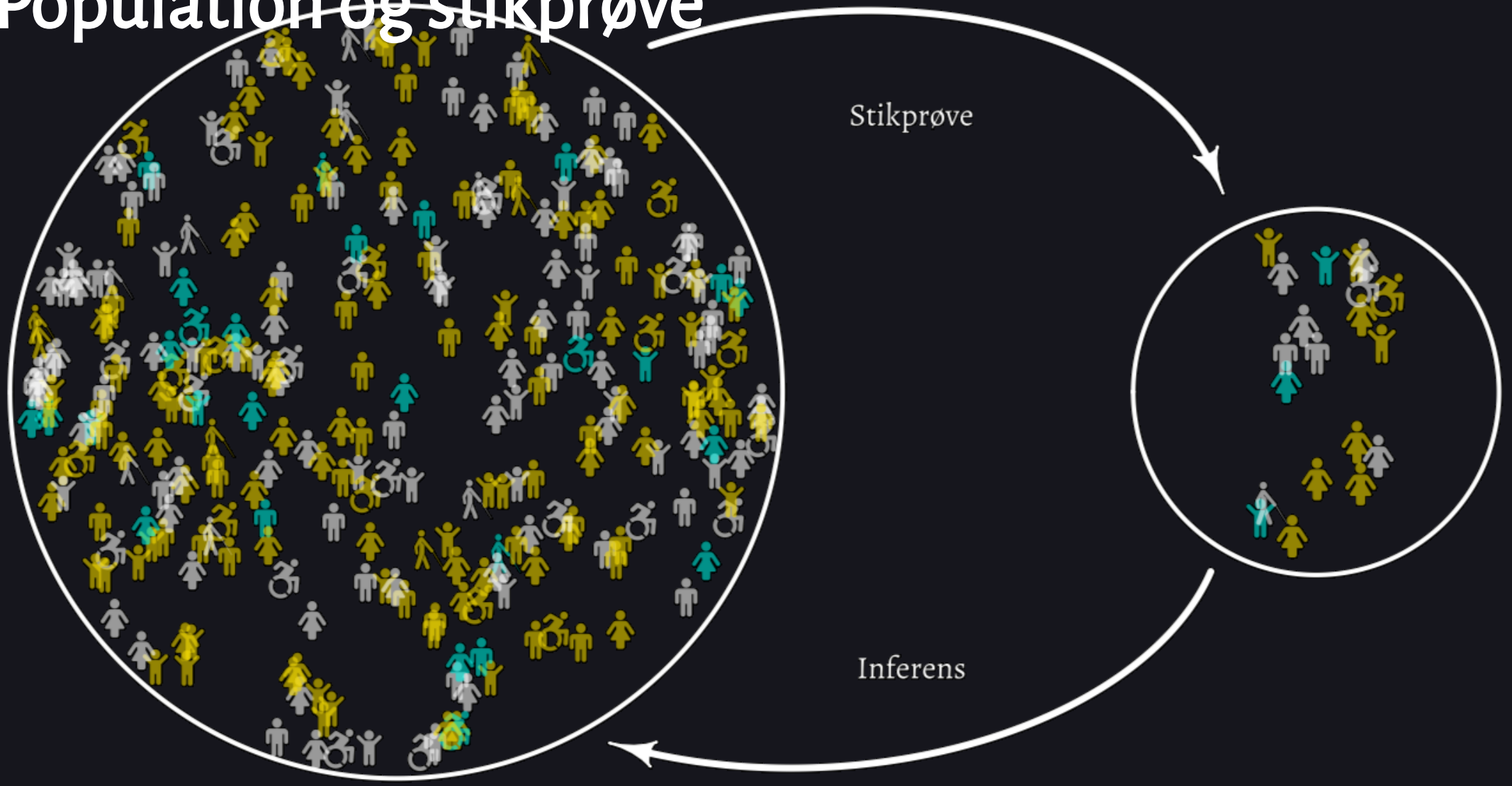
# Konfidensintervaller og vurdering af estimaters usikkerhed

