



## To stikprøver, hypotesetest og t-teststørrelser

Anders Tolver  
Institut for Matematiske Fag

Statistisk Dataanalyse 1, Kursusuge 3, onsdag  
Dias 1/40



## Dagens program

### Formiddag

- Tostikprøve-problemet: parrede vs. uparrede test
- Hypotesetest: introduceret via t-test
- Eksempel: t-test i forbindelse med lineær regression
- Eksempel: t-test i forbindelse med ensidet ANOVA

### Videoer

- Gennemgang af Quiz 1 + Quiz 2 (på Absalon)
- Gennemgang af Quiz 3 (på Absalon sidst på ugen)
- (Muligvis) lidt kommentarer til slides fra mandag (20/9)
- Analyse af datasæt med gæt på antal punkter på en figur
  - Hypotesetest: Gætter man systematisk for højt/lavt?

Statistisk Dataanalyse 1, Kursusuge 3, onsdag  
Dias 2/40



## Overblik

Vi skal have „udfyldt“ følgende skema over modeller (rækker) og statistiske begreber (søjler):

	Intro	Model	Est.+SE	KI	Test	Kontrol	Præd.
En stikprøve	✓	✓	✓	✓	nu	✓	
Ensidet ANOVA	✓	✓	✓	✓	(nu)		
Lineær regr.	✓	(✓)	(✓)	(✓)	nu		
To stikprøver	nu	nu	nu	nu	nu		
Multipel regr.							
Tosidet ANOVA							

Statistisk Dataanalyse 1, Kursusuge 3, onsdag  
Dias 3/40



## Statistiske begreber

### Statistiske grundbegreber indtil videre:

- Population og stikprøve
- Gennemsnit, stikprøvespredning, median, kvartiler
- Statistisk model og parametre
- Estimer og standard error (SE) for estimer
- Konfidensinterval
- **Hypotesetest**

Statistisk Dataanalyse 1, Kursusuge 3, onsdag  
Dias 4/40



## To stikprøver



## Uparrede vs parrede stikprøver

**Data:**  $x_1, \dots, x_{n_1}$  og  $y_1, \dots, y_{n_2}$ . Den samme slags respons målt med både  $x$  og  $y$ , men under to forskellige "omstændigheder".

### Eksempel 1:

- 100 kvinders hhv. 42 mænds gæt på punktplot 1
- Interesseret i forskel på mænd og kvinder (om nogen)

### Eksempel 2:

- 142 studerendes gæt på hhv. punktplot 1 og punktplot 2
- Interesseret i forskel på gæt mellem de to punktplot

Der er en væsentlig forskel mellem de to situationer. Hvilken?

Opgave HS.16: Afgør hvilket set-up i tre forskellige situationer.



## To uafhængige stikprøver

### Egenskaber:

- **Alle** observationer kan antages at være uafhængige
- Man kan ændre rækkefølgen af  $x$ 'er og  $y$ 'er hver for sig uden at ændre datasættet
- Stikprøvestørrelserne kan være **forskellige**

Under antagelse af **ens spredninger**: Ensidede ANOVA med  $k = 2$

- Kan bruge `lm` som i ANOVA
- Alternativ: `t.test(x, y, var.equal=TRUE)`

Kan godt analysere data **uden at antage ens spredninger**, se afsnit 5.4. R: `t.test(x, y)`.

Se kommenteret eksempel i R-kode til i dag, `fore1210922.Rprog`



## To parrede stikprøver

### Egenskaber:

- Observationerne kommer i **par**. Parrene, men ikke enkeltobservationerne, kan antages at være uafhængige
- Man kan ikke ændre rækkefølgen af  $x$ 'er og  $y$ 'er hver for sig uden at ændre datasættet
- Stikprøver er nødvendigvis **lige store**

Analysere **forskellen** mellem  $x$  og  $y$  som **en enkelt stikprøve**.

- Kan bruge `lm(x-y ~ 1)`
- Alternativ: `t.test(x, y, paired=TRUE)` eller `t.test(x-y)`

Se kommenteret eksempel i R-kode til i dag, `fore1210922.Rprog`



## $t$ -test for en enkelt stikprøve og parrede stikprøver



## Eksempel 6.1: Hormonkoncentration

Et forsøg skal vise om et nyt foder ændrer konc. af et hormon.

