

Elementær statistik

Lektion 4

Peter Tibert Stoltze
stat@peterstoltze.dk

1. marts 2010

Dagens program

- ▶ Kapitel 4: Sandsynlighed og statistiske modeller
 - ▶ Opsamling
 - ▶ Opgave 4
- ▶ Kapitel 5: Estimation
 - ▶ Gennemgang (fortsat)
 - ▶ Opgave 6 og 7
- ▶ Kapitel 6: Signifikanstestning
 - ▶ Gennemgang
- ▶ Kapitel 7: Forskelle mellem centraltendenser
 - ▶ Gennemgang
 - ▶ Opgave 8 og 9

Del I

Kapitel 4: Sandsynlighed og statistiske modeller

Oversigt

Opsamling

Opgave 4

Opsamling

- ▶ **Normalfordelingen** er en ofte benyttet **model** for delvist observerede populationer, idet fordelingsparametre kan estimeres fra en stikprøve
- ▶ Normalfordelingen har to **parametre**, middelværdi μ og spredning σ , og vi skriver $N(\mu, \sigma^2)$
- ▶ **Standardnormalfordelingen** $N(0, 1)$ kan benyttes til beregninger i andre normalfordelinger via en z-værdi

$$z = \frac{x - \mu}{\sigma}$$

- ▶ Således beregnes **sandsynligheden** $P(X < x)$, hvor $X \sim N(\mu_0, \sigma_0^2)$ ved

$$p = \Phi\left(\frac{x - \mu_0}{\sigma_0}\right)$$

hvor Φ er **fordelingsfunktionen** for standardnormalfordelingen.

Opgave 2

Gennemgang af opgave 4...

Del II

Kapitel 5: Estimation

Oversigt

Opsamling

Konfidensintervaller

Opgaver

Estimation og konfidensintervaller

- ▶ Stikprøvegennemsnittet er et **estimat** for middelværdien
- ▶ Estimer har en **fordeling**
- ▶ Stikprøvegennemsnittet er et **centralt estimat** for middelværdien
- ▶ Spredningen på middelværdien kaldes **standardfejlen** og beregnes som

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- ▶ Ved at afskære de nederste og øverste $\alpha/2$ procent af fordelingen kan vi konstruere et $1 - \alpha$ procent **konfidensinterval** for estimatet

Beregning af konfidensinterval

- ▶ Følger estimatet en normalfordeling kan vi lave konfidensinterval ved at beregne z-værdier og bruge Φ
- ▶ I eksemplet med vokalvarighed er $n = 40$, $\bar{x} = 208,9$ og $s = 9,79$, så standardfejlen er

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{9,79}{\sqrt{40}} = 1,548$$

- ▶ Vi løser en ligning for at finde nedre grænse:

$$\begin{aligned}\Phi\left(\frac{x_{nedre} - 208,9}{1,548}\right) &= 0,025 \\ \Rightarrow \frac{x_{nedre} - 208,9}{1,548} &= \Phi^{-1}(0,025) = -1,96 \\ \Rightarrow x_{nedre} &= 208,9 - 1,96 \cdot 1,548 = 205,9\end{aligned}$$

Beregning af konfidensinterval

- ▶ Vi løser en anden ligning for at finde øvre grænse:

