

Kapitel 4

Sandsynlighed og statistiske modeller

Peter Tibert Stoltze
stat@peterstoltze.dk

Elementær statistik
F2011

1 / 22

Generalisering fra stikprøve til population

- ▶ Idé: Opstil en model for populationen og estimér modellens parametre på baggrund af stikprøven
- ▶ Kontrollér at stikprøven ikke er i modstrid med modellen
- ▶ Eksempel: 95% konfidensinterval for middelværdien i en normalfordeling

2 / 22

Binomialfordelingen - uformelt

- ▶ Lyttetest: En person har i tre ud af tre sætninger korrekt hørt forskel på **bas** og **pas** - kan det være tilfældigt?
- ▶ Vi gentager et eksperiment tre gange, hvor der hver gang er 50% sandsynlighed for at få succes ved en tilfældighed (fx. få krone)
- ▶ Hvad er sandsynligheden for at få krone tre gange i træk?
- ▶ Og hvorfor er det interessant?

3 / 22

Uformelt. . . *fortsat*

- ▶ Der er otte mulige udfald ved tre kast: KKK, KKP, KPK, PKK, KPP, PKP, PPK, PPP
- ▶ Alle otte udfald er lige sandsynlige og netop ét udfald svarer til tre gange krone
- ▶ Laplaces lov: Sandsynlighed er antal gunstige divideret med antal mulige
- ▶ Sandsynligheden for netop tre gange krone er således $1/8 = 0,125 = 12,5\%$

4 / 22

- ▶ Tilbage til lyttetest: Der er altså en sandsynlighed (risiko) på 12,5% for, at personen ikke kan høre forskel på bas og pas selvom der blev svaret rigtigt i 3 ud af 3 tilfælde.
- ▶ Er dette acceptabelt og hvis ikke: Hvordan kan man så lave eksperimentet bedre?

Binomialfordelingen — formelt

- ▶ n Bernoulli-forsøg med sandsynligheden p for sandt (og følgelig sandsynligheden $1 - p$ for falsk)
- ▶ Punktsandsynligheder er givet ved

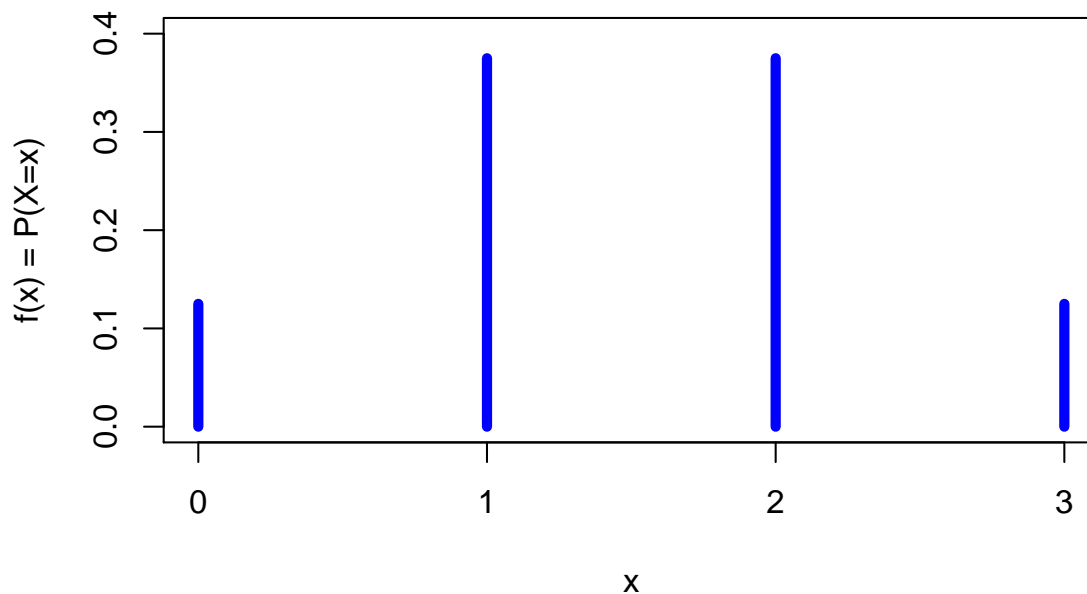
$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$$

hvor $K(n,x)$ er binomialkoefficienten

$$\binom{n}{x} = \frac{n \cdot (n-1) \cdot \dots \cdot (n-x+1)}{x \cdot (x-1) \cdot \dots \cdot 1} = \frac{n!}{x!(n-x)!}$$

