1}^{2}\sum_{s=1}^{n_r}(x_{rs}-\bar{x}_r)^2$, og T_{n-2} er T–fordelt med $n-2$ frihedsgrader.

Bemærk: Vi beregner differencen på de to gennemsnit, trækker den formodede middelværdi fra (=0) og dividerer med et estimat af standardafvigelsen på differencen. Vi har altså en teststørrelse, der under hypotesen har middelværdi 0 og varians 1. Også her kan vi direkte se fordelingen af vores teststørrelse udfra fordelingerne af de enkelte elementer og definitionen af en t-fordeling.

Eksempel

Glukosekoncentrationen i blodet under faste er blevet målt hos en gruppe på 14 normalvægtige personer og en gruppe på 12 overvægtige personer. Vi ønsker at teste om middelværdien af glukosekoncentrationen afhænger af om BMI er over eller under 25 kg/m^2. Vi antager at data er uafhængige og normalfordelte med samme varians.

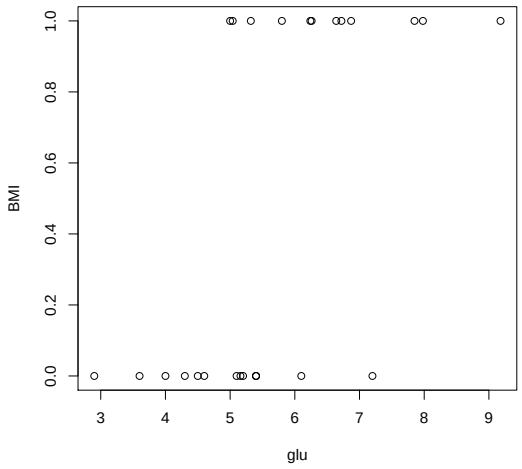
Data ligger i en dataframe:

```
> data
    glu BMI
1  4.30   0
2  3.60   0
[...]
26 7.98   1
```


Først ønsker vi at få et overblik over data ved et scatterplot:

```
> plot(data)
```

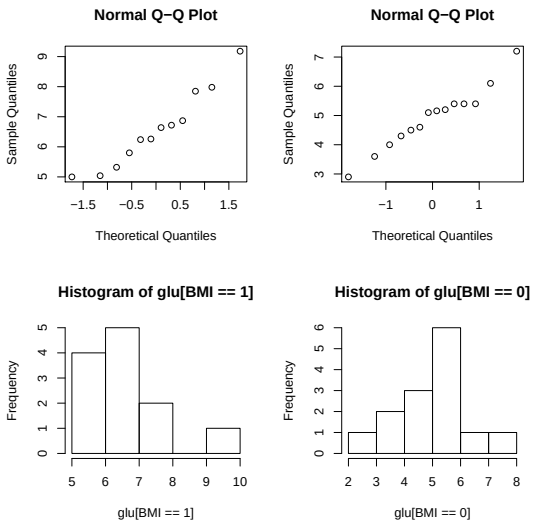
Læg mærke til at data ligger i et samlet datasæt med oplysninger om to variable, nemlig glukose og BMI. De to grupper er altså defineret ved den anden variabel (BMI), og ikke ved to enkeltstående vektorer. Dette er den bedste måde at organisere data på hvis mere komplicerede analyser skal foretages.



Dernæst vil vi teste normalfordelingsantagelsen i de to grupper.

```
> attach(data)
> par(mfrow=c(2,2))
> qqnorm(glu[BMI == 1])
> qqnorm(glu[BMI == 0])
> hist(glu[BMI == 1])
> hist(glu[BMI == 0])
> par(mfrow=c(1,1))
> detach(data)
```

Hverken QQ-plots eller histogrammer får os til at tvivle på normalfordelingsantagelsen. Der er dog få punkter at bedømme ud fra.



Dernæst vil vi lave et test for at se om vi kan antage at variansen i de to grupper er ens:

```
> var.test(glu ~ BMI, data = data)

F test to compare two variances

data:  glu by BMI
F = 0.7058, num df = 13, denom df = 11, p-value = 0.5441
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2081009 2.2568617
sample estimates:
ratio of variances
         0.7058215
```


Læg mærke til funktionskaldet:

```
> var.test(glu ~ BMI, data = data)
```

Vi ønsker at teste om variansen i to grupper kan antages at være ens. Det er variansen af glukoseobservationerne, og grupperne er defineret ved deres værdi af variablen BMI. Vi kan derfor skrive `glu ~ BMI` der læses som "glukose ved BMI". Denne skrivemåde er mere optimal når man ønsker at lave mere komplicerede analyser. Ideen er at tænke i modeller af hele datasæt i stedet for enkeltstående vektorer. Dette kan kun lade sig gøre hvis data er organiseret i et samlet datasæt.

