02402: Introduktion til Statistik

Forelæsning 6: Sammenligning af to grupper og styrkeberegninger/stikprøvestørrelser

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

Overview

Motiverende eksempel – ernæringsstudie

Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge om der er en forskel i energiforbrug for forskellige typer (moderat fysisk krævende) arbejde.

I studiet har 9 sygeplejersker fra hospital A og 9 (andre) sygeplejersker fra hospital B fået målt deres energiforbrug. Målingerne ses i følgende tabel, målt i MJ:

Stikprøve fra hver hospital, $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

Eksempel – ernæringsstudie

Hypotesen om ingen forskel (i gennemsnitligt energiforbrug) ønskes undersøgt:

$$H_0: \ \mu_A = \mu_B$$

Gennemsnit og standardafvigelse for stikprøven:

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

Er data i overensstemmelse med nulhypotesen H_0?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Nulhypotese: $H_0: \ \mu_B - \mu_A = 0$

NYT: ***p*-værdi for forskel:**

$$p\text{-value} = 0.0083$$

(Beregnet under det scenarie at H_0 er sand.)

NYT: **Konfidensinterval for forskel**:

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

Overview

Definition af hypotesetest og signifikans (repetition)

Definition 3.24. Hypotesetest:

Når vi *udfører et hypotesetest*, da beslutter vi for eller imod en (nul)hypotese, ved brug af dataene.

En nulhypotese *afvises* hvis p-værdien, efter at data er observeret, er mindre end α – dvs., hvis p-værdien $< \alpha$, hvor α et på forhånd valgt *signifikansniveau*.

Hvis nulhypotesen ikke afvises, da siges den at være *godkendt*.

Definition 3.29. Statistisk signifikans:

En *effekt* siges at være *(statistisk) signifikant* hvis p-værdien er mindre end signifikansniveauet α.

Ofte (og hvis ikke andet angivet) bruges $\alpha = 0.05$.

Steps ved et hypotesetest – overblik (repetition)

Helt generelt består et hypotesetest af følgende trin:

1. Formulér hypotesen og vælg signifikansniveau α (vælg "risk-level").
2. Udregn, ud fra de observerede data, værdien af teststørrelsen.
3. Udregn p-værdien ud fra teststørrelsen holdt op imod den rette fordeling. Sammenlign p-værdien med signifikansniveauet α og konkludér.

   ELLER:

   Alternativt, konkludér ud fra de relevante kritiske værdier.

Definition og fortolkning af p-værdien (repetition)

p-værdien udtrykker *evidens* imod nulhypotesen – Tabel 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Definition 3.22 af p-værdien:

p-værdien er sandsynligheden for at observe en teststørrelse som er **mindst lige så ekstrem** som den faktiske observerede testværdi. Denne sandsynligheden udregnes under antagelse om at nulhypotesen er sand.

Kritisk værdi, konfidensinterval og hypotesetest (repetition)

Theorem 3.33: Kritisk-værdi-metode = Konfidensinterval-metode

Vi betragter et $(1-\alpha)\cdot 100\%$ konfidensinterval for μ:

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

Konfidensintervallet svarer til acceptområdet for H_0 når man tester (non-directional) hypotesen

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

(Ny) fortolkning af konfidensintervallet:

Konfidensintervallet indeholder de værdier, som vi tror på i de givne data. (De værdier som accepteres ved det tilsvarende hypotesetest.)

Overview

Method 3.49: Two-sample t-test

Beregning af teststørrelsen:

Hvis vi betragter følgende nulhypotese om forskellen i middelværdi mellem to *uafhængige* stikprøver:

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

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

så er (Welch) two-sample t-teststørrelsen:

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

Theorem 3.50: Fordelingen af (Welch) t-teststørrelsen

Welch t-teststørrlsen er (tilnærmelsesvis) t-fordelt:

Under nulhypotesen er (Welch) two-sample teststørrlsen set som stokastisk variabel:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Den følger approksimativt en t-fordeling med ν frihedsgrader, hvor

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

hvis de to populationer er normalfordelte eller stikprøvestørrelserne tilstrækkeligt store.

