Intro til statistik

II Inferens og lineære modeller I

Claus Thorn Ekstrøm
Biostatistik, KU
ekstrom@sund.ku.dk

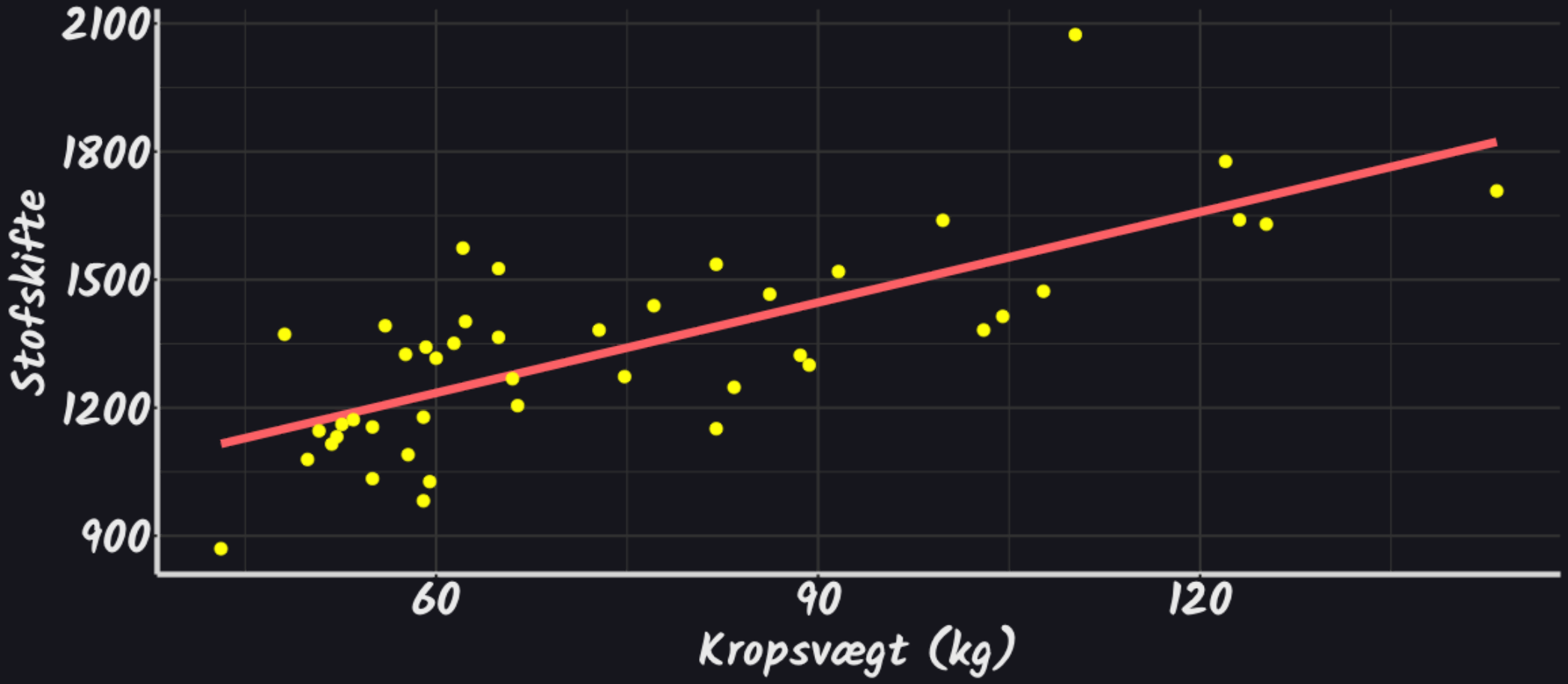
Torsdag 7. maj 2020

Slides @ biostatistics.dk/puff/

Plan for i dag

- Lineær regression
- Normalfordelingen
- Inferens: konfidensintervaller og hypotesetests

Hvilende stofskifte og kropsvægt



Lineær regression

Vil beskrive en sammenhæng mellem to kvantitative variable.