outputtet skal læses som følger

```
F test to compare two variances

data:  glu by BMI
```

Burde være selvforklarende.

```
F = 0.7058, num df = 13, denom df = 11, p-value = 0.5441
```

Her angives først F-teststørrelsen, nemlig ratioen mellem de estimerede varianser i de to grupper. Dernæst angives frihedsgraderne: der er 13 frihedsgrader i tælleren, dvs variansestimatet baserer sig på 14 observationer og er χ^2_{13} fordelt, da vi jo ikke kender middelværdien. Tilsvarende for nævneren, hvor der er 11 frihedsgrader, dvs 12 observationer. Til sidst angives p-værdien, der er sandsynligheden for at observere en F-teststørrelse som den angivne, eller noget der er længere væk fra 1, vurderet i en F-fordeling med (13,11) frihedsgrader. Da denne sandsynlighed er stor tror vi på at varianserne kan være ens i de to grupper. Med andre ord: hvis varianserne er ens, er der 54% sandsynlighed for at varianserne i to stikprøver af størrelse 12 og 14 vil være mindst så forskellige som det observerede.


```
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2081009 2.2568617
sample estimates:
ratio of variances
         0.7058215
```

Vi får oplyst hvad hypotesen er (at ratioen er lig 1), og der angives et 95% konfidensinterval for den sande ratio mellem varianserne. Det ses at intervallet indeholder 1, som vi vidste det ville da p-værdien jo var større en 0.05. Til slut angives estimatet for ratioen, som jo netop er F-teststørrelsen.

Vi er nu klar til at lave vores t-test, da antagelserne er blevet efterprøvet.

```
> t.test(glu ~ BMI, data = data, var.equal = TRUE)

Two Sample t-test

data:  glu by BMI
t = -3.618, df = 24, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.6013325 -0.7115247
sample estimates:
mean in group 0 mean in group 1
       4.918571        6.575000
```


Outputtet skal læses som ved et one-sample t-test. Bemærk at R ved selve funktionskaldet ved at det er et two-sample t-test, der ønskes.

Da p-værdien er mindre end 0.05 vil vi afvise hypotesen om ens middelværdi i de to grupper. Vores slutmodel bliver derfor en model hvor fasteglukosekoncentrationen er normalfordelt med varians

$$s^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{T^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

dvs

```
> (4.918571-6.575)^2/((-3.618)^2*(1/14 + 1/12)
[1] 1.354392
```

og middelværdier angivet til sidst i outputtet:

```
mean in group 0 mean in group 1
       4.918571        6.575000
```

Vi skriver at vores slutmodel er

$$\begin{aligned} X &\sim N(4.92, 1.35) \\ Y &\sim N(6.58, 1.35) \end{aligned}$$

hvor X er fasteglukosen hos en person med BMI ≤ 25 og Y er fasteglukosen hos en person med BMI > 25.

Bemærk at vi også her kunne have lavet analysen ved

```
> summary(lm(glu ~ BMI, data = data))

Call:
lm(formula = glu ~ BMI, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0186 -0.7359  0.1050  0.4814  2.6050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9186     0.3110  15.814 3.42e-14 ***
BMI           1.6564     0.4578   3.618  0.00137 **
```



```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.164 on 24 degrees of freedom
Multiple R-Squared: 0.3529,Adjusted R-squared: 0.326
F-statistic: 13.09 on 1 and 24 DF,  p-value: 0.001374
```

Outputtet skal læses som før. Interceptet angiver estimatet for middelglukosen i gruppen hvor BMI-variablen = 0, dvs for personer med BMI ≤ 25. BMI estimatet angiver hvor meget middelværdien ændrer sig hvis BMI > 25, dvs at estimatet for middelværdien vil være summen af de to estimater: $4.9186 + 1.6564 = 6.575$, der netop er middelværdien angivet i `t.test`. Bemærk at p-værdien udfor BMI er den samme som i t-testet. Vi får også angivet et estimat for σ, nemlig $\sqrt{1.354392} = 1.164$.

Parret t-test

Dette test benyttes hvis man har sammenhørende par af observationer, for eksempel før og efter et indgreb på samme subjekt, og man ønsker at teste om indgrebet ændrer middelværdien. I praksis udføres testet ved at lave et one-sample t-test på differencerne.