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

Marts 18, 2019

Slides @ biostatistics.dk/talks/

# Population og stikprøve



# Stikprøvevariation

Hvad er danskernes gennemsnitshøjde? $N = 10$

$$\bar{X}_1 = 169 \text{ cm}$$

$$\bar{X}_2 = 183 \text{ cm}$$

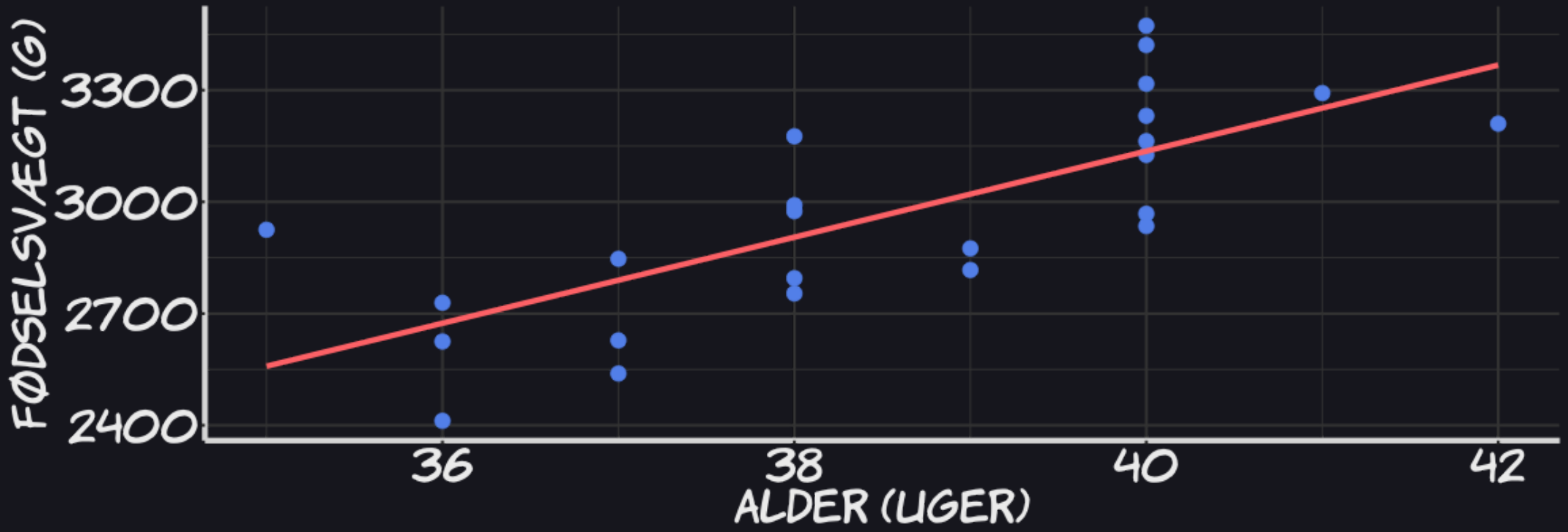
$$\bar{X}_3 = 171 \text{ cm}$$

$$\bar{X}_4 = 113 \text{ cm}$$

$$\bar{X}_5 = 174 \text{ cm}$$

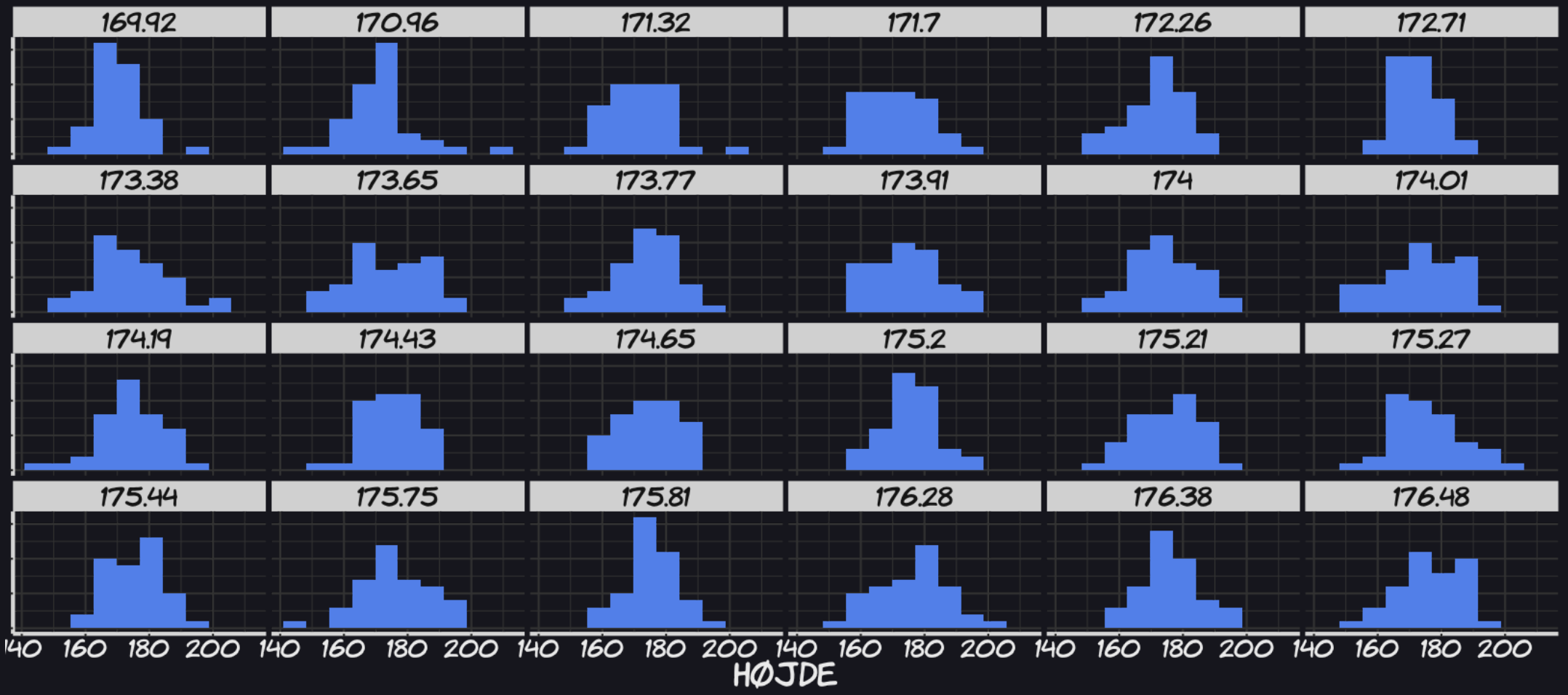
# Hvorfor er et estimates præcision vigtig?