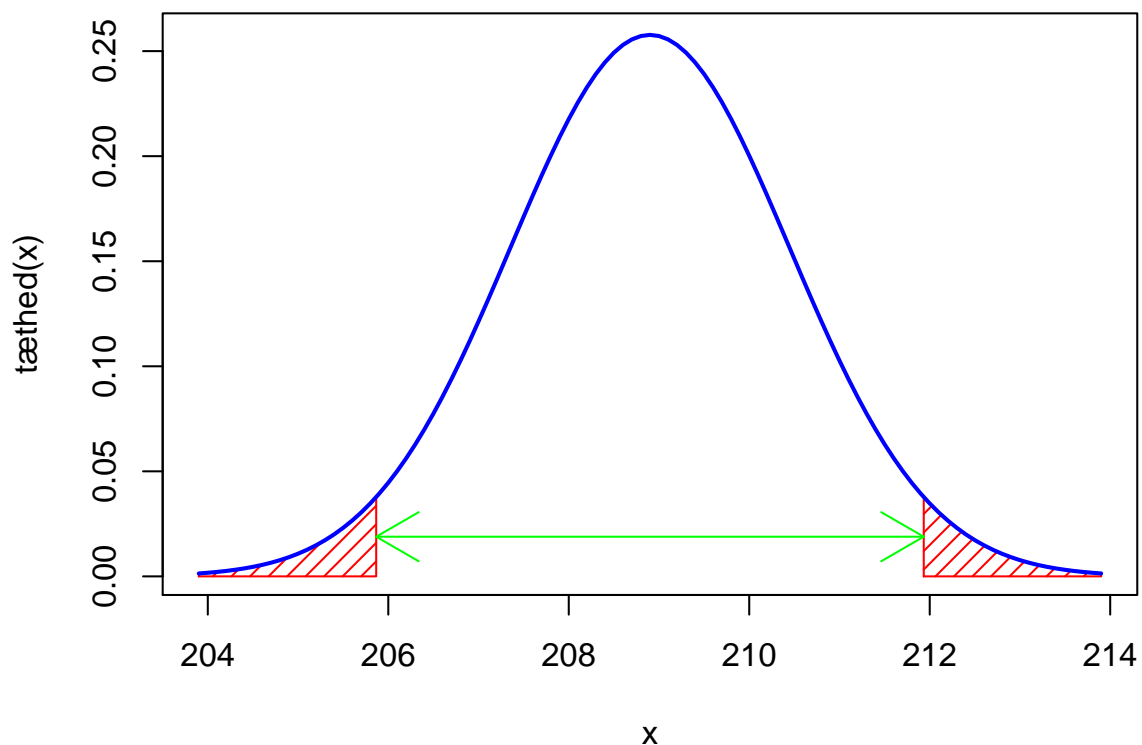
$$\begin{aligned}\Phi\left(\frac{x_{\text{øvre}} - 208,9}{1,548}\right) &= 0,975 \\ \Rightarrow \frac{x_{\text{øvre}} - 208,9}{1,548} &= \Phi^{-1}(0,975) = 1,96 \\ \Rightarrow x_{\text{øvre}} &= 208,9 + 1,96 \cdot 1,548 = 211,9\end{aligned}$$

- ▶ Et 95% konfidensinterval for middelværdien er altså

$$208,9 \pm 1,96 \cdot 1,548 = [205,9; 211,9]$$

- ▶ På baggrund af stikprøven kan vi altså sige, at med en sikkerhed på 95% ligger den sande middelværdi i intervallet fra 205,9 til 211,9

Beregning af konfidensinterval



Beregning af nødvendigt n

- ▶ Bredden af det beregnede konfidensinterval er på $211,9 - 205,9 = 6,0$ ms
- ▶ Hvor stor skal n være for at bredden af konfidensintervallet er højst 3?
- ▶ Intervallet beregnes som $\bar{x} \pm u_{1-\alpha/2} \frac{s}{\sqrt{n}}$, dvs. bredden af intervallet er $L = 2u_{1-\alpha/2} \frac{s}{\sqrt{n}}$
- ▶ For vilkårligt L_0 klares dette ved at løse ligningen

$$2u_{1-\alpha/2} \frac{s}{\sqrt{n}} < L_0 \Rightarrow n > \left(\frac{2u_{1-\alpha/2}s}{L_0} \right)^2$$

- ▶ I vores tilfælde findes

$$n > \left(\frac{2 \cdot 1,96 \cdot 9,79}{3} \right)^2 = 163,64$$

så n skal altså mindst være 164

Andre konfidensgrader $(1 - \alpha)$

- ▶ Vi kan også beregne konfidensintervaller med andre konfidensgrader $1 - \alpha$ efter opskriften