Binomialfordelingen — formelt og grafisk

$X \sim \text{Bin}(3, 0.5)$



7 / 22

Opgave 3

Opgave 3

199L
07.02.09

Sandsynligheden for at svare korrekt
n gange i træk er

$$p = 0.5^n$$

Vi skal løse uligheden

$$p = 0.5^n < 0.05$$

Vi bruger regnereglen $\log(y^x) = x \cdot \log(y)$:

$$0.5^n < 0.05$$

$$\Downarrow n \cdot \log(0.5) < \log(0.05)$$

$$\Downarrow n > \frac{\log(0.05)}{\log(0.5)} = 4,32$$

Bemærk at ulighedstegnet bliver vendt
da $\log(0.5) < 0$

Vi skal altså mindst spørge fem
for at sandsynligheden bliver
mindre end 5 pct.

Er kravet 0,1 pct. finder vi

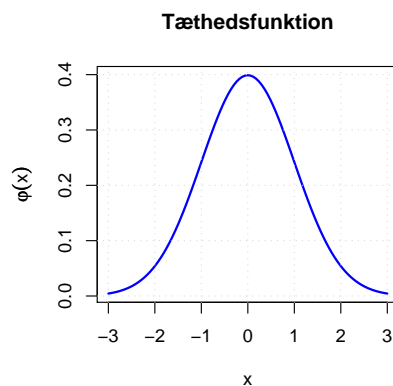
$$n > \frac{\log(0.001)}{\log(0.5)} = 9,97$$

Dvs. vi skal mindst spørge ti

8 / 22

Normalfordelingen

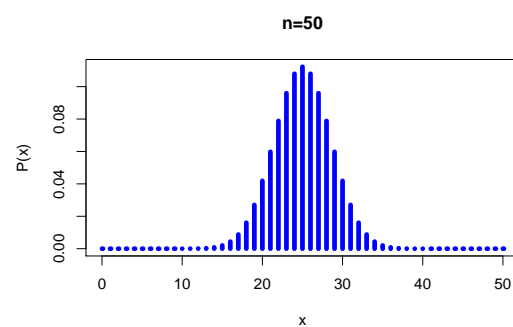
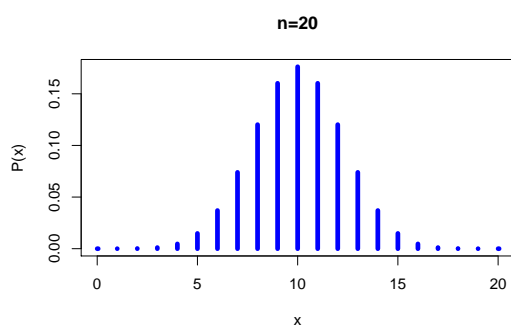
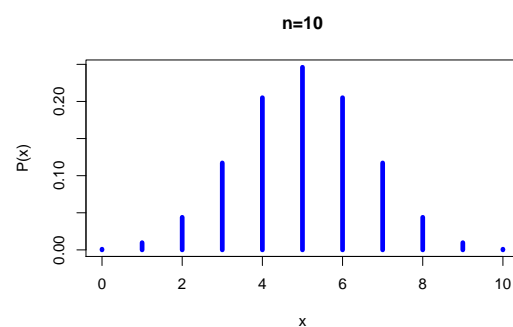
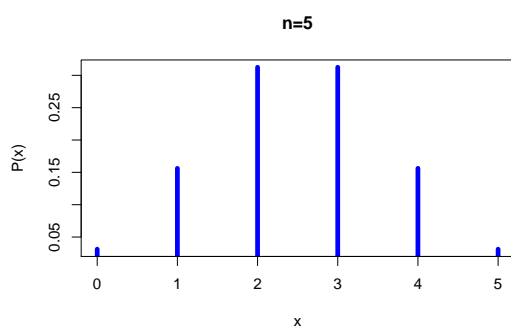
- ▶ Normalfordelingen er en kontinuert fordeling mens binomialfordelingen er en diskret fordeling