Metode: lineær regression

$$y \leftarrow x$$

$$\text{afhængig variabel} \leftarrow \text{uafhængig variabel}$$

$$\text{udfald/outcome} \leftarrow \text{kovariat/prædiktor/forklarende var.}$$

Vi fokuserer på lineære sammenhænge (dette er et valg!)

Lineære regression

Vi fokuserer på *lineære* sammenhænge:

$$y_i = a \cdot x_i + b + \varepsilon_i$$

Hvad der er den afhængige og den uafhængige variabel skal bestemmes ud fra den videnskabelige kontekst!

Regressionen af y på x er ikke det samme som regressionen af x på y!

Fortolkning af estimaterne

$$y_i = a \cdot x_i + b + \varepsilon_i$$

- b er værdien af y, når $x = 0$ (**skæringspunktet** med y-aksen)
- a er linjens **hældning** (gnst. ændring i y når x ændrer sig med $+1$)

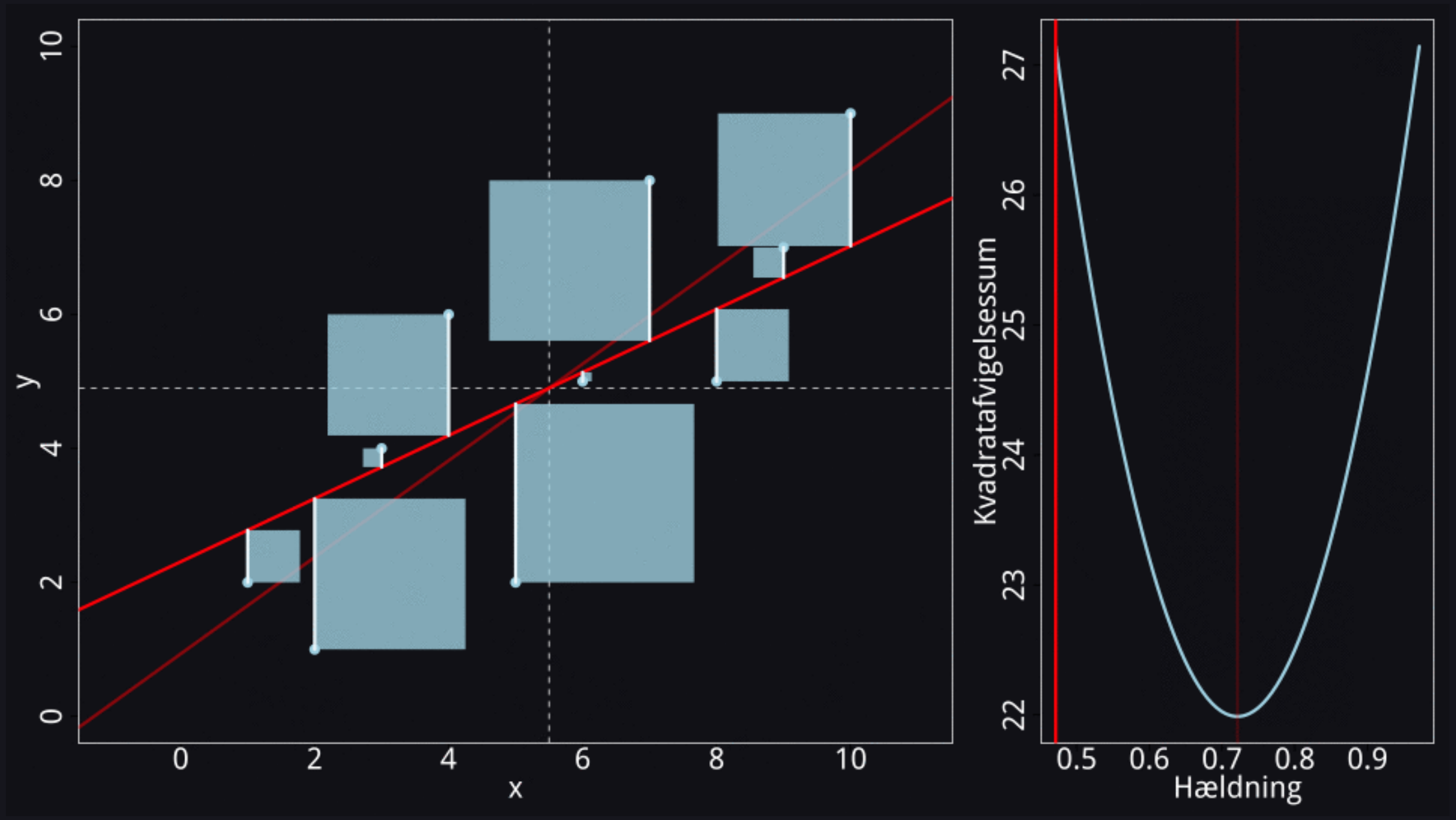
Den rette linje beskriver den *ideelle* eller overordnede sammenhæng mellem de to variable.