$$\bar{x} \pm u_{1-\alpha/2} \cdot s_{\bar{x}}$$

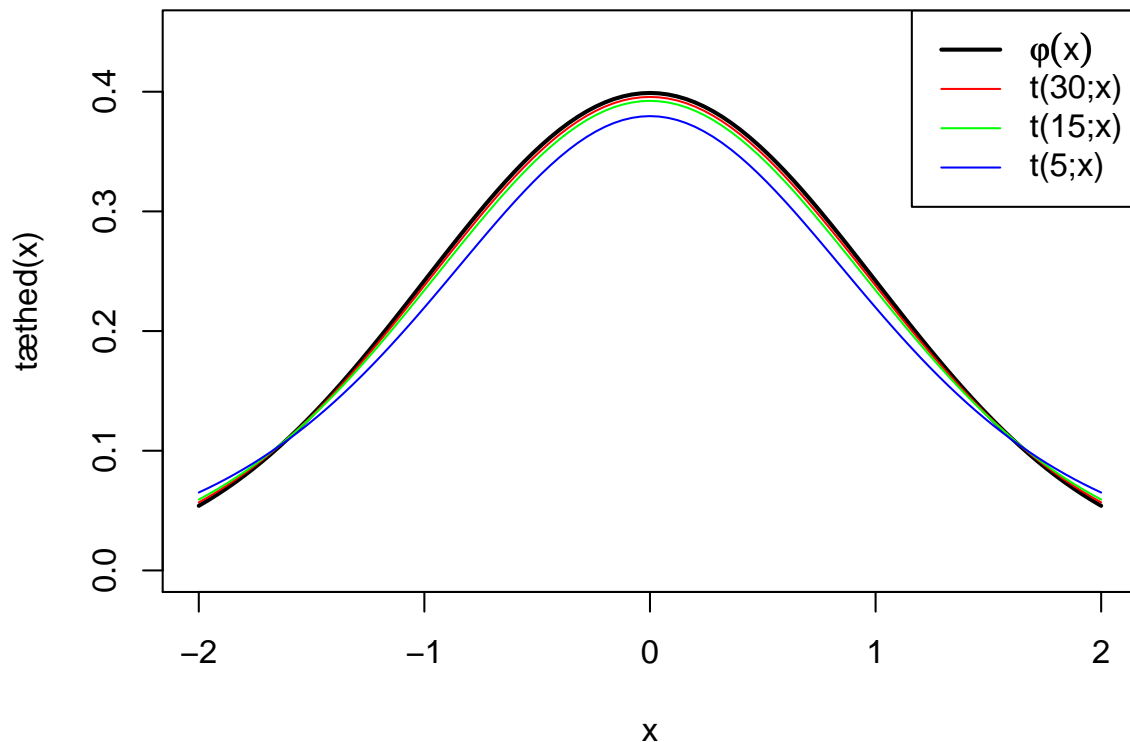
hvor $u_{1-\alpha/2}$ er en fraktil fra normalfordelingen

- ▶ Hvis $n < 30$ er konfidensintervaller baseret på normalfordelingen for optimistiske
- ▶ Vi benytter stadig samme opskrift, men nu bruger vi en fraktil fra en t -fordeling med $n - 1$ frihedsgrader i stedet for:

$$\bar{x} \pm t_{n-1;1-\alpha/2} \cdot s_{\bar{x}}$$

- ▶ Selvom $n > 30$ er dette den rigtige måde at regne på, såfremt spredningen s er estimeret

t-fordelingen



Opgave 6

$$x = \{4, 2, 5, 3, 4, 3, 5, 3, 6, 4, 4\}$$

a) $n = 11$, $\sum x = 43$, $\sum x^2 = 181$

$$\bar{x} = \frac{\sum x}{n} = \frac{43}{11} = 3,91$$

$$s = \sqrt{\frac{S_{AK}}{n-1}} = \sqrt{\frac{181 - \frac{1}{11}(43^2)}{11-1}} = 1,14$$

b) Generelt udtryk for $1-\alpha$ interval:

$$\bar{x} \pm t(n-1, 1-\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}$$

Nødvendige t-fraktiler

$1-\alpha$	α	$1-\frac{\alpha}{2}$	$t(10, 1-\frac{\alpha}{2})$
0,90	0,10	0,95	1,812
0,95	0,05	0,975	2,228
0,99	0,01	0,995	3,169