Sammenhængen mellem fødselsvægt og fostrets alder (i uger). $\hat{\beta} = 116$.

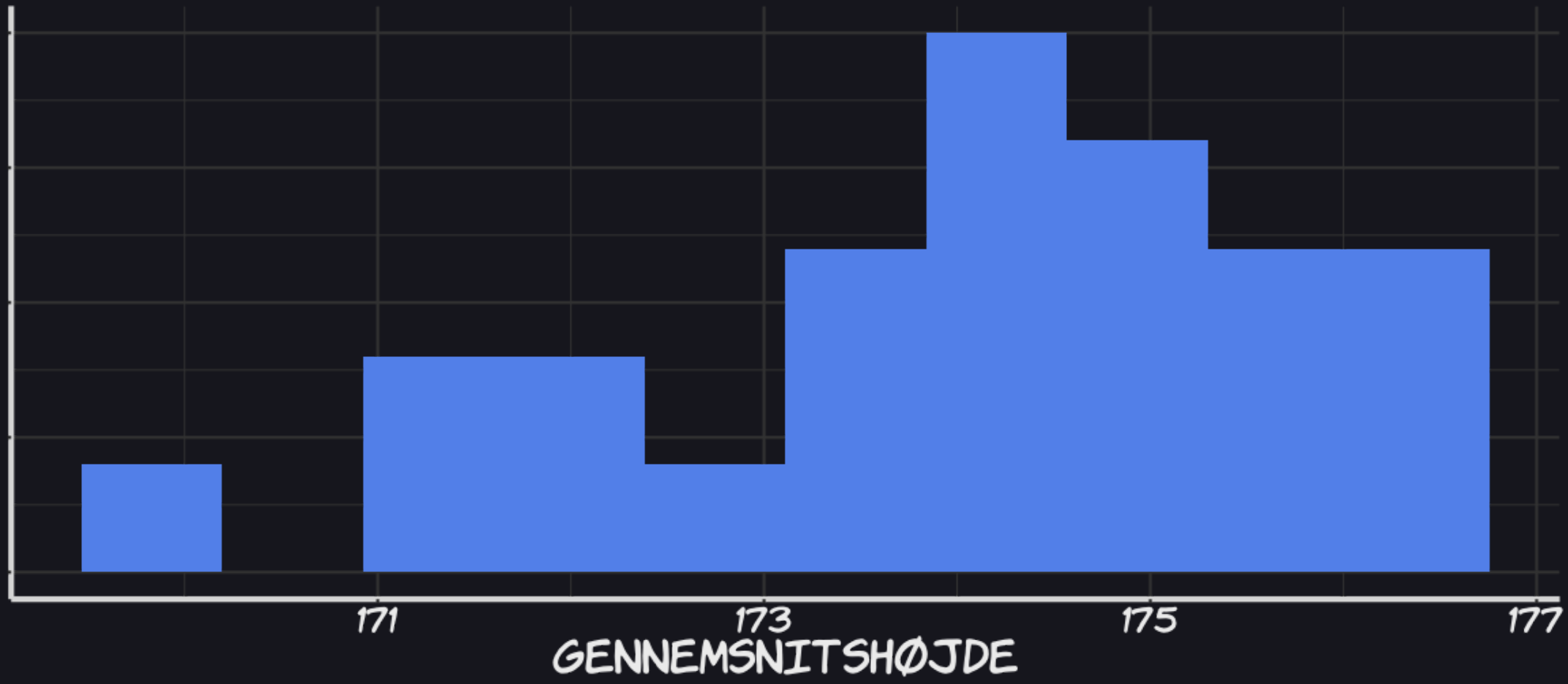



Estimater er de (biologisk/fysisk/…) relevante parametre.