I praksis ligger vores datapunkter ikke perfekt på en linje.

Hvilende stofskifte og kropsvægt

```
library("tidyverse")
library("broom")
lm(rmr ~ bw, data=indata) %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   811.      77.0       10.5  2.29e-13
## 2 bw              7.06     0.978      7.22 7.03e- 9
```


y
x
0
2
4
6
8
10
Kvadratafvigelsessum
22
23
24
25
26
27
0.5
0.6
0.7
0.8
0.9
Hældning

Mindste kvadraters metode

Minimér de kvadratiske residualer

$$r_i = y_i - f_\beta(x_i)$$

$$\arg\min \sum_{i=1}^{N} (y_i - f_\beta(x_i))^2$$

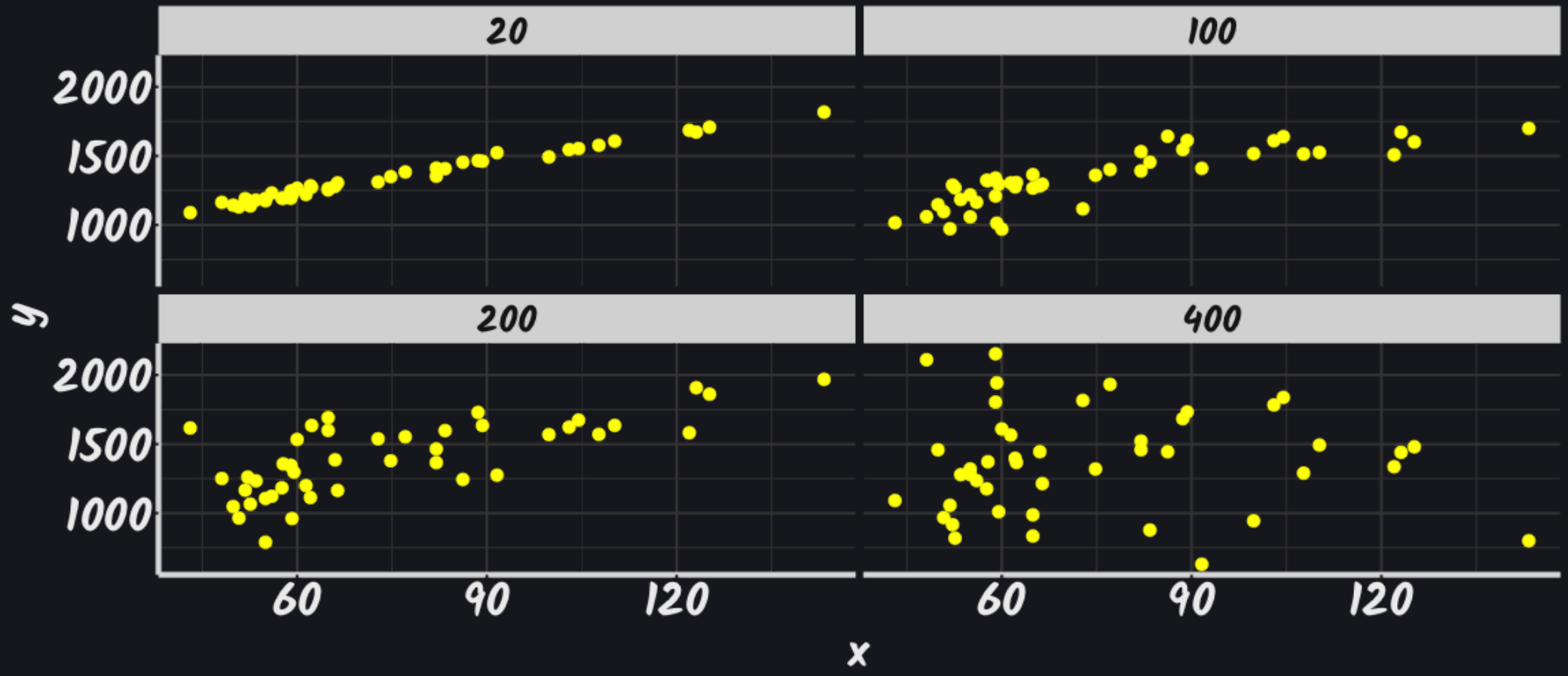
Hvor god er modellen? Prædiktionen?