Eksempel

Data:

Eksperimentel enhed	1	2	3	4	5	6	7	8
1. måling	4.15	6.02	6.28	5.57	5.73	7.44	5.67	5.32
2. måling	4.30	6.47	7.04	5.50	6.18	7.70	6.25	5.44

Vi vil teste om der er forskel på middelværdierne mellem første og anden måling. Dette kan i R gøres på to måder:

```
> t.test(x1,x2,paired=TRUE)

Paired t-test

data:  x1 and x2
t = -3.5327, df = 7, p-value = 0.009564
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5670571 -0.1123125
sample estimates:
mean of the differences
             -0.3396848
```


I stedet for at give de to vektorer og angive at data er parret kan differencerne analyseres:

```
> t.test(x1-x2,var.equal=TRUE)

One Sample t-test

data:  x1 - x2
t = -3.5327, df = 7, p-value = 0.009564
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5670571 -0.1123125
sample estimates:
 mean of x
-0.3396848
```


Outputtet skal læses som før. Det ses at konklusionerne er nøjagtig de samme da det er samme analyse, der er foretaget.

Estimatet for differencen er -0.34 med 95% konfidensinterval (-0.57 – -0.11). Da intervallet ikke indeholder 0 kan vi ikke acceptere hypotesen om samme middelværdi i de to målinger. Dette stemmer overens med p-værdien på 0.0096, der angiver sandsynligheden for at observere en forskel i gennemsnit af differencerne i en stikprøve på 8 par på 0.34 eller større, givet at middelværdien af differencen er 0.

Da denne sandsynlighed er lille afviser vi hypotesen om ens middelværdier.

Statistik og Sandsynlighedsregning 2

Repetition og eksamen

Lineær regression

Overheads til forelæsninger, onsdag 7. uge

Lineær regression

Observationssæt

t	x
t_1	x_1
.	.
.	.
.	.
t_n	x_n

Realisationer af stokastiske variable X_r, $r = 1, \ldots, n$

X_r'erne er indbyrdes uafhængige.

$X_r \sim N(\nu + \beta t_r, \sigma^2)$

Lineær regression

$$X_r \sim N(\nu + \beta t_r, \sigma^2)$$

Ny parametrisering

$$EX_r = \alpha + \beta(t_r - \bar{t}) \text{ for } r = 1, \ldots, n$$

Regressionslinien bliver

$$y(t) = \alpha + \beta(t - \bar{t})$$

og liniens skæring med y–aksen bliver $\alpha - \beta\bar{t}$.

Statistisk model

Linearitetsmodel

$$M_l \quad : \quad \mathrm{E}X_r = \alpha + \beta(t_r - \bar{t}), \quad (\alpha, \beta) \in \mathbf{R}^2,$$

Parameterområde under modellen $\Theta_0 = \mathbf{R}^2 \times]0, \infty[$

x er observation fra den statistiske model

$$(\mathbf{R}^n, (N_{\alpha,\beta,\sigma^2})_{(\alpha,\beta,\sigma^2) \in \mathbf{R}^2 \times]0,\infty[})$$

hvor $N_{\alpha,\beta,\sigma^2}$ har tæthed

$$\varphi_{\alpha,\beta,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{r=1}^{n} (x_r - \alpha - \beta(t_r - \bar{t}))^2 \right\}$$

MLE for $(\alpha, \beta, \sigma^2)$ er entydigt givet ved

$$\begin{aligned} \hat{\alpha} &= \bar{x} \\ \hat{\beta} &= \frac{\sum_{r=1}^{n} (x_r - \bar{x})(t_r - \bar{t})}{\mathrm{SSD}_t} \\ \hat{\sigma}_l^2 &= \frac{1}{n} \sum_{r=1}^{n} (x_r - \bar{x} - \hat{\beta}(t_r - \bar{t}))^2 \end{aligned}$$

Dog benyttes $\quad s_l^2 = \dfrac{1}{n-2} \displaystyle\sum_{r=1}^{n} (x_r - \bar{x} - \hat{\beta}(t_r - \bar{t}))^2$

$\hat{\alpha}$, $\hat{\beta}$ og $\hat{\sigma}_l^2$ (eller s_l^2) er uafhængige og

$$\begin{aligned} \hat{\alpha} &\sim N(\alpha, \frac{1}{n}\sigma^2) \\ \hat{\beta} &\sim N(\beta, \frac{\sigma^2}{\mathrm{SSD}_t}) \\ \mathrm{SSD}_l = (n-2)s_l^2 = n\hat{\sigma}_l^2 &\sim \sigma^2\chi^2_{n-2} \end{aligned}$$