- Ni dyr har fået foderet i en periode. Hormonkoncentrationen målt ved forsøgets start og slutning. Enhed:  $\mu\text{g/ml}$ .
- Spørgsmål: **Har foderet en effekt på hormonconc.?**

```
> hormData
  feed initial final
1    1     207   216
2    1     196   199
.
.
9    1     190   182
```

**Parrede data!**



## Foreløbig analyse af forskelle

Vi ser på **differenserne**, og betragter dem som en enkelt stikprøve:

$$y = \text{diff} = \text{final} - \text{initial}$$

Analyse:

- Statistisk model:  $y_1, \dots, y_n$  uafhængige, normalfordelte med middelværdi  $\mu$  og spredning  $\sigma$ .
- Fortolkning af  $\mu$ ? **Hvilken værdi er særligt interessant?**
- Konfidensinterval? Fortolkning?

Konfidensintervallet svarer til en vis grad på vores spørgsmål, men man plejer at lave et **hypotesetest** i stedet.



## Hypotese

Hvis foderet ikke har nogen effekt, så er der ikke systematisk forskel på "før og efter". Dette svarer til at  $\mu = 0$ .

Vil derfor teste **hypotesen** (nulhypotesen)

$$H_0 : \mu = 0$$

Hypotesen er en **restriktion** på den statistiske model.

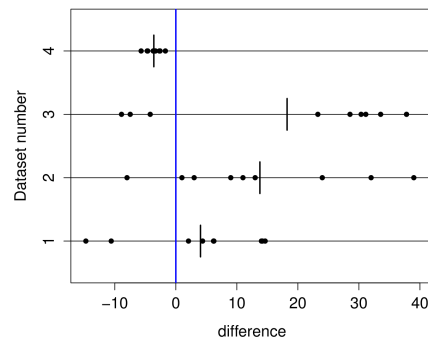
- Under modellen:  $y_i \sim N(\mu, \sigma^2)$ , uafhængige.
- Hvis  $H_0$  er sand:  $y_i \sim N(0, \sigma^2)$ , uafhængige.

NB: Nullet i " $H_0$ " henviser **ikke** til nullet i " $\mu = 0$ ". Nulhypotesen også havde været  $H_0 : \mu = 5$ , hvis dette var interessant.



## Hvad passer bedst/dårligst?

Fire datasæt med hver 9 differencer. Hvilke datasæt stemmer bedst/dårligst med hypotesen om, at middelværdien er 0?



- Sorte lodrette streger: stikprøvegennemsnit.
- Blå lodret streg: Hypoteseværdien (nul)



## Ideen i et hypotesetest

Hypotese  $H_0 : \mu = 0$

Vi har estimeret — “bedste gæt” —  $\hat{\mu} = \bar{y}$ . Alt andet lige:

- Hvis  $\hat{\mu} = \bar{y}$  ligger **langt fra nul**, tyder det på at  $H_0$  er falsk.
- Hvis  $\hat{\mu} = \bar{y}$  ligger **tæt på nul**, tyder det ikke på at  $H_0$  er falsk.

Men hvad er “langt fra” og hvad er “tæt på”?

- Værdien  $\hat{\mu} = 13.78$  alene er ikke nok!  
Hvis vi målte i  $\mu g/l$  i stedet ville vi have fået 0.01378 i stedet.  
Det lyder lille, men er jo helt den samme forskel.
- Skal tage højde for **variationen i data!**

Skyldes forskellen i stikprøven en **reel effekt** eller blot **tilfældigheder**?  
Hvad hvis vi gentog eksperimentet?



## Ideen i et hypotesetest

Er data i overensstemmelse med hypotesen?

- Data stemmer **godt** med hypotesen, hvis hypotesen gør det **sandsynligt** at en gentagelse af eksperimentet resulterer i observationer, der passer dårligere med  $H_0$  end dem vi har.
- Data stemmer **dårligt** med hypotesen, hvis hypotesen gør det **usandsynligt** at en gentagelse af eksperimentet resulterer i observationer, der passer dårligere med  $H_0$  end dem vi har.