Observeret - forventet

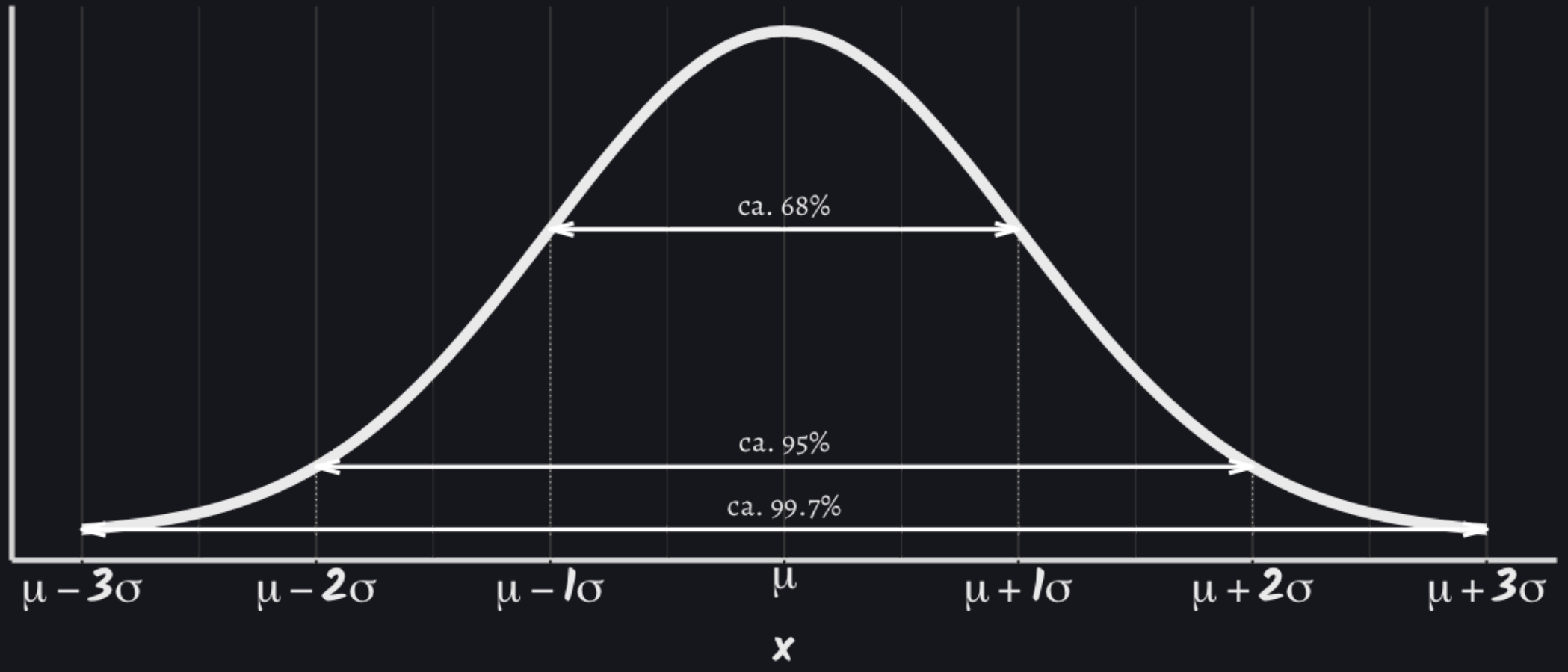
Fejlleddet

$$y_i = a \cdot x_i + b + \varepsilon_i, \quad \varepsilon \sim N(0, \sigma^2)$$

Fejlleddets betydning



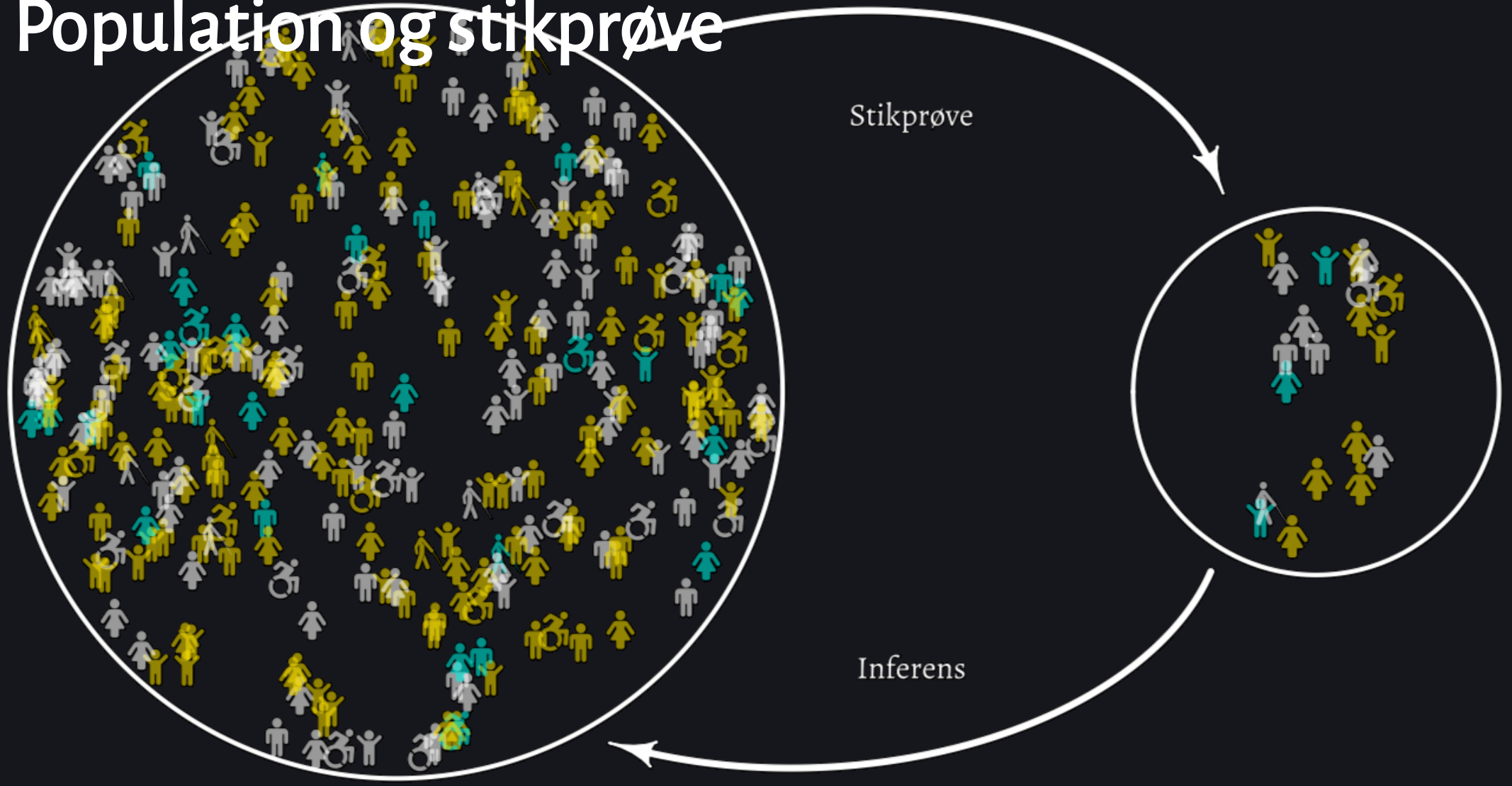
Egenskaber ved normalfordelingen