Estimatet for regressionslinien $y(t)$ bliver

$$\hat{y}(t) = \bar{x} + \hat{\beta}(t - \bar{t}).$$

Den stokastiske variabel

$$Y(t) = \bar{X} + \hat{\beta}(t_r - \bar{t})$$

har fordeling

$$Y(t) \sim N(\alpha + \beta(t-\bar{t}), \sigma^2(\frac{1}{n} + \frac{(t-\bar{t})^2}{\mathrm{SSD}_t})$$

Variansen på den estimerede regressionslinie vokser med afstanden til $\bar{t}$, således at regressionslinien er bedst bestemt nær $\bar{t}$.

I praktiske anvendelser indsættes $(\bar{x}, \hat{\beta}, s_l^2)$ i stedet for parameterværdierne, når man skal angive estimatorernes og den estimerede regressionslinies fordelinger.

Test for β under linearitetsmodellen

Hypotese:

$$H_\beta \quad : \quad EX_r = \alpha + \beta_0(t_r - \bar{t}), \quad r = 1, \ldots, n, \quad \alpha \in \mathbf{R}$$

Parameterområde under hypotesen: $\Theta_\beta = \mathbf{R}\times]0, \infty[$

Statistisk model

$$(\mathbf{R}^n, (N_{\alpha,\sigma^2})_{(\alpha,\sigma^2)\in\mathbf{R}\times]0,\infty[})$$

hvor N_{α,σ^2} har tæthed

$$\varphi_{\alpha,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{r=1}^n (x_r - \alpha - \beta_0(t_r - \bar{t}))^2\right\}$$

$$\begin{aligned} \text{MLE under } H_\beta \quad \hat{\alpha} &= \bar{x} \\ \hat{\sigma}_\beta^2 &= \frac{1}{n}\sum_{r=1}^n (x_r - \bar{x} - \beta_0(t_r - \bar{t}))^2 \\ \text{Dog benyttes} \quad s_\beta^2 &= \frac{1}{n-1}\sum_{r=1}^n (x_r - \bar{x} - \beta_0(t_r - \bar{t}))^2 \end{aligned}$$

- $\hat{\alpha}$ og $\hat{\sigma}_\beta^2$ er uafhængige
- $\hat{\alpha} \sim N(\alpha, \frac{1}{n}\sigma^2)$
- $\mathrm{SSD}_\beta = (n-1)s_\beta^2 = n\hat{\sigma}_\beta^2 \sim \sigma^2\chi^2_{n-1}$

$$\text{Testsandsynlighed:} \quad \epsilon_\beta(x) = 2P\left(\mathrm{T}_\beta \geq \frac{\sqrt{\mathrm{SSD}_t}|\hat{\beta} - \beta_0|}{s_l}\right)$$

hvor $\mathrm{T}_\beta = \frac{\sqrt{\mathrm{SSD}_t}(\hat{\beta}(X) - \beta_0)}{s_l(X)}$ er T–fordelt med $n-2$ frihedsgrader.

Eksempel på eksamen

Fedtsyreprocenten er den fundamentale kvalitetsegenskab ved sæbe. Den bestemmes sædvanligvis ved langsomme kemiske laboratoriemålinger. Til lettelse af produktionskontrollen i sæbefabrikker har man foreslået at bestemme fedtsyreprocenten ved at måle sæbens elektriske ledningsevne. Ledningsevnen er let at måle, og målingerne kan udføres på produktionsstedet.

I nedenstående tabel findes en række uafhængige bestemmelser af ledningsevnen målt i milli-Siemens (mS) for en bestemt sæbetype og forskellige fedtsyreprocenter.

Fedtsyre-procent	Ledningsevne i mS			
81.3	1.40	1.20	0.90	1.00
82.2	1.75	1.50	1.70	1.80
82.3	1.52	1.52	1.67	1.67
83.0	2.10	1.95	1.85	1.90

Tabel 1: Sammenhæng mellem ledningsevne og fedtsyreprocent i sæbe

1. I R-udskriften nedenfor er data analyseret ved hjælp af en lineær regressionsmodel. Opstil den statistiske model. Redegør for forudsætningerne for analysen, og diskuter om disse kan antages at være opfyldte i det foreliggende tilfælde.
2. Angiv estimater for parametrene under regressionsmodellen og disses fordeling.
3. Er data forenelige med en hypotese om at ledningsevnen ikke afhænger af fedtsyreprocenten?
4. Er data forenelige med en hypotese om at regressionslinien har en hældning på 0.6?