# Hvad sker der, hvis vi gentager forsøget?



# Histogram af middelværdier



# Hvad gør man i praksis?

Hvis man nu kendte den data-genererende proces …

Hvis $X$ stokastisk var. med $\mathbb{E}(X) = \mu$ og $\mathbb{V}(X) = \sigma^2$ så vil $a + bX$ have

$$\mathbb{E}(X) = a + b\mu, \text{ og } \mathbb{V}(X) = b^2\sigma^2$$

Hvis $X_1, \ldots, X_N$ har middelværdi $\mu_1, \ldots, \mu_N$ og spredning $\sigma_1, \ldots, \sigma_N$

$$\mathbb{E}(\sum_i X_i) = \sum_i \mu_i$$

$$\mathbb{V}(\sum_i X_i) = \sum_i \sigma_i^2 \text{ (hvis uafh.)}$$

# Den centrale grænseværdisætning

Hvis $X_1, \ldots, X_N$ er uafhængige og identisk fordelte med samme middelværdi $\mu$ og spredning $\sigma$ så vil der gælde for gennemsnittet,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i,$$

og at

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{N})$$

Approksimationen bliver bedre jo større $N$.

# Distribution of Sample Means

**Distribution type:**

Normal

Uniform

Log-normal

Exponential

Add samples one at a time

**Sample size:**

1 29 200

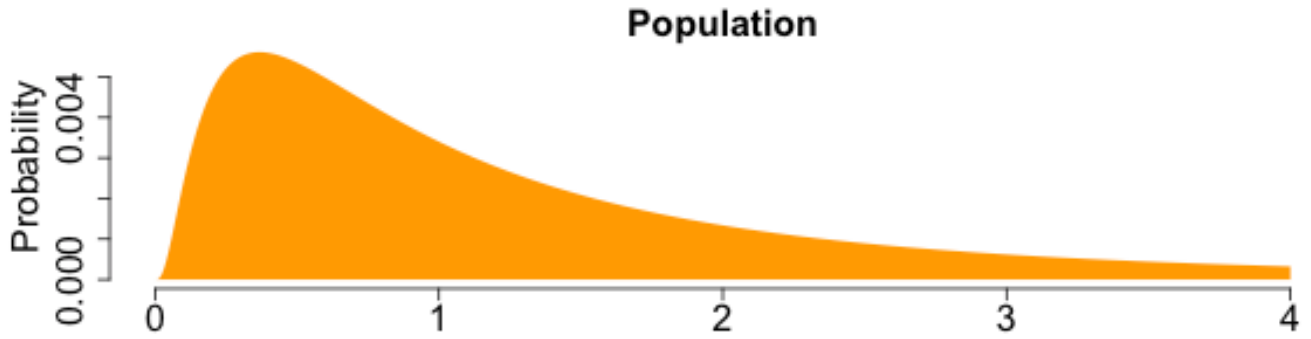
1 21 41 61 81 101 121 141 161 181 200

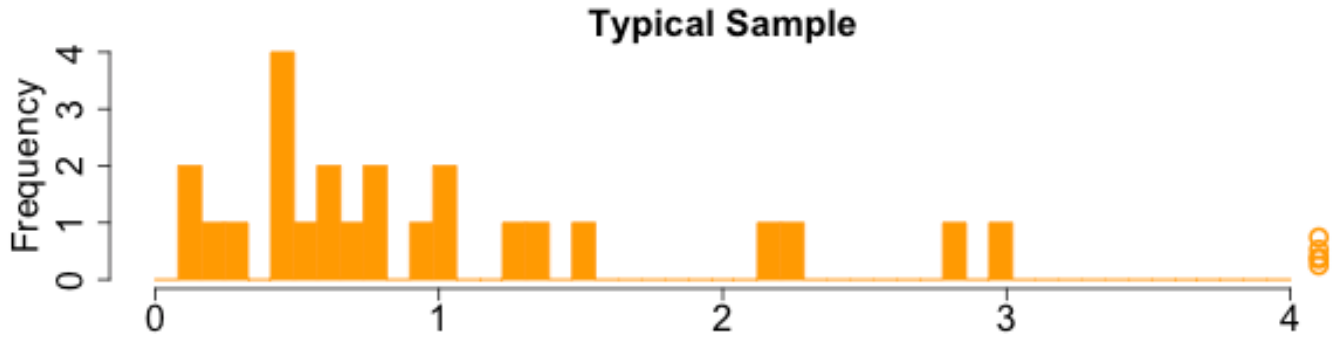
**Number of repetitions:**

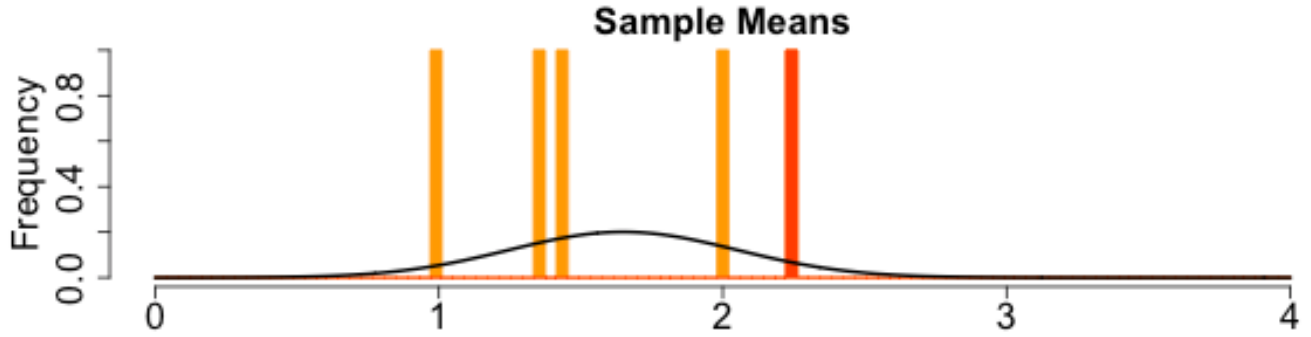
1 1,000

1 101 201 301 401 501 601 701 801 901 1,000

Draw New Sample



# Måleusikkerhed

Hvis den data-genererende proces er

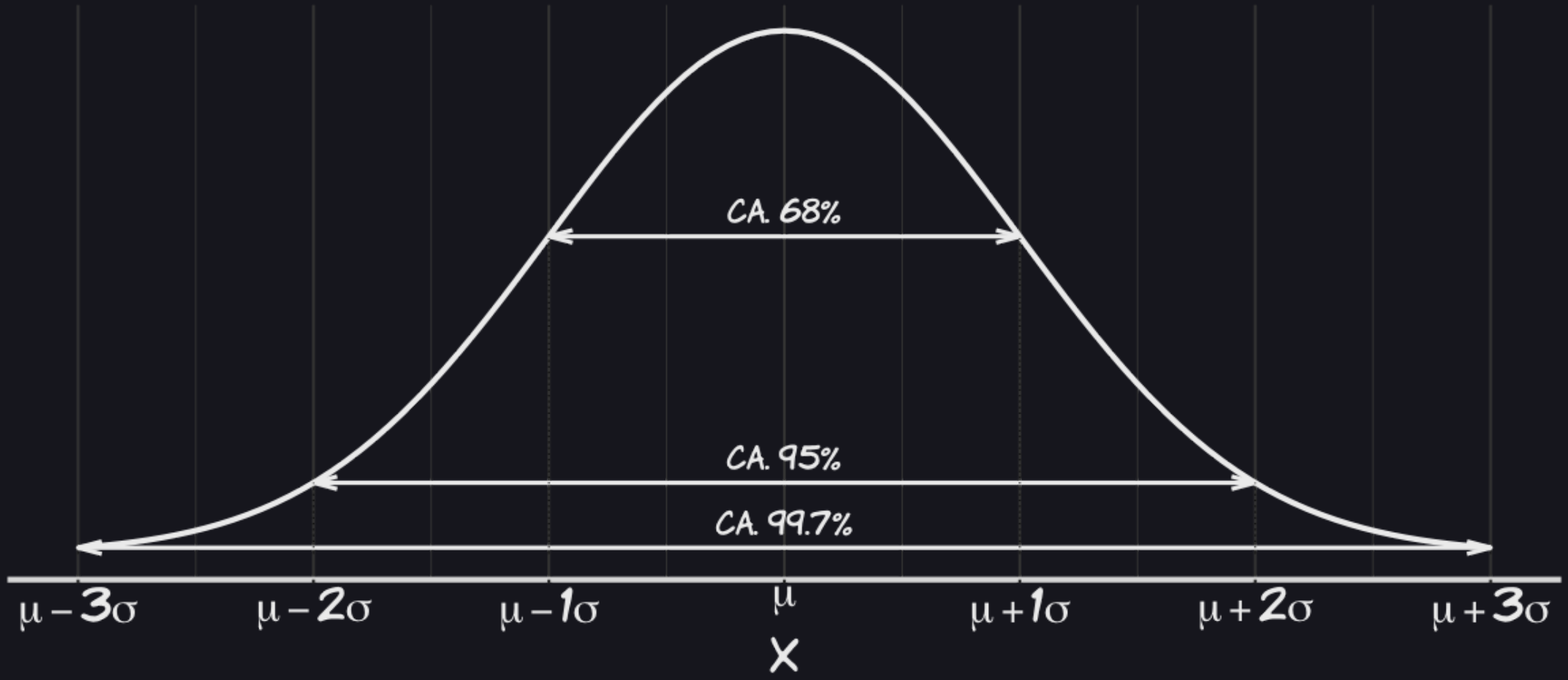
$$\text{observation} = \underbrace{\mu}_{\text{sand værdi}} + \underbrace{\varepsilon}_{\text{støj}}$$

hvor $\mathbb{E}(\varepsilon) = 0$ og $\mathbb{V}(\varepsilon) = \sigma^2$ så vil (for fast grænse $\tau$)