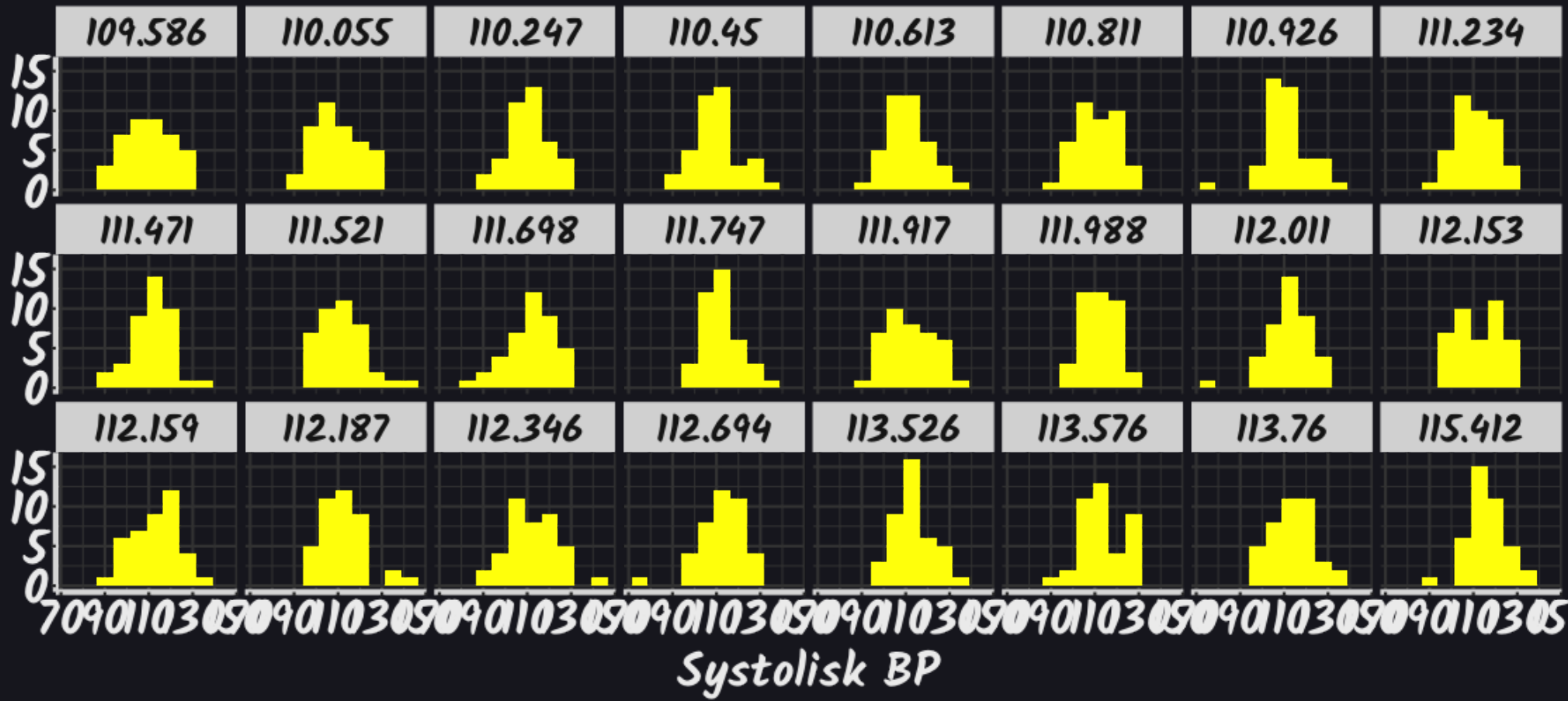
Hvorfor er normalfordelingen så normal?

```
library("shiny")
runGitHub('ShinySampleMean', 'ekstroem')
```