Method 3.51: Two-sample t-test

Et niveau α test er:

1. Udregn t_{obs} og ν, givet ved ovenstående.
2. Beregn evidensen mod *nulhypotesen*[a] $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. den *alternative hypotese* $H_1 : \mu_1 - \mu_2 \neq \delta_0$:
   $$p\text{-værdi} = 2 \cdot P(T > |t_{\text{obs}}|)$$
   hvor vi benytter t-fordelingen med ν frihedsgrader.
3. Hvis p-værdi $< \alpha$: Så afvises H_0, ellers accepteres H_0.

   ELLER

   Vi kan som alternativ (men ækvivalent) konkludere ud fra de kritiske værdier $\pm t_{1-\alpha/2}$:

   Hvis $|t_{\text{obs}}| > t_{1-\alpha/2}$ så afvises H_0, ellers accepteres H_0.

[a] Vi er som regel interesseret i $\delta_0 = 0$

Eksempel -– ernæringsstudie

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0 : \ \delta = \mu_B - \mu_A = 0$$

versus det tosidede alternativ:

$$H_1 : \ \delta = \mu_B - \mu_A \neq 0$$

Først beregninger af t_{obs} og ν:

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

og

$$\nu = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

Eksempel -– ernæringsstudie

Dernæst findes p-værdien:

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

```
## Nutrition study example: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Vurdér evidensen (Tabel 3.1):

Der er stærk evidens imod nulhypotesen.

Konklusion baseret på $\alpha = 0.05$:

Vi forkaster nulhypotesen. Der er signifikant forskel på de to grupper – sygeplejersker på Hospital B kan siges at have et større (middel)energiforbrug end sygeplejersker på Hospital A.

Overview

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

Konfidensintervallet for forskellen i middelværdi:

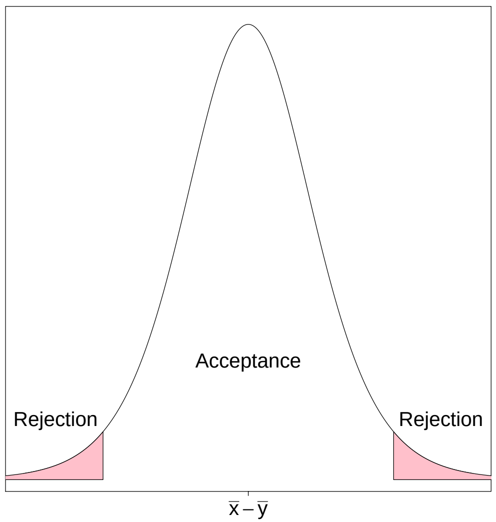
For to stikprøver $x_1, \ldots, x_{n_1}$ og $y_1, \ldots, y_{n_2}$ er $100(1-\alpha)\%$ konfidensintervallet for $\mu_1 - \mu_2$ givet ved:

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

hvor $t_{1-\alpha/2}$ er $100(1-\alpha/2)\%$-fraktilen i t-fordelingen med ν frihedsgrader givet ved Theorem 3.50 (see ovenfor).

Konfidensinterval og hypotesetest (Repetition)

Acceptområdet er de mulige værdier for $\mu_1 - \mu_2$ som ikke ligger for langt væk fra data:



Eksempel – ernæringsstudie:

Lad os finde 95% konfidensintervallet for $\mu_B - \mu_A$. Med $\nu = 15.99$ er den relevante t-fraktil givet ved:

$$t_{0.975} = 2.120$$

så konfidensintervallet bliver

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}.$$

Udregnet giver dette:

$$[0.59;\ 3.42]$$

Eksempel – ernæringsstudie – det hele i R:

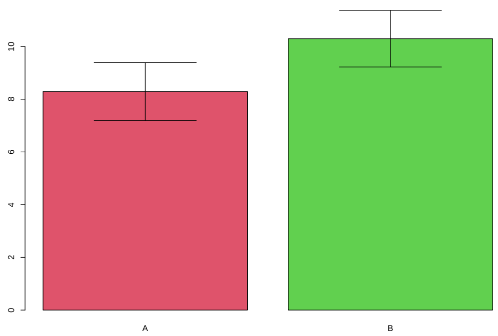
```
# Read the two samples into R
xA = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB = c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)