Sandsynlighed for at en gentagelse passer dårligere kaldes **p-værdien**.

Troværdigheden af hypotesen måles vha. p-værdien.



## Teststørrelse

Skal altså beregne sandsynligheden for at en gentagelse passer dårligere med hypotesen end de givne data — hvis hyp. er sand.

Skal have en metode til at måle hvor godt/dårligt data passer med hypotesen. Vi skal bruge en **teststørrelse** (eng.: test statistic).

Teststørrelsen for en hypotese skal opfylde tre kriterier:

- Det er en **talværdi**, som kan beregnes ud fra data.
- Den skal være et (godt) **mål for hvor godt data stemmer med hypotesen**.

Skal kunne skelne om hypotesen passer godt til data eller ej.

- Under forudsætning af at hypotesen er sand, skal **teststørrelsens sandsynlighedsfordeling** kunne bestemmes.



## T-teststørrelsen for en enkelt stikprøve

Statistisk model:  $y_1, \dots, y_n \sim N(\mu, \sigma^2)$

Husk:

- $\hat{\mu} = \bar{y}$  er normalford. med middelværdi  $\mu$  og spredning  $\sigma/\sqrt{n}$ .
- Fra konstruktion fra konfidensinterval:

$$T = \frac{\hat{\mu} - \mu}{SE(\hat{\mu})} = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Hypotese,  $H_0 : \mu = 0$ . Hvis hypotesen er sand, kan vi erstatte  $\mu$  med 0:

$$T = \frac{\hat{\mu} - 0}{SE(\hat{\mu})} = \frac{\bar{y} - 0}{s/\sqrt{n}} \sim t_{n-1}$$

Opfylder  $T$  kriterierne?



## Kan $T$ bruges som teststørrelse?

De tre kriterier:

- Det er en **talværdi**, som kan beregnes ud fra data ✓
- Den skal være et godt mål for hvor godt data stemmer med hypotesen ✓  
Værdier tæt på 0 passer godt, værdier langt fra 0 passer skidt
- Under forudsætning af at hypotesen er sand, skal teststørrelsens sandsynlighedsfordeling kunne beregnes ✓  
**Hvis**  $H_0$  er sand vil  $T$  være  $t$ -fordelt med  $n - 1$  frihedsgrader.

Tilsammen: Vi kan nu beregne **ssh. for at få en  $T$ -værdi der passer dårligere med hypotesen end den vi fik fra vores data.**



## $p$ -værdi for eksemplet med hormonkoncentration

Vi fik  $\hat{\mu} = \bar{y} = 13.78$ ,  $s = 15.24$  og har  $n = 9$ . Dermed

$$SE(\hat{\mu}) = \frac{15.24}{\sqrt{9}} = 5.08, \quad T_{\text{obs}} = \frac{13.78 - 0}{5.08} = 2.71$$

**$p$ -værdien** er sandsynligheden for at få en værdi af  $T$  der ligger lige så langt eller længere væk fra nul end det vi fik. **Se figur!**

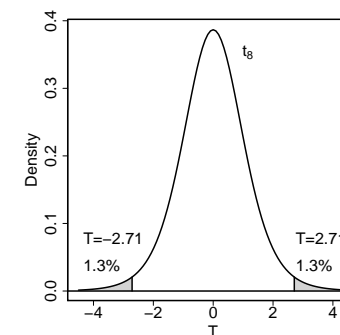
$$p = P(|T| \geq |T_{\text{obs}}|) = P(|T| \geq 2.71) = 2 \cdot P(T \geq 2.71) = 0.026,$$

Hvis  $H_0$  er sand er det altså ikke særligt sandsynligt at få en så stor værdi af  $T$  som vi fik  $\rightarrow H_0$  virker ikke troværdig  $\rightarrow H_0$  **afvises**.



## $p$ -værdi for eksemplet med hormonkoncentration

$T_{\text{obs}} = 2.71$  skal evalueres i  $t$ -fordelingen med 8 frihedsgrader.