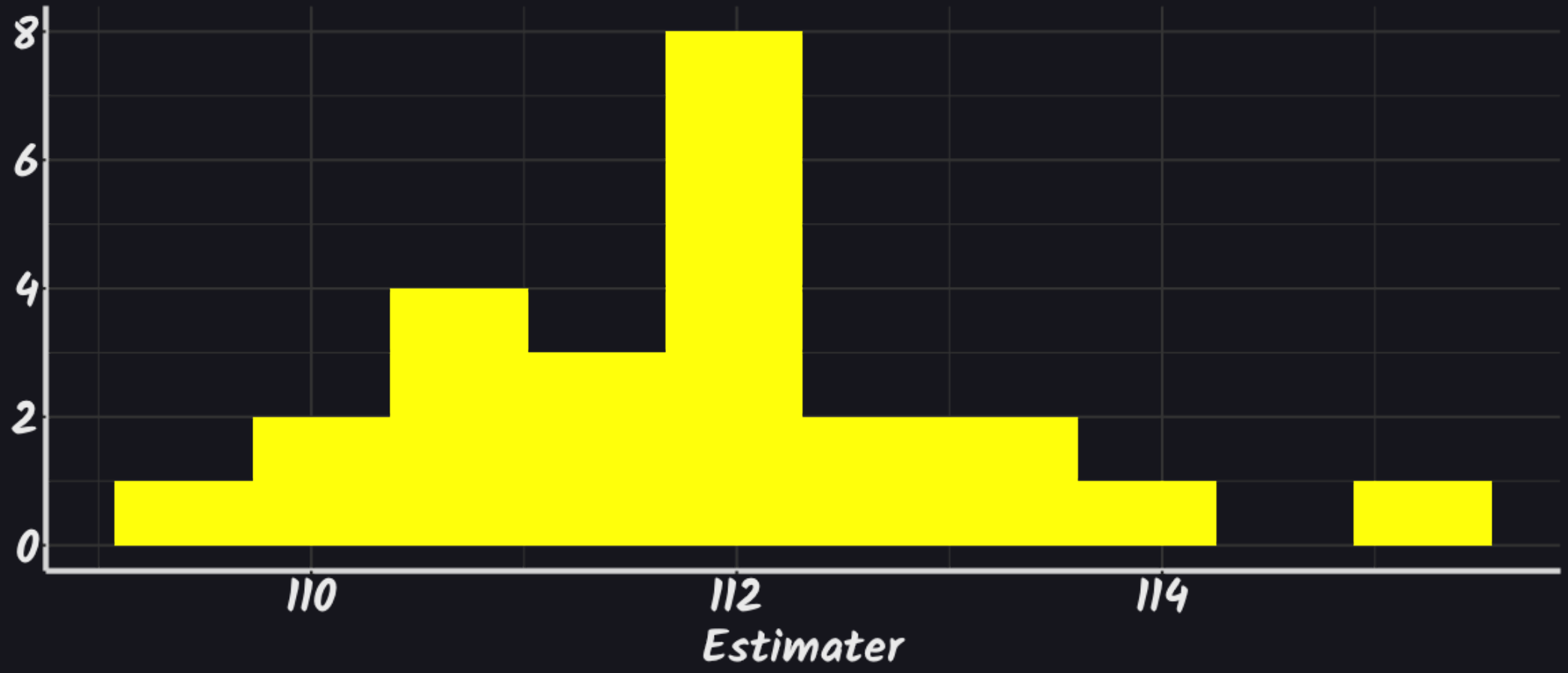
Population og stikprøve



Blodtryk


109.586
110.055
110.247
110.45
110.613
110.811
110.926
111.234
111.471
111.521
111.698
111.747
111.917
111.988
112.011
112.153
112.159
112.187
112.346
112.694
113.526
113.576
113.76
115.412
15
10
5
0
Systolisk BP

Histogram over blodtryksestimater



Konfidensintervaller

Hvad kan vi bruge standardfejlen til?

Standardfejlen hjælper os til at vurdere, hvor god vores stikprøve er til at ramme rigtigt - hvor stor tillid (konfidens) kan vi have til stikprøven?

Et **konfidensinterval**: En procedure hvormed stikprøven anvendes til at indkredse beliggenheden af en **parameter**.

Endepunkterne for konfidensintervallet kaldes også sikkerhedsgrænserne.

Konfidensintervaller

Et 95%-konfidensinterval for middelværdien er *ca* givet ved

$$[\overline{X} - 2 \cdot SE; \overline{X} + 2 \cdot SE]$$

Udregnes i R med funktionen `confint()` og `lm()`.

- Enten indeholder intervallet den sande værdi for middelværdien eller også er der sket en hændelse med lille sandsynlighed ($\leq 5\%$)
- Alt.: Hvis vi trak 100 nye stikprøver, forventer vi at (i snit) 95% af dem indeholdt den sande middelværdi i populationen.

Hypotesetests

Naturvidenskabelig tilgang har en indbygget skepsis - vi vil se stærke beviser for, at der *er* en effekt før vi faktisk tror på det.