← Aflest i Tabel C (p.122) under t_{α} -højt signifikansniveau og $\alpha = \{0,10; 0,05; 0,01\}$

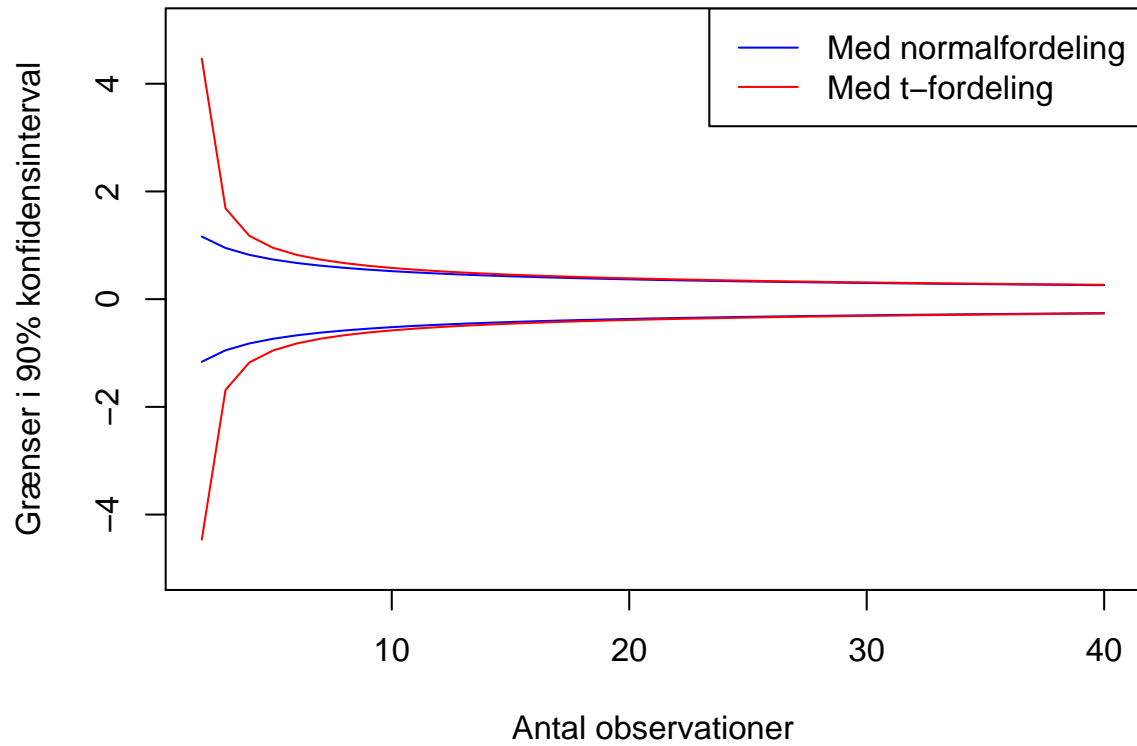
$$90\% \text{ C.I.: } 3,91 \pm 1,812 \cdot \frac{1,14}{\sqrt{11}} = [3,29; 4,53]$$

$$95\% \text{ C.I.: } 3,91 \pm 2,228 \cdot \frac{1,14}{\sqrt{11}} = [3,15; 4,67]$$

$$99\% \text{ C.I.: } 3,91 \pm 3,169 \cdot \frac{1,14}{\sqrt{11}} = [2,82; 5,00]$$

c) Konfidensinterval: Det område, hvor populationens sande middelværdi (med en vis sandsynlighed) må forventes at befinde sig, på et stikprøven...

Opgave 7



Del III

Kapitel 6: Signifikanstestning

Oversigt

Indledning

Hypoteser og virkelighed

Valg af testtype

Retningsbestemt alternativ

Indledning

- ▶ Generalisering af forskelle mellem stikprøver
- ▶ Signifikanstestning handler ofte om, hvor sandsynligt det er, at en observeret forskel skyldes en tilfældighed
- ▶ Eksempel: I en læsetest scorer drengene i gennemsnit 51,9 mens pigerne scorer 59,8
- ▶ Men kan vi herfra udlede at pigerne generelt laver højere score end drengene?