###  $P(T > 2.71)$ . Skal ganges med 2 for at få  $p$ -værdien  
> 1-pt(2.71,df=8)  
[1] 0.01332905



## $p$ -værdi og konklusion på test

Forvirret...?

Så hold fast i følgende, som **altid** gælder:

- En (meget) **lille  $p$ -værdi** tyder (stærkt) på, at hypotesen er falsk, så vi **afviser** hypotesen
- En **moderat eller stor  $p$ -værdi** siger, at hypotesen stemmer godt med vores data, så vi **afviser ikke hypotesen**

Men hvor lille er lille?



## Konventionelle grænser

Fra gamle dage med stat. tabeller har man tre signifikansgrænser:

- \*\*\*  $p < 0.001$ . Signifikans på 0.1% niveau. Meget stærk evidens mod hypotesen.
- \*\*  $p < 0.01$ . Signifikans på 1% niveau. Temmelig stærk evidens mod hypotesen.
- \*  $p < 0.05$ . Signifikans på 5% niveau. Nogen evidens mod hypotesen.
- NS  $p > 0.05$ . Ikke signifikant (Not Significant). Ingen overbevisende evidens mod hypotesen.

Grænser bruges stadig, selvom de er temmelig arbitrære.

Evidensen mod hypotesen er så godt som den samme for en  $p$ -værdi på 5.1% som for 4.9%. **Angiv altid  $p$ -værdien.**



## Hormonkoncentration: Konklusion

Spørgsmål: **Har foderet en effekt på hormonkoncentrationen?**

- Hypotese,  $H_0 : \mu = 0$  hvor  $\mu$  er den forventede ændring for et tilfældigt dyr (populationsgennemsnittet).
- Vi har med rimelig sikkerhed påvist, at hypotesen ikke holder, og dermed påvist en effekt af foderet ( $p = 0.026$ ).
- Stigningen i hormonkoncentrationen estimeres til 13.78 med 95% konfidensinterval (2.06, 25.49).



## R: "Manuelt"

```
> library(isdals)
> data(hormone)
> hormData <- subset(hormone, feed=="1")
> hormData <- transform(hormData, dif = final-initial)

> mean(hormData$dif)
[1] 13.77778
> sd(hormData$dif)
[1] 15.23793
> 13.77778 / 15.23793 * sqrt(9)
[1] 2.71253
> 2*(1 - pt(2.71253, df=8))
[1] 0.02655391
```



## R: Med lm

```
> model <- lm(dif ~ 1, data=hormData)
> summary(model)
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.778      5.079   2.713  0.0266 *
```

---

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

```
Residual standard error: 15.24 on 8 degrees of freedom
```



## R: t.test

```
> t.test(hormData$dif)
```

One Sample t-test

data: hormData\$dif

t = 2.7125, df = 8, p-value = 0.02655

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

2.06487 25.49069

sample estimates:

mean of x

13.77778



## Konfidensinterval og hypotesetest

I eksemplet gav konfidensintervallet og hypotesetestet samme konklusion:

- Nul ligger ikke i **95%**-konfidensintervallet
- Vi afviser  $H_0$  med en  $p$ -værdi mindre end **5%**

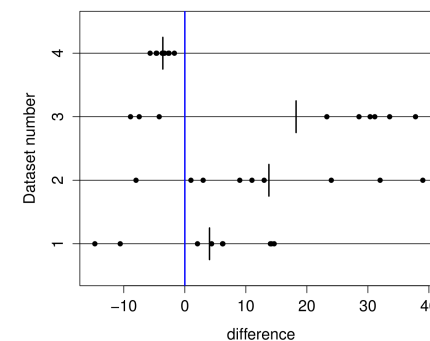
Sådan er det altid for  $t$ -tests:

**0 er ikke i 95%-konfidensinterval hvis og kun hvis hypotesen  $H_0 : \mu = 0$  kan afvises på 5% signifikansniveau.**



## Hvad passer bedst/dårligst?

Fire datasæt med hver 9 differencer. Hvilke stemmer bedst/dårligst med hypotesen om at middelværdien er 0?