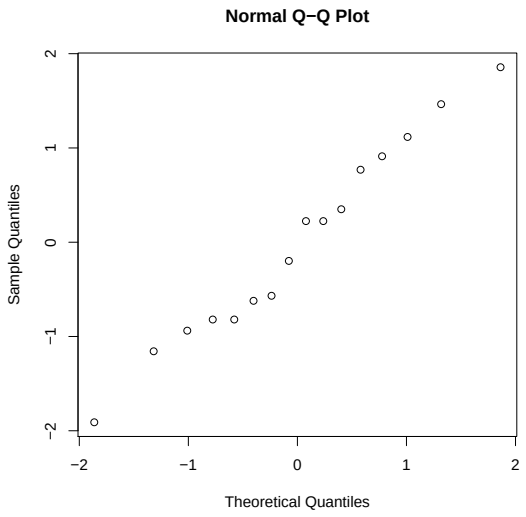
Ved besvarelsen kan nedenstående uddrag af et R-udskrift og et QQ-plot af de standardiserede residualer anvendes. Data antages at ligge i datasættet `ledning` med de to variable `fedtpct` og `ledning`.

Udskrift 1:

```
Call:
lm(formula = ledning ~ I(fedtpct - mean(fedtpct)), data = ledning)

Residuals:
      Min        1Q    Median        3Q       Max
-0.253553 -0.117800  0.002714  0.113776  0.246447

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.58938    0.03713  42.804 3.03e-16
I(fedtpct - mean(fedtpct))   0.48425    0.06146   7.879 1.63e-06
---
Residual standard error: 0.1485 on 14 degrees of freedom
```



Besvarelse

1. Opstil den statistiske model.

Data består af 16 observationer af ledningsevnen, hvor fedtsyreprocenten også er angivet. Vi angiver den $r$te måling af ledningsevnen som x_r med tilhørende fedtsyreprocent t_r. Det antages at ledningsevnen X_r er normalfordelt med middelværdi $\alpha + \beta(t_r - \bar{t})$, hvor $\bar{t}$ er gennemsnittet af de angivne fedtsyreprocenter, og varians σ^2. Den statistiske model bliver således

$$(\mathbf{R}^{16}, (N_{\alpha,\beta,\sigma^2})_{(\alpha,\beta,\sigma^2)\in\mathbf{R}^2\times]0,\infty[})$$

hvor $N_{\alpha,\beta,\sigma^2}$ har tæthed

$$\varphi_{\alpha,\beta,\sigma^2}(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^{16}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{r=1}^{16}(x_r - \alpha - \beta(t_r - \bar{t}))^2\right\}$$

1. Redegør for forudsætningerne for analysen, og diskuter om disse kan antages at være opfyldte i det foreliggende tilfælde.

Det antages at data er uafhængige. Det angives at det er uafhængige bestemmelser, så denne antagelse vil vi godtage. Derudover antages data at være normalfordelt med den givne middelværdi. Dette kan efterprøves ved at se på fordelingen af residualerne. Fra udskriftet kan vi bruge informationen om residualerne. Her bør henholdsvis min og max og 1. og 3. kvartil være nogenlunde lige store i absolut værdi. Det lader til at være fint opfyldt. Derudover bør medianen være tæt på 0, der er gennemsnittet af residualerne. Dette lader også til at være opfyldt, og vi godtager således normalfordelingsantagelsen.

QQ-plottet af de standardiserede residualer indikerer også fin overensstemmelse med normalfordelingsantagelsen, da punkterne ligger tæt på en ret linie.

2. Angiv estimater for parametrene under regressionsmodellen og disses fordeling.

Bemærk først at regressionen er foretaget på de centrerede værdier af fedtprocenten, dvs gennemsnittet af t_r er fratrukket alle fedtprocentangivelser inden analysen. Vi skal angive estimater for de 3 parametre α, β og σ og deres fordelinger. Vi har

$$\hat{\alpha} = \frac{1}{n}\sum_{r=1}^{n} x_r \quad \text{og} \quad \hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n})$$

$$\hat{\beta} = \frac{\sum_{r=1}^{n}(x_r - \bar{x})(t_r - \bar{t})}{\sum_{r=1}^{n}(t_r - \bar{t})^2} \quad \text{og} \quad \hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{r=1}^{n}(t_r - \bar{t})^2})$$

$$s^2 = \frac{1}{n-2}\sum_{r=1}^{n}(x_r - \bar{x} - \hat{\beta}(t_r - \bar{t}))^2 \quad \text{og} \quad (n-2)s^2 \sim \sigma^2\chi^2_{n-2}$$

hvor s^2 er estimatet for σ^2. Vi benytter estimaterne for α, β og σ når fordelingerne skal vurderes.