$$|\bar{X} - \mu| \leq \tau \Leftrightarrow -\tau \leq \bar{X} - \mu \leq \tau$$

Men CLT giver, at $\bar{X} - \mu \approx N(0, \sigma^2/N)$

# Egenskaber ved normalfordelingen



# Intervaller

For $X \sim N(\mu, \sigma^2)$ vil

$$P(|X - \mu| \leq 2\sigma) \approx 0.95$$

så

$$\begin{aligned} P(-2\sigma \leq X - \mu \leq 2\sigma) \Leftrightarrow \\ P(-X - 2\sigma \leq -\mu \leq -X + 2\sigma) \Leftrightarrow \\ P(X + 2\sigma \geq \mu \geq X - 2\sigma) \Leftrightarrow \\ P(X - 2\sigma \leq \mu \leq X + 2\sigma) = 0.95 \end{aligned}$$

# Konfidensintervaller

Konfidensinterval for en parameter $\mu$:

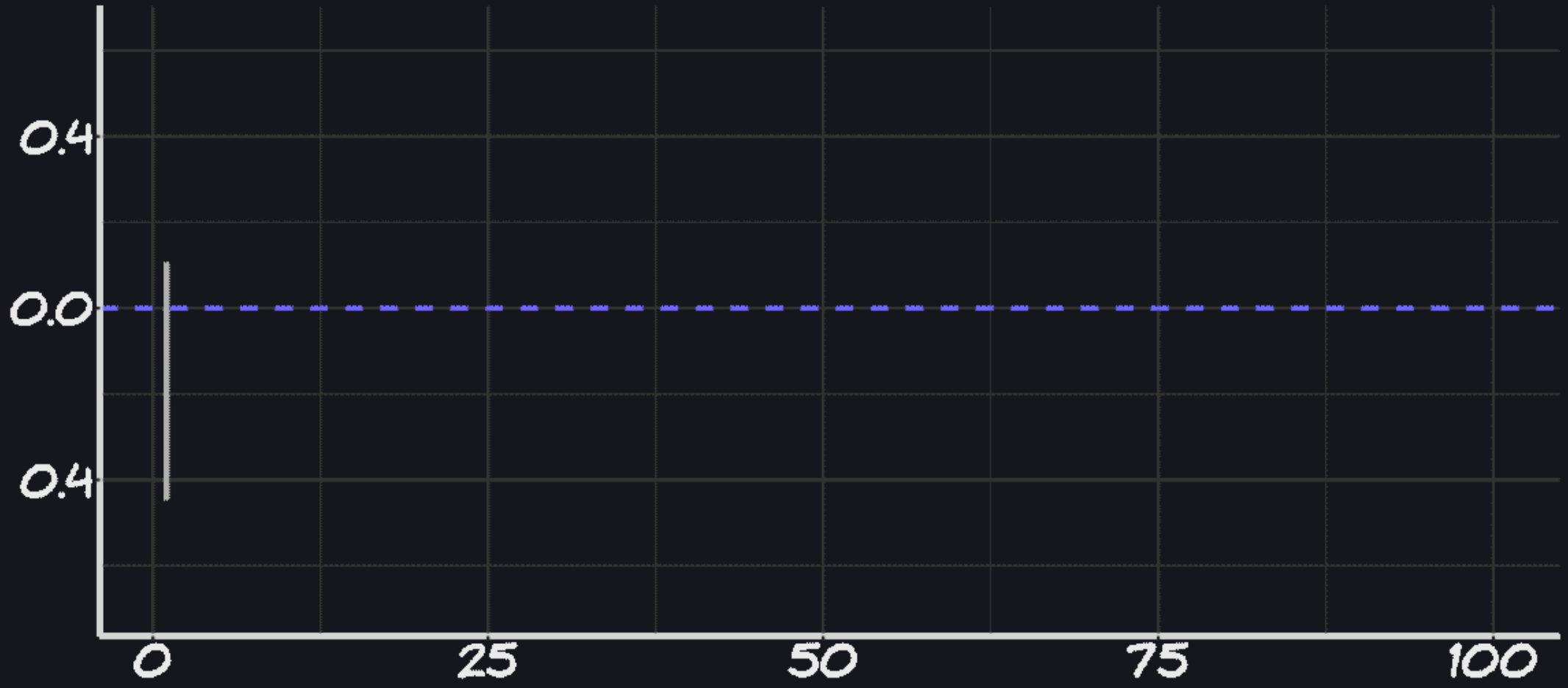
# Konfidensintervaller

Hvis vi hver gang vi udfører et eksperiment hævder, at den ukendte parameter ligger i det beregnede 95% interval, så tager vi kun fejl i 5% af tilfældene.

Et konfidensinterval er altid for en parameter.

Kan gøre intervallerne bredere for at være mere sikre (men også mere upræcise).