Kender vi også fra fx. uskyldsformodningen ("uskyldig til det modsatte er bevist").

Vi starter med at formulere en *nulhypotese*, som beskriver fraværet af en effekt (f.eks. hældningskoefficienten er lig 0).

Så antager vi, at nulhypotesen passer og undersøger hvor godt vores stikprøve stemmer overens med den.

Test af hypotese

Størrelsen af beviset kaldes for p-værdien.

> Sandsynligheden for at få et estimat, der ligger lige så langt eller længere fra nulhypotesen end den værdi, som vi faktisk fik, *hvis* nulhypotesen er sand (og alle andre antagelser omkring modellen *også* er sande).

- Hvis estimatet er meget langt væk fra, hvad man ville forvente under nulhypotesen, tyder det på, at hypotesen er falsk.
- Hvis estimatet er tæt på, hvad man forventer under nulhypotesen, så er vi ikke så villige til at forkaste hypotesen.

p-værdi

Jo *mindre* p-værdien er, jo større evidens er der for, at nulhypotesen ikke stemmer overens med vores stikprøve.

Tommelfingerregel:

$$p \text{ værdien } \begin{cases} < 0.05 & \text{vi forkaster nulhypotesen} \\ \geq 0.05 & \text{vi forkaster ikke nulhypotesen} \end{cases}$$

Bemærk: Grænsen på 5% er et arbitrært valg.

Summary fra lm

```
lm(rmr ~ bw, data=indata) %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   811.      77.0       10.5  2.29e-13
## 2 bw              7.06     0.978      7.22 7.03e- 9
```

Kan også bruge `summary()`.