- **Blå lodret streg:** Hypoteseværdien (nul)
- $p$ -værdier:  $p = 0.00002$ ,  $p = 0.022$ ,  $p = 0.027$ ,  $p = 0.28$



## $t$ -test: intuition og eksempler



## Hvad er en statistisk hypotese?

Husk vores *opskrift*

- **Data:** hvor mange variable og hvilke typer?
- Fører til **statistisk model** for variationen i vores stikprøve
- Populationsparametre fra modellen har vores primære interesse
- Hidtil: fokus på estimater og konfidensintervaller

En **statistisk hypotese** er et spørgsmål / en antagelse omkring værdien af nogle af populationsparametrene.

Hvordan afgøres om data understøtter hypotesen eller ej?

Vi *måler* om data understøtter hypotesen vha. en **teststørrelse**.



## Generel form for $t$ -teststørrelser

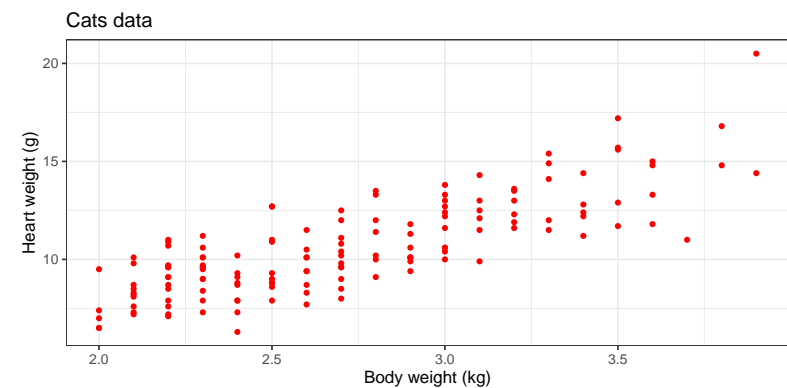
$T$ -teststørrelser har **altid** formen

$$T_{\text{obs}} = \frac{\text{estimat} - \text{hypoteseværdi}}{\text{SE}(\text{estimat})}$$

og skal vurderes i "den relevante"  $t$ -fordeling.



## Test i lineær regression



- Er der faktisk en sammenhæng mellem kropsvægt og hjertevægt?





## Test i lineær regression

**Data:** Par  $(x_1, y_1), \dots, (x_n, y_n)$

**Statistisk model:**

- $y_1, \dots, y_n$  uafhængige
- $y_i$  normalfordelt med middelværdi  $\alpha + \beta x_i$  og spredning  $\sigma$ .

Hypotesen er at  $x$  ikke har nogen effekt på  $y$ , at der ikke er nogen sammenhæng mellem de to variable.

Hvordan kan det udtrykkes vha.  $\alpha$  og/eller  $\beta$ ?



## Test i lineær regression

Den relevante hypotese er ofte (men ikke altid)  $H_0 : \beta = 0$ :

$$T_{\text{obs}} = \frac{\hat{\beta} - 0}{\text{SE}(\hat{\beta})}$$

Skal vurderes i  $t$ -fordelingen med  $\text{df} = n - 2$ .

**Eksempel:** Der er data fra  $n = 144$  katte:

$$T_{\text{obs}} = \frac{\hat{\beta} - 0}{\text{SE}(\hat{\beta})} = \frac{4.0341}{0.2503} = 16.12$$

der skal vurderes i  $t_{142}$ . Dette giver en  **$p$ -værdi**  $< 2 \cdot 10^{-16}$ . **Tegn!**