# Simulerede konfidensintervaller



# Fortolkning af konfidensintervaller

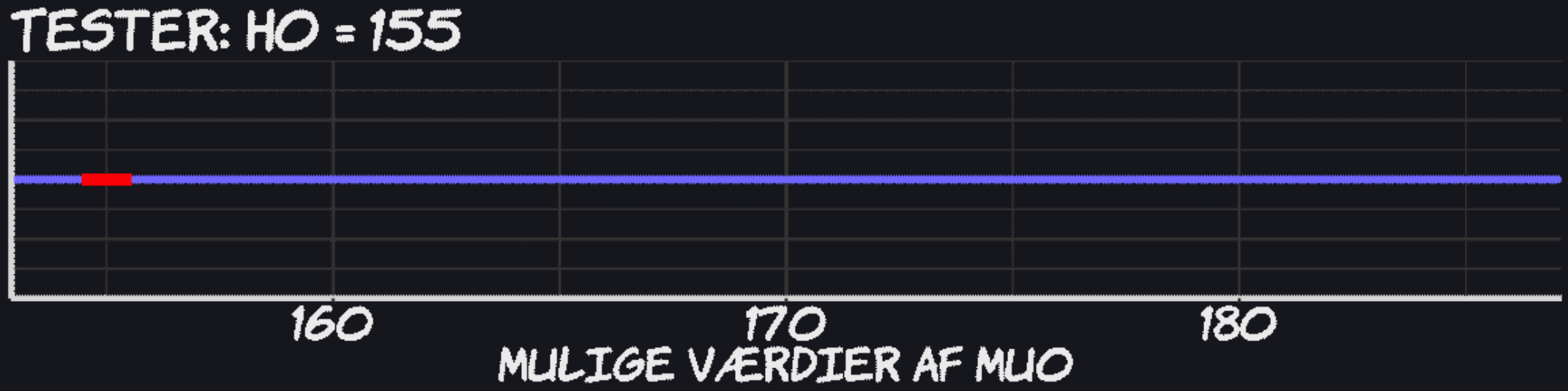
> Jeg er 95% sikker på, at intervallet fra [165 ; 175] indeholder den sande gennemsnitlige højde for danskere.

I virkeligheden: enten 0% eller 100%, men vi ved ikke hvilken.

De 95% henviser derfor til den generelle procedure med at lave konfidensintervaller.