Opstilling af hypoteser

- ▶ Ved at se på data opstiller vi en relevant hypotese, som vi vil teste
- ▶ Hypotesen er todelt og består af en nulhypotese og en alternativ hypotese
- ▶ Eksempel: Sammenligning af to middelværdier

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Hypoteser og virkelighed

- ▶ Vores konklusion er korrekt hvis
 - ▶ H_0 er sand, og vi ikke forkaster den
 - ▶ H_0 er falsk, og vi forkaster den
- ▶ Vores konklusion er forkert hvis
 - ▶ H_0 er sand, men vi forkaster den (Type I-fejl)
 - ▶ H_0 er falsk, men vi ikke forkaster den (Type II-fejl)
- ▶ Vi kan altså drage forkerte (og korrekte) konklusioner på to principielt forskellige måder:

Nulhypotese	Accepteres	Forkastes
Sand	Korrekt konklusion	Type I-fejl (α)
Falsk	Type II-fejl (β)	Korrekt konklusion

Strategi ved test

- ▶ Vi starter med at antage, at H_0 er sand
- ▶ Under H_0 beregnes sandsynligheden p (givet alternativet) for en forskel mellem μ_1 og μ_2 svarende til det observerede (eller større)
- ▶ Hvis den observerede forskel er usandsynlig (for eksempel $p < 5\%$) forkaster vi H_0 og antager i stedet H_1
- ▶ Hvis den observerede forskel ikke er specielt usandsynlig ($p > 5\%$) forkaster vi ikke H_0
- ▶ Bemærk, at vi ved at undlade at forkaste H_0 formelt set ikke har bevist at, den faktisk er korrekt. . .

Valg af testtype

- ▶ Korrelerede og ukorrelerede data
- ▶ Korrelerede data kan parres
 - ▶ Læsescore før og efter specialundervisning for samme person
 - ▶ Patient før og efter behandling
- ▶ Ukorrelerede data kan ikke parres
 - ▶ Læsescore for piger og drenge i en klasse
 - ▶ Patienter og kontrolgruppe

Valg af testtype

- ▶ Parametriske tests forudsætter ofte, at data kan beskrives ved normalfordeling (kan data ikke det benyttes ikke-parametriske alternativer)
- ▶ Er responset målt på ordinalskala eller nominalskala skal der benyttes ikke-parametriske tests
- ▶ I nogle tilfælde kan parametriske test på ordinale data godt forsvares. . .

Énkelt- eller dobbeltsidet alternativ

- ▶ Uden at have set på data kan man formulere et dobbeltsidet alternativ:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- ▶ Et meningsfyldt enkeltsidet (retningsbestemt) alternativ kræver at data er undersøgt først:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

- ▶ Generelt giver enkeltsidede tests de stærkeste konklusioner

Eksempler med læsescores

- ▶ Det første hypotesesæt er

$$H_0 : \mu_{piger} = \mu_{dreng}$$

$$H_1 : \mu_{piger} \neq \mu_{dreng}$$

- ▶ Men vi ved at $\bar{x}_{dreng} = 51,9$ og $\bar{x}_{piger} = 59,8$ så vi kan skærpe vores hypotese:

$$H_0 : \mu_{piger} = \mu_{dreng}$$

$$H_1 : \mu_{piger} > \mu_{dreng}$$

- ▶ Af samme årsag er følgende hypotese ikke relevant:

$$H_0 : \mu_{piger} = \mu_{dreng}$$

$$H_1 : \mu_{piger} < \mu_{dreng}$$

Del IV

Kapitel 7: Forskelle mellem centraltendenser

Oversigt

Indledning

Parametriske tests for ukorrelerede data

Parametriske tests for korrelerede data

Opgaver

Indledning

1. z-test for ukorrelerede data
2. t -test for ukorrelerede data med ens varianser
3. t -test for ukorrelerede data med uens varianser
4. z-test for korrelerede data
5. t -test for korrelerede data
6. Mann-Whitney U -test for ukorrelerede data
7. Wilcoxon's rangtest for korrelerede data