- ▶ Histogrammet for en binomialfordeling med $p = 0,5$ og meget højt n ligner tæthedsfunktionen for en normalfordeling

9 / 22

Normalfordelingen som grænsefordeling for binomialfordelingen med $p = 0,5$



10 / 22

Normalfordelingen

- ▶ Der findes uendeligt mange normalfordelinger, der hver især er karakteriseret ved deres **middelværdi** μ og deres **spredning** σ
- ▶ Middelværdi μ og spredning σ er parametre i normalfordelingen, og vi skriver $N(\mu, \sigma^2)$
- ▶ **Tæthedsfunktionen** er en klokkeformet kurve:

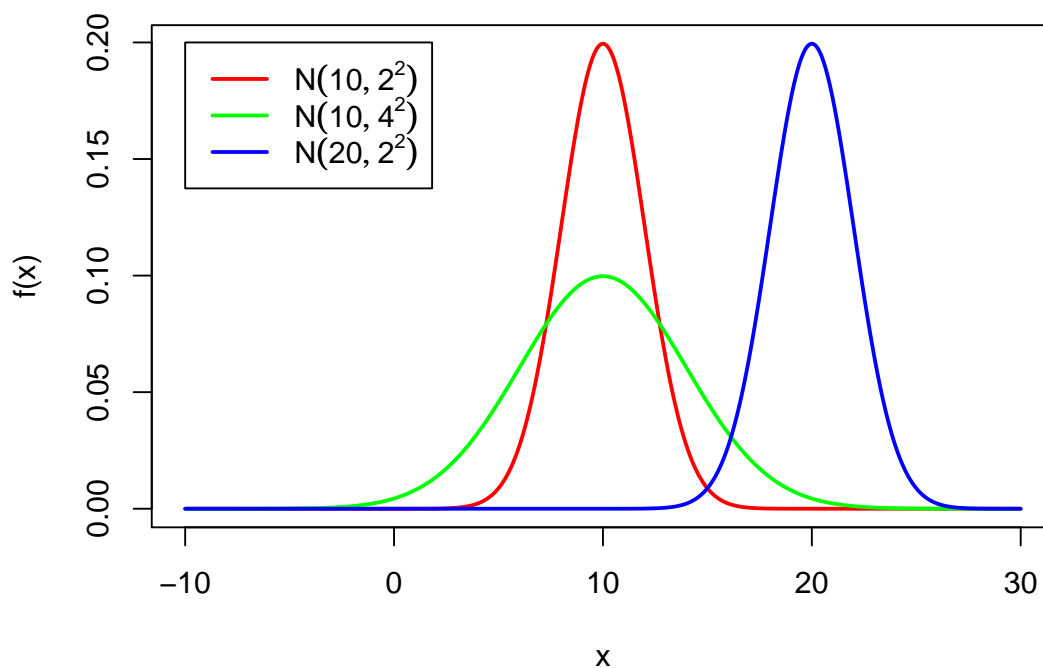
$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

(vi bruger heldigvis næsten altid tabeller)

- ▶ Kurven har toppunkt for $x = \mu$
- ▶ Større spredning giver fladere tæthedsfunktion