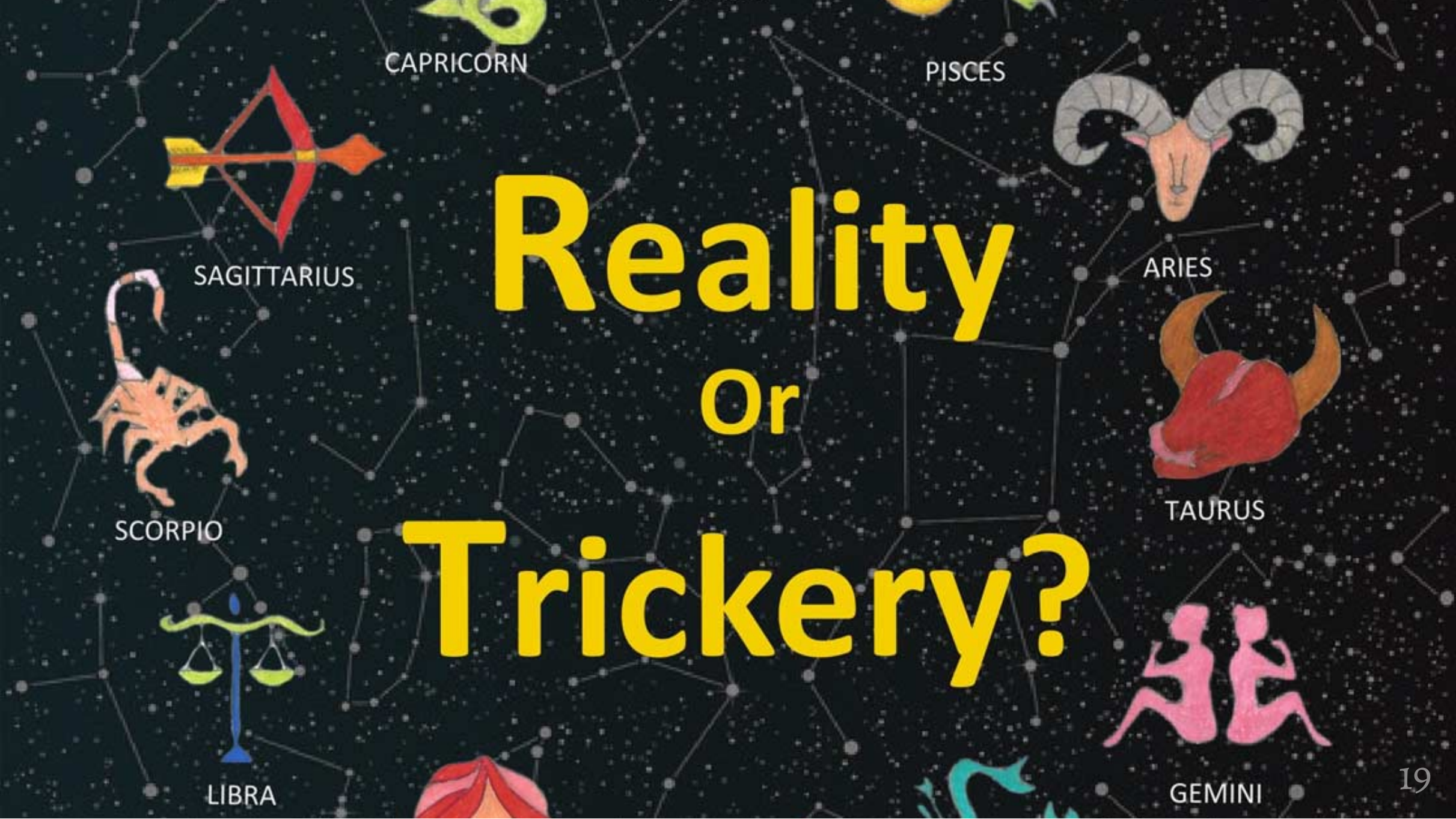
# Nulhypotesen og konfidensintervaller

Når man tester en nulhypotese, $H_0 : \mu = \mu_0$ så er 95% konfidensintervallet netop de værdier, der *ikke* bliver forkastet.




De værdier for nulhypotesen, som data ikke er i modstrid med.

# Binomialfordelingen


CAPRICORN
PISCES
SAGITTARIUS
ARIES
Reality
Or
Trickery?
SCORPIO
TAURUS
LIBRA
GEMINI

# Binomialfordelingen

Antagelser om en binomialfordelt variabel

- $N$ uafhængige forsøg
- To mulige udfald: succces og fiasko
- Samme successandsynlighed, $\theta$, i hvert forsøg

F S S S S F F S F S S S S S F F S F F F

Estimat:

$$\hat{y} = \frac{\#\ \text{Gunstige}}{\#\ \text{Mulige}}$$

# Binomialfordelingen

Antagelser om en binomialfordelt variabel

- $N$ uafhængige forsøg
- To mulige udfald: succces og fiasko
- Samme successandsynlighed, $\theta$, i hvert forsøg

```
0 1 1 1 1 0 0 1 0 1 1 1 1 1 0 0 1 0 0 0
```

Estimat:

$$\hat{\theta} = \frac{\#\text{ Gunstige}}{\#\text{ Mulige}} = \frac{\sum_i y_i}{N}$$

# Approksimativt KI for binomialfordelingen

For binomialfordelt variabel er $\hat{\sigma}^2 = \hat{\theta}(1-\hat{\theta})$ så et 95% KI for $\theta$ er ca.

$$[\hat{\theta} - 1.96\frac{\hat{\sigma}}{\sqrt{N}}; \hat{\theta} + 1.96\frac{\hat{\sigma}}{\sqrt{N}}]$$

# Generel formel

Et 95% konfidensinterval for en parameter $\mu$ har generelt formen

$$[\hat{\mu} - 1.96 \cdot SE(\hat{\mu}); \hat{\mu} + 1.96 \cdot SE(\hat{\mu})]$$

Standardfejlen - *standard error* - er **spredningen på estimatet**.

For horoskopdata: $N = 87, Y = 27$ så $\hat{\theta} = 0.32$ og

$$0.32 \pm 1.96 \cdot \sqrt{\frac{0.32 \cdot (1 - 0.32)}{84}} = [0.22; 0.42]$$

# Udvidelser

# Lineær regression

Antag $Y_1, \ldots, Y_N$ følger en regressionsmodel

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

hvor $x_1, \ldots, x_N$ er kendte og $\varepsilon_i \sim N(0, \sigma^2)$.

LS giver estimaterne

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \quad \text{og} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Disse estimater er normalfordelte (lineære funktioner af data)!

# Varianser ifm lineær regression

$\hat{\alpha}$ og $\hat{\beta}$ har varianser

$$\mathbb{V}(\hat{\alpha}) = \sigma^2 \frac{\sum_i x_i^2}{N \sum_i (x_i - \bar{x})^2} \quad \text{og} \quad \mathbb{V}(\hat{\beta}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

$\sigma^2$ estimeres ved

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_i \underbrace{\left(y_i - (\hat{\alpha} + \hat{\beta} x_i)\right)^2}_{\text{residual}}$$

Så følger KI direkte.

# Fødselsdata

```
lm(weight ~ age, data=birthweight) %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -1485.     853.      -1.74 0.0955
## 2 age             116.      22.1      5.23 0.0000304
```

95% KI for $\beta$ : $116 \pm 1.96 \cdot 22.1 = [72.7; 159.3]$

# Konfidensintervaller og prædiktionsintervaller

Et *konfidensinterval* siger noget om realistiske værdier for en parameter. Et *prædiktionsinterval* siger noget om realistiske værdier for en enkelt observation.