# Perform Welch two-sample t-test
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data:  xB and xA
## t = 3, df = 16, p-value = 0.008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5923 3.4166
## sample estimates:
## mean of x mean of y
##    10.298     8.293
```


Overview

Eksempel – ernæringsstudie – preæentation af resultater

Barplots med *error bars* ses ofte:

Et grupperet barplot med nogle "error bars" – herunder vises 95%-konfidensintervallerne for hver gruppe:



Vær varsom med at bruge "overlappende konfidensintervaller"

Man bruger faktisk så ikke den rigtige variation til at vurdere forskellen:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\mathrm{Var}(\bar{X}_A - \bar{X}_B) = \mathrm{Var}(\bar{X}_A) + \mathrm{Var}(\bar{X}_B)$$

Antag at de to standard errors er 3 og 4: Summen er 7, men $\sqrt{3^2 + 4^2} = 5$

Det korrekte forhold mellem standard errorsne er således:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

Vær varsom med at bruge "overlappende konfidensintervaller"

Remark 3.59. Regel for brug af "overlappende konfidensintervaller":

Hvis to CIer IKKE overlapper: De to grupper er signifikant forskellige

Hvis to CIer HAR overlap: Vi ved ikke hvad konklusionen er – men vi kan, f.eks, lave et CI for *forskellen* i middelværdi for at undersøge.

Overview

Motiverende eksempel – sovemedicin

Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemidler, A og B. Fra 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid i timer (forskellen på effekten af de to midler er angivet):

Sample, $n = 10$:

Person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

$\bar{x} = 1.67$ (gennemsnit)
$\bar{s} = 1.13$ (standardafvigelse)

Parret setup og analyse = one-sample analyse

```
# Read the two samples into R
x1 = c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2 = c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)

# Compute differences to get a paired t-test
dif = x2 - x1

# Perform paired t-test
t.test(dif)

##
##  One Sample t-test
##
## data:  dif
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##      1.67
```

Parret setup og analyse = one-sample analyse

```
# Another way to perform the paired t-test
t.test(x2, x1, paired = TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of the differences
##                    1.67
```


Parret ift. uafhængigt eksperiment

Fuldstændigt tilfældigt (uafhængige stikprøver)

Vi har 20 patienter, som tilfældigt fordeles på to grupper (normalt lige mange i hver gruppe). Dvs. at der er forskellige (uafhængige) patienter i de to grupper.

Parrede observationer (afhængige stikprøver):

Vi har 10 patienter, som alle får begge behandlinger (typisk noget tid imellem og tilfældig rækkefølge af behandlingerne).
Dvs. de samme patienter i de to grupper.

Eksempel – Sovemedicin – FORKERT analyse

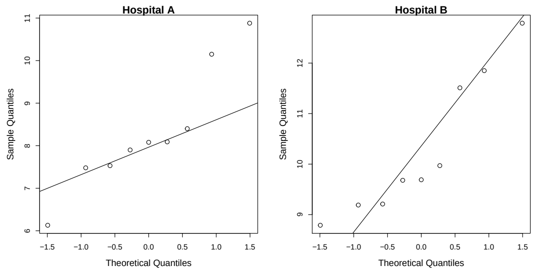
```
# WRONG analysis
t.test(x1, x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -1.9, df = 18, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.4854  0.1454
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```

Vi kan IKKE (kun) ud fra data afgøre om det er et parret eller et uparret setup – vi skal vide noget om forsøget også

Overview

Eksempel – Q-Q plot for hver af stikprøverne:

```r
# Q-Q plots separately for each sample
qqnorm(xA, main = "Hospital A")
qqline(xA)
qqnorm(xB, main = "Hospital B")
qqline(xB)
```

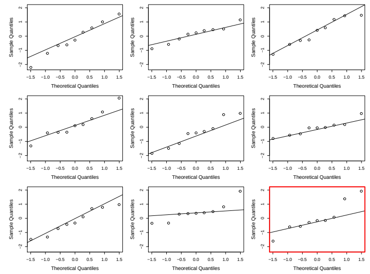


Bemærk!!