11 / 22

Tæthedsfunktion for 3 forskellige normalfordelinger



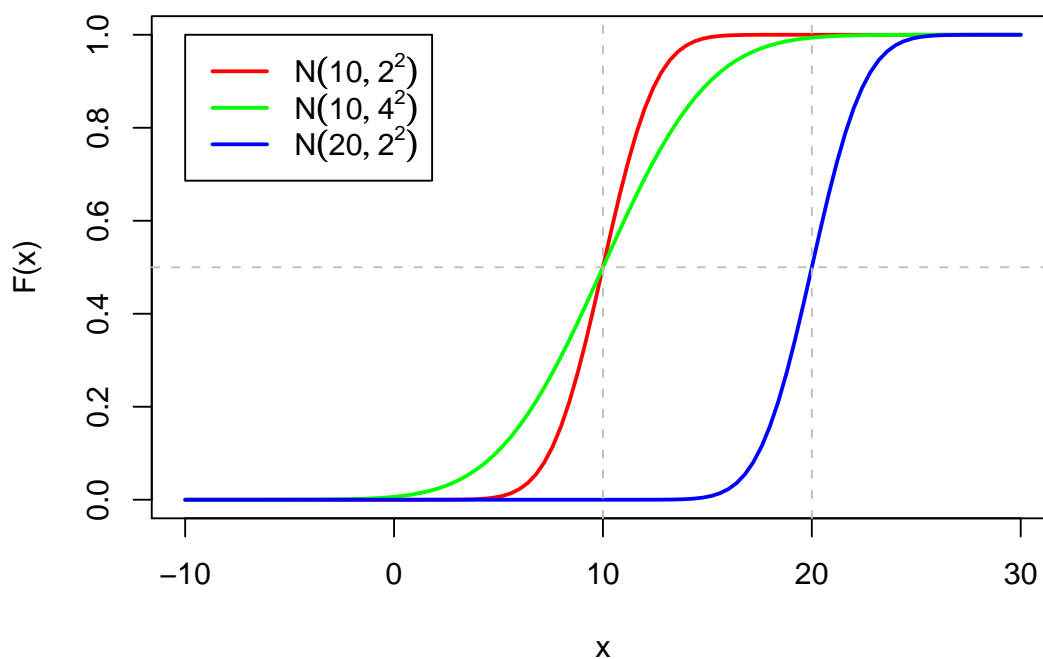
12 / 22

Fordelingsfunktion og standardnormalfordeling

- ▶ Ved bestemt integration af tæthedsfunktionen kommer vi frem til fordelingsfunktionen, der er en slags kummuleret frekvensfordeling
- ▶ Fraktiler i en normalfordeling er nyttige ifm udsagn af typen:
 - ▶ 50% af eleverne kan forventes at score mellem 22 og 87 i den forelagte prøve
 - ▶ 5% af eleverne forventes at score mindre end 12
- ▶ Fordelingsfunktionen går gennem $(\mu, 0.5)$
- ▶ Lavere spredning giver stejlere fordelingsfunktion

13 / 22

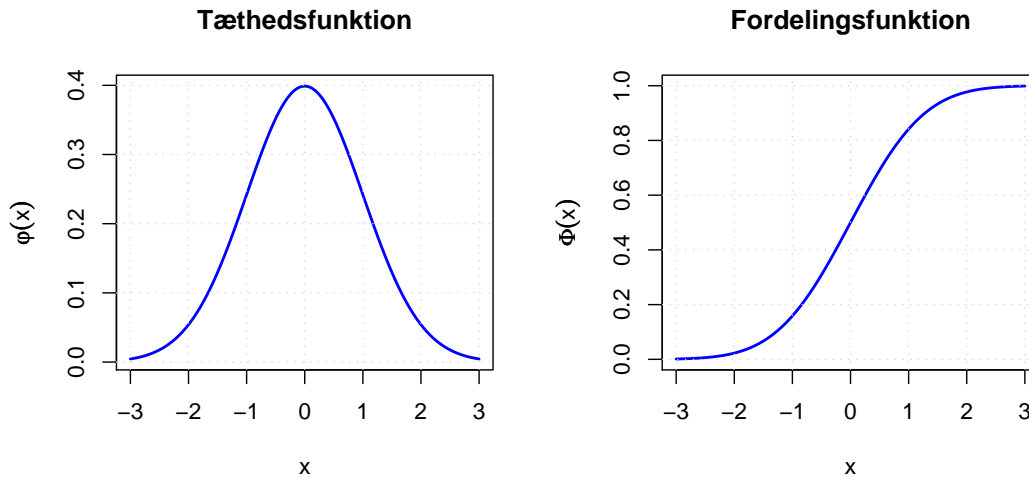
Fordelingsfunktion for 3 forskellige normalfordelinger