I udskriftet under `Coefficients` er α betegnet som interceptet og estimeret til 1.58938. Dette estimat er gennemsnittet af ledningsevnemålingerne. Standardfejlen for estimatet er angivet til 0.03713. Denne kunne også findes i sidste linie hvor s er angivet til 0.1485. Antallet af målinger er $n = 16$. Bemærk at $s/\sqrt{n} = 0.1485/\sqrt{16} = 0.03713$. Vi får således følgende bud på fordelingen af $\hat{\alpha}$:

$$\hat{\alpha} \sim N(1.58938, 0.03713^2)$$

I udskriftets sidste linie angives et estimat for σ til $s = 0.1485$ og frihedsgraderne er $n = 2 = 14$. Vi har følgende bud på fordelingen af s^2:

$$s^2 \sim \frac{0.1485^2}{14}\chi^2_{14} = 0.001575\,\chi^2_{14}$$

I udskriftet under `Coefficients` findes estimatet for β under `I(fedtpct - mean(fedtpct))` og er estimeret til 0.48425 med en standard fejl på 0.06146. I udskriftets sidste linie er s angivet til 0.1485. Vi kan således se at $\mathrm{SSD}_t = \sum_{r=1}^{n}(t_r - \bar{t})^2 = 0.1485^2/0.06146^2 = 5.83805$. Vi har følgende bud på fordelingen af $\hat{\beta}$:

$$\hat{\beta} \sim N(0.48425, 0.06146^2)$$

3. Er data forenelige med en hypotese om at ledningsevnen ikke afhænger af fedtsyreprocenten?

Vi skal teste hypotesen

$$H\,:\,\beta = 0$$

Dette kan gøres med t-teststørrelsen

$$T = \frac{\sqrt{\mathrm{SSD}_t}|\hat{\beta} - 0|}{s} = \frac{\sqrt{5.83805}|0.48425|}{0.1485} = 7.879$$

der under hypotesen er T-fordelt med $n - 2 = 14$ frihedsgrader. Den er allerede regnet ud i udskriftet og kan findes på linien for β. Testsandsynligheden er opgivet til at være 1.63e-06. Der er altså en meget lille sandsynlighed for at observere en værdi for $\hat{\beta}$ på 0.48425 eller længere væk fra 0 i en stikprøve af denne størrelse, hvis den sande værdi af β er 0. Vi afviser således hypotesen om at ledningsevnen ikke afhænger af fedtsyreprocenten.

4. Er data forenelige med en hypotese om at regressionslinien har en hældning på 0.6?

Vi skal teste hypotesen

$$H \,:\, \beta = 0.6$$

Dette kan gøres med t-teststørrelsen

$$T = \frac{\sqrt{\mathrm{SSD}_t}|\hat{\beta} - 0.6|}{s} = \frac{\sqrt{5.83805}|0.48425 - 0.6|}{0.1485} = 1.883$$

der under hypotesen er T-fordelt med $n - 2 = 14$ frihedsgrader. Testsandsynligheden er givet ved $2P(T \geq 1.883)$ og kan slås op i R med ordren

```
> 2*(1-pt(1.883339, df=14))
[1] 0.080597
```

Da testsandsynligheden er større end 0.05 kan vi acceptere hypotesen om en hældning på 0.6 på 5% niveau.

Hvis man ikke har mulighed for at slå testsandsynligheden op i R kan en tilnærmelse findes i MS s. 306. Her angives at $P(T_{14} \geq 2.145) = 0.025$, dvs at $P(|T_{14}| \geq 2.145) = 0.05$. Da $2.145 > 1.883$ kan vi konkludere at vi accepterer hypotesen på 5% niveau.

En endnu grovere tilnærmelse kan findes udfra betragtningen:

$$P(|T_n| \geq 1.96) > P(|Y| \geq 1.96) \quad = \quad 0.05$$

for alle $n = 1, 2, \ldots$, hvor Y er standard normalfordelt.

Konklusion: Data er forenelige med en hypotese om at regressionslinien har en hældning på 0.6.