**Konklusion:** Der er meget stærk evidens mod hypotesen. Der **er** sammenhæng mellem kattes kropsvægt og hjertevægt.



## R: sammenhæng mellem kropsvægt og hjertevægt for katte

```
summary(lm(Hwt ~ Bwt, data = cats))

##
## Call:
## lm(formula = Hwt ~ Bwt, data = cats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923   -0.515   0.607
##              Bwt       4.0341     0.2503   16.119 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```



## Lineær regression: Test for $H_0 : \beta = \beta_0$

Antag at en **teori** siger at 1 kg ekstra på kropsvægten i gennemsnit (i populationen af katte) fører til 4 g ekstra hjertevægt.

Dette svarer til **hypotesen**  $H_0 : \beta = 4$ .

Mere generelt: Hypotese  $H_0 : \beta = \beta_0$  for en **præ-specificeret værdi**  $\beta_0$  (kendt inden vi indsamlede data).

Teststørrelse

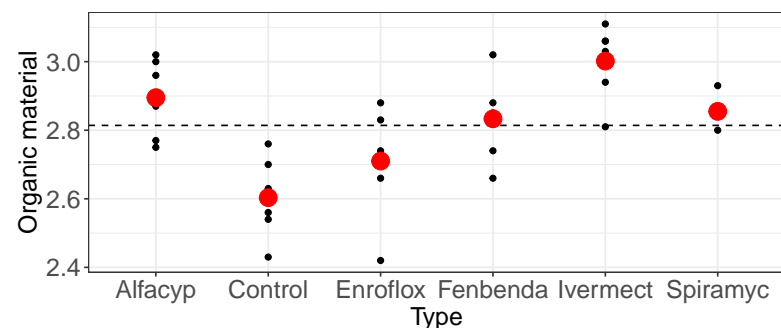
$$T_{\text{obs}} = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})} = \frac{4.0341 - 4}{0.2503} = 0.136.$$

Sammenligning med  $t_{142}$  giver  $p$ -værdien 0.89. **Tegn!**

**Konklusion:** Data er ikke i modstrid med hypotesen.



## Sammenligning af to grupper i ensidet ANOVA



- Hæmmer Fenbendazole nedbrydningen af organisk materiale?
- Laver testet i modellen for alle data; ikke som to stikprøver



## Statistisk model

**Data:**  $y_1, \dots, y_n$  fra  $k$  grupper med  $n_j$  obs. i gruppe  $j$ .

### Statistisk model:

- $y_1, \dots, y_n$  uafhængige
- $y_i$  normalfordelte med middelværdi  $\alpha_{g(i)}$  og spredning  $\sigma$

**Hypotese:**  $H_0 : \alpha_{\text{Con}} = \alpha_{\text{Fen}}$

$t$ -teststørrelse

$$T_{\text{obs}} = \frac{\hat{\alpha}_{\text{Fen}} - \hat{\alpha}_{\text{Con}}}{\text{SE}(\hat{\alpha}_{\text{Fen}} - \hat{\alpha}_{\text{Con}})} = \frac{0.230}{0.070} = 3.27$$

Skal evalueres i  $t_{28}$ . Dette giver  **$p$ -værdien** 0.0028.

**Konklusion:** Vi har med stor sikkerhed påvist at Fenbendazole hæmmer nedbrydningen.



## R

```
> antibio$myType <- relevel(antibio$type, ref="Control")
> model3 <- lm(org ~ myType, data=antibio)
> summary(model3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.60333	0.04970	52.379	< 2e-16	***
myTypeAlfacyc	0.29167	0.07029	4.150	0.000281	***
myTypeEnroflox	0.10667	0.07029	1.518	0.140338	
myTypeFenbenda	0.23000	0.07029	3.272	0.002834	**
myTypeIvermect	0.39833	0.07029	5.667	4.5e-06	***
myTypeSpiramyc	0.25167	0.07858	3.202	0.003384	**

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

Residual standard error: 0.1217 on 28 degrees of freedom



## Opsummering

$t$ -test for hypotesen  $H_0 : \mu = 0$  i en enkelt stikprøve udføres sådan:

- Hypotese,  $H_0 : \mu = 0$
- Beregn  $t$ -teststørrelsen,

$$T_{\text{obs}} = \frac{\hat{\mu} - 0}{\text{SE}(\hat{\mu})} = \frac{\sqrt{n}(\bar{y} - 0)}{s}$$

- Sammenlign teststørrelsen med  $t$ -fordelingen med  $n - 1$  frihedsgrader og beregn  $p$ -værdien
- Konkluder