Parametriske tests

- ▶ Vi starter med at opstille nulhypotesen H_0 og en relevant alternativhypotese H_1
- ▶ Vi beregner en **teststørrelse** med kendt fordeling på baggrund af data
- ▶ Ved opslag i tabel omsættes teststørrelsen til en **signifikanssandsynlighed** kaldet p
- ▶ For n mindre end 30 foretrækkes t -fordelingen, og ellers z -fordelingen
- ▶ På baggrund af p konkluderes det, om H_0 kan antages eller bør forkastes

Parametriske tests for ukorrelerede data

- ▶ Disse tests er aktuelle når data er
 - ▶ målt på en ratio- eller intervallskala
 - ▶ ukorrelerede, dvs. hænger ikke naturligt sammen i par
- ▶ Når $n \geq 30$ kan vi anvende et z -test
- ▶ Når $n < 30$ anvender vi et t -test, idet vi først skal teste om der kan antages ens varians

1. z-test for ukorrelerede data

- ▶ I eksemplet i Tabel 7.1 beregnes først spredningen på forskellen mellem gennemsnittene til 4,636
- ▶ Dernæst beregnes teststørrelsen z til 1,711 og så bestemmes p idet der benyttes enkeltsidet alternativ

$$H_0 : \mu_{piger} = \mu_{dreng}$$

$$H_1 : \mu_{piger} > \mu_{dreng}$$

- ▶ Af Tabel A finder vi signifikanssandsynligheden
 $p = P(z > 1,71) = 1 - \Phi(1,71) = 1 - 0,956 = 0,044$

Præ 2 og 3: F -test for ens varians

- ▶ Eksempel i Tabel 7.2
- ▶ Starter med at beregne stikprøvevarianserne $s_1^2 = 3,0$ og $s_2^2 = 2,2$ idet $n_1 = 4$ og $n_2 = 5$
- ▶ Dernæst beregnes F som

$$F = \frac{s_{max}^2}{s_{min}^2} = \frac{3,0}{2,2} = 1,36$$

der følger en F -fordelingen med $4-1 = 3$ frihedsgrader i tæller og $5-1 = 4$ frihedsgrader i nævner

- ▶ Vi finder kritisk værdi for $\alpha = 0,05$ til 6,59 hvilket betyder at $p > 0,05$ og at H_0 ($s_1^2 = s_2^2$) derfor ikke kan forkastes

2. t -test for ukorrelerede data med ens varians

- ▶ Samme princip som i z -testet men nu beregnes teststørrelsen t som

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

hvor

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

- ▶ t -værdien vurderes i en t -fordeling med antal frihedsgrader svarende til $n_1 + n_2 - 2$

3. t -test for ukorrelerede data med uens varians

- ▶ Nu beregnes teststørrelsen t som

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

- ▶ t -værdien vurderes i en t -fordeling med antal frihedsgrader svarende til

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (n_1 - 1)(1 - c)^2}$$

hvor

$$c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$$

Parametriske tests for korrelerede data

- ▶ Disse tests er aktuelle når data er
 - ▶ målt på en ratio- eller intervallskala
 - ▶ korrelerede, dvs. hænger naturligt sammen i par
- ▶ Når $n \geq 30$ kan vi anvende et z -test
- ▶ Når $n < 30$ anvender vi et t -test
- ▶ I begge tilfælde regnes på differenserne $d_i = x_i - y_i$

4. z -test for korrelerede data

- ▶ Først bestemmes spredningen på differenserne som

$$s_d^2 = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

- ▶ Herefter beregnes teststørrelsen z som

$$z = \frac{\bar{d}}{s_d / \sqrt{n}}$$

- ▶ Signifikanssandsynligheden p bestemmes ved opslag i Tabel A

5. t -test for korrelerede data

- ▶ Når $n < 30$ benyttes et t -test i stedet for z -testet
- ▶ Teststørrelsen t beregnes præcist som ved z -testet, men p bestemmes ved opslag i en t -fordeling med $n - 1$ frihedsgrader

Opgaver

- ▶ Opgave 8
- ▶ Opgave 9