14 / 22

Standardnormalfordelingen

- ▶ Der findes uendeligt mange normalfordelinger, men vi kan i praksis klare os med én, nemlig **standardnormalfordelingen** $N(0, 1)$
- ▶ Fordelingsfunktionen $\Phi(x)$ fremkommer ved integration af tæthedsfunktionen $\varphi(x)$



15 / 22

Eksempel på brug af Φ

- ▶ Antag at vi har lavet en undersøgelse, hvor gennemsnittet af scorene er 17 og standardafvigelsen er 3. Vi antager desuden, at scorene følger en normalfordeling.
- ▶ Vi vil nu gerne kende sandsynligheden for, at en tilfældig score er mindre end 14.
- ▶ Vi normaliserer ved at beregne den såkaldte z-værdi:

$$z = \frac{x - \bar{X}}{s} = \frac{14 - 17}{3} = -1$$

- ▶ Ved opslag i Tabel A kan vi nu se at

$$p = P(x \leq 14) = \Phi(-1) = 0,159$$

- ▶ Sandsynligheden for at en tilfældig score er mindre en 14 er altså cirka 16%

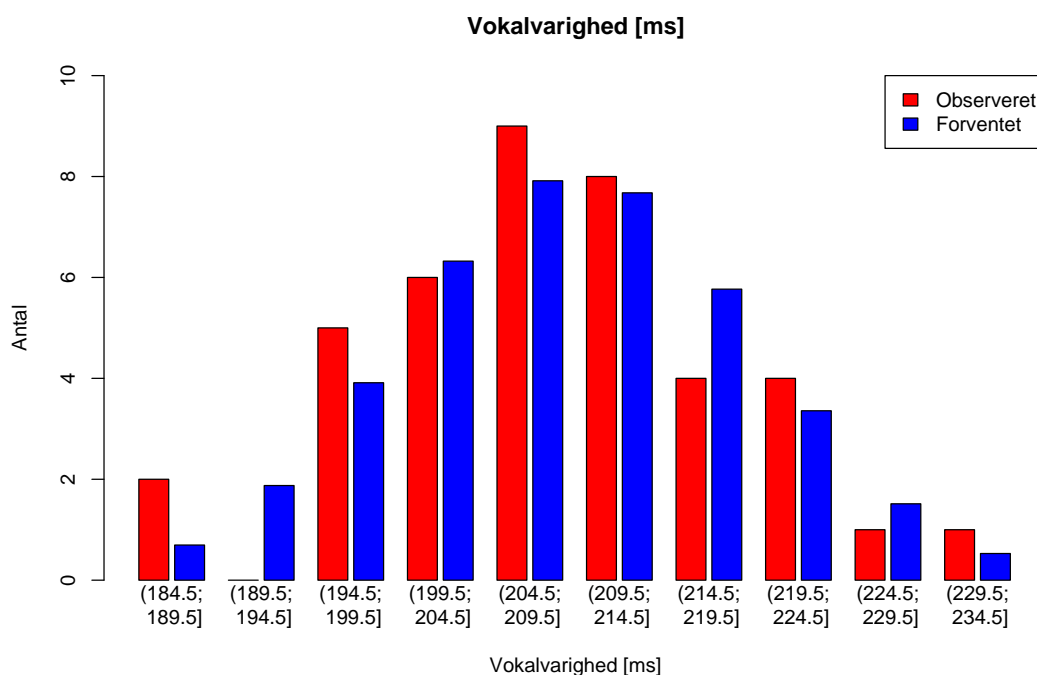
16 / 22

Modelkontrol

- ▶ Vi er ofte interesserede i at se, hvor godt vores stikprøve egentlig stemmer overens med normalfordelingsantagelsen
- ▶ For Tabel 2.5 (vokalvarighed i ms) beregner vi det forventede antal observationer i et bestemt interval under antagelsen om normalitet og sammenligner med det observerede
- ▶ Vi beregner $\bar{x} = 208,9$ og $s = 9,79$
- ▶ For klassen afgrænset ved $(204,5; 209,5]$ beregnes to z-værdier til $-0,45$ og $0,06$
- ▶ Via Tabel A findes tilhørende sandsynligheder p som $0,326$ og $0,524$
- ▶ Sandsynligheden for at være i intervallet er derfor $0,524 - 0,326 = 0,198$
- ▶ Da stikprøven omfatter 40 enheder forventer vi at finde $40 \cdot 0,198 = 7,92$ enheder i intervallet
- ▶ Der var faktisk 9...