Brug IKKE et samlet Q-Q plot for alle data, det skal gøres separat på hver stikprøve (eller på residualerne)

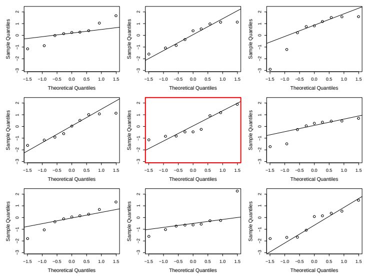
Eksempel – Sammenlign med simulerede data, A

```r
# Multiple (simulated) Q-Q plots and sample A
require(MESS)
fitA <- lm(xA ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fitA, FUN = qqnorm.wally, main = "")
```



Eksempel – Sammenlign med simulerede data, B

```r
# Multiple (simulated) Q-Q plots and sample B
fitB <- lm(xB ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fitB, FUN = qqnorm.wally, main = "")
```



Overview

Forsøgsplanlægning med krav til præcisionen

Vi har lært at $100 \cdot (1-\alpha)\%$ CI ved én stikprøve er givet ved: $\bar{x} \pm t_{1-\alpha/2} \cdot s/\sqrt{n}$.

Fejlmargenen (margin of error, ME) er defineret som

$$t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Method 3.63: The one-sample CI sample size formula:

Hvis σ er kendt, eller vurderet til at være en bestemt værdi, så kan vi beregne den stikprøvestørrelse, som kræves for at opnå en given fejlmargin, ME, med sandsynlighed $1-\alpha$.

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

Eksempel, højdedataene igen

Middelværdi og standardafvigelse for stikprøven:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimater for populationens middelværdi og -standardafvigelse:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

Hvis vi ønsker $ME = 3$ cm med 95% konfidens, hvor stort skal n så være?

$$n = \left(\frac{1.96 \cdot 12.21}{3}\right)^2 = 63.64 \approx 64$$

Forsøgsplanlægning, styrke

Hvad er styrken af et fremtidigt eksperiment:

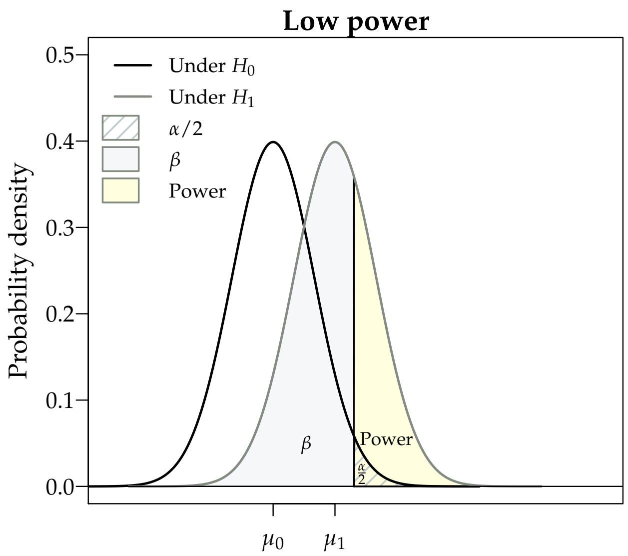
- Sandsynligheden for at detektere en (påstået) effekt.
- $P(H_0$ afvises$)$ når H_1 er sand.
- Sandsynligheden for en korrekt afvisning of H_0.
- MEN: En nulhypotese kan være forkert på mange måder!
- I praksis: brug en scenarie-baseret tilgang
  - E.g. "Hvis $\mu = 86$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"
  - E.g. "Hvis $\mu = 84$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"
  - etc.

Forsøgsplanlægning, styrke

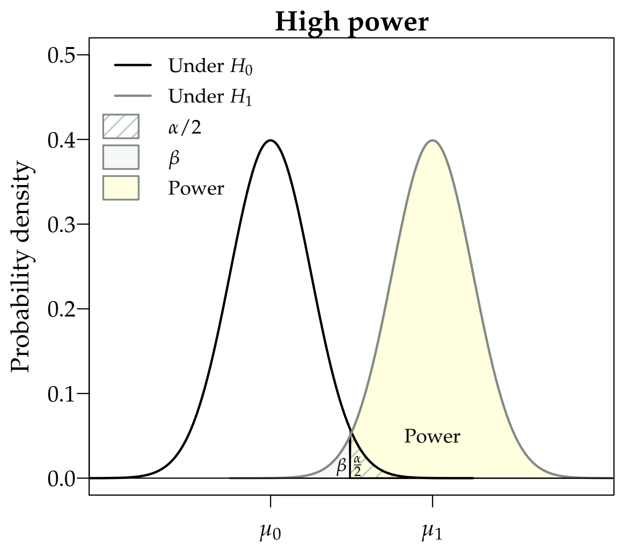
Hvis vi kender (eller antager) fire ud af de fem følgende størrelser, så kan vi finde den manglende:

- Stikprøvestørrelsen (sample size) n.
- Signifikansniveauet α, som vi tester på.
- Forskellen i middelværdi (effekt-størrelsen) $\mu_0 - \mu_1$.
- Populationsstandardafvigelsen, σ.
- Styrken (power), $(1-\beta)$.

Eksempel med lav styrke



Eksempel med høj styrke



Planning, sample size n

Det store spørgsmål: Hvor stort skal n være?

Vi skal have nok observationer til at kunne detektere en relevant effekt med høj styrke $1-\beta$ (typisk mindst 80%):

Metode 3.65: The one-sample sample size formula:

For one-sample t-testet med givne α, β and σ, vælg:

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

Her er $\mu_0 - \mu_1$ den forskel i middelværdi, som vi ønsker at måle. $z_{1-\beta}$ og $z_{1-\alpha/2}$ er fraktiler i standard normalfordelingen.

Eksempel – styrken hvis $n = 40$

```
# Power calculation (one-sample)
power.t.test(n = 40, delta = 4, sd = 12.21, type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.5242
##     alternative = two.sided
```

Eksempel – stikprøvestørrelsen hvis styken $= 0.80$

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

##
##      One-sample t test power calculation
##
##               n = 75.08
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```


Styrke og stikprøvestørrelse – two-sample

Find styrken ved at detektere en forskel/effektstørrelse på 2 med $\sigma = 1$ for $n = 10$:

```
# Power calculation (two-sample)
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9882
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```


Power and sample size - two-sample

Find stikprøvestørrelsen hvis vi skal detektere en forskel/effektstørrelse på 2 med $\sigma = 1$ og styrke $= 0.9$:

```
# Sample size calculation (two-sample)
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 6.387
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```


Power and sample size - two-sample

Fund den effektstørrelse δ der svarer til $\sigma = 1$, $n = 10$ and power $= 0.9$:

```
## Detectable effect size (two-sample)
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 1.534
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Overview

Den poolede two-sample t-teststørrelse

Det *sammenvejede* (poolede) variansestimat (her antages $\sigma_1^2 = \sigma_2^2$)

Method 3.52

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

Den sammenvejede two-sample t-teststørrelse, Method 3.53

Betragter vi nulhypotesen om forskellen i middelværdi mellem to *uafhængige* stikprøver:

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

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

så er den sammenvejede two-sample t-teststørrelse givet ved:

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

Theorem 3.54: Fordelingen af den poolede teststørrelse

... er en t-fordeling:

Den sammenvejede two-sample teststørrelse set som en stokastisk variabel:

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

Under nulhypotesen og under antagelsen $\sigma_1^2 = \sigma_2^2$, så følger T en t-fordeling med $n_1 + n_2 - 2$ frihedsgrader, hvis de to populationer er normalfordelte.

Vi bruger altid "Welch"-versionen

Nogenlunde (idiot)sikkert at bruge Welch-versionen altid:

- Hvis $s_1^2 = s_2^2$, så er Welch og den poolede t-teststørrelse are the same. Hvis det er tilfældet, så foretrækker vi ikke den poolede version, da antagelsen om ens varianser er højst tvivlsom.
- Kun hvis de to varianser er meget forskellige, kan det ske at de to teststørrelser bliver meget forskellige. I dette tilfælde foretrækker vi ikke poolede version, da antagelsen om ens varianser er højst tvivlsom.
- I tilfælde med lille stikprøvestørrelse i mindst en af grupperne, kan den poolede version give en højere styrke (under antagelse om ens varianser). I disse tilfælde er Welch-versionen en "forsigtig" tilgang.

Overview