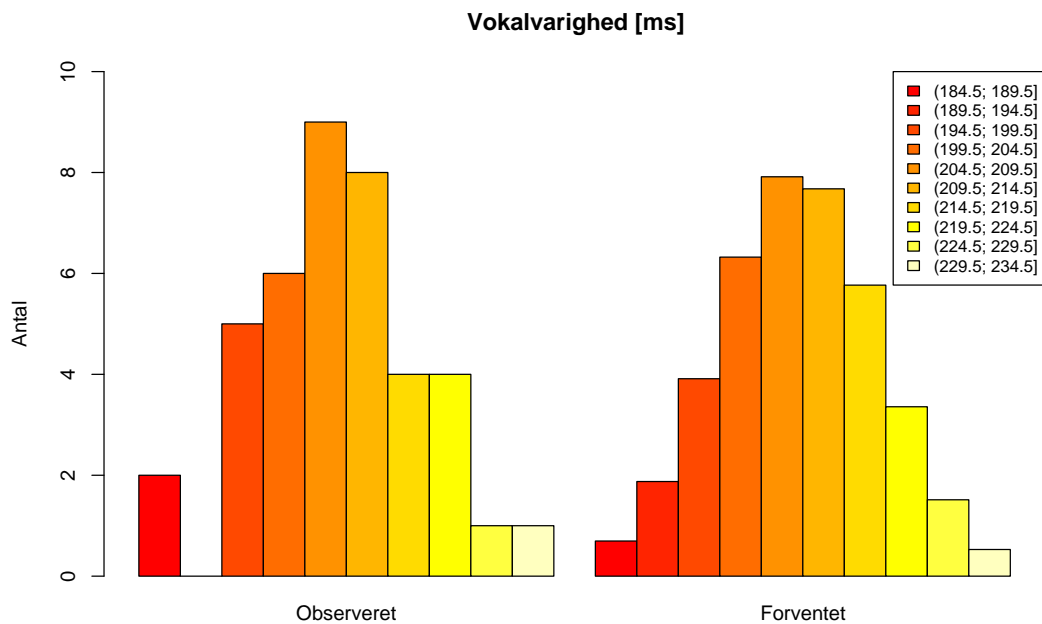
17 / 22

Grafisk modelkontrol



18 / 22

Grafisk modelkontrol



19 / 22

Normalfordeling i Excel

- ▶ Der kan beregnes værdier for både $f(x)$ og $F(x)$ for vilkårlige normalfordelinger med funktionen `normfordeling(...)`, der tager fire argumenter:
 - ▶ værdi af x
 - ▶ middelværdi μ
 - ▶ spredning σ
 - ▶ kumulativ: 0 betyder nej (der regnes med f) og 1 betyder ja (der regnes med F)
- ▶ Der kan findes fraktiler for vilkårlige normalfordelinger med funktionen `norminv(...)`, der tager tre argumenter:
 - ▶ sandsynlighed p
 - ▶ middelværdi μ
 - ▶ spredning σ

Dette svarer til at finde x_p i ligningen

$$F(x_p) = p \Rightarrow x_p = F^{-1}(p)$$

20 / 22

Normalfordeling i Excel

- ▶ Ønsker man at finde værdier i standardnormalfordelingen kan man benytte funktionerne `standardnormfordeling(...)` og `standardnorminv(...)`, der tager ét argument hver
- ▶ ... men det er nok lige så nemt at angive $\mu = 0$ og $\sigma = 1$ i de generelle funktioner
- ▶ I praksis bruger vi stort set kun tæthedsfunktionen når vi skal tegne pæne klokkeformede kurver — det er næsten altid fordelingsfunktionen, der er den interessante

21 / 22

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.

22 / 22