02323 Introduktion til statistik

Uge 7: Simulation og bootstrapping

Nicolai Siim Larsen
DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

1. Introduktion til simulation
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrapping
   - Introduktion til bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
4. Ikke-parametrisk bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Dagsorden

Motivation

- Mange relevante stikprøvefunktioner har komplicerede fordelinger. Det kunne f.eks. være:
  - Medianen
  - Fraktiler
  - Den interkvartile variationsbredde (IQR)
  - Variationskoefficienten (coefficient of variation)
  - Ikke-lineære funktioner af en eller flere variable
  - Variansen (el. spredningen)
- Vi mangler værktøjer, når antagelserne for vores test ikke er opfyldte.

Motivation

- Mange relevante stikprøvefunktioner har komplicerede fordelinger. Det kunne f.eks. være:
  - Medianen
  - Fraktiler
  - Den interkvartile variationsbredde (IQR)
  - Variationskoefficienten (coefficient of variation)
  - Ikke-lineære funktioner af en eller flere variable
  - Variansen (el. spredningen)
- Vi mangler værktøjer, når antagelserne for vores test ikke er opfyldte.
- **Løsning**: Simulation og bootstrapping – R er et super værktøj til dette!

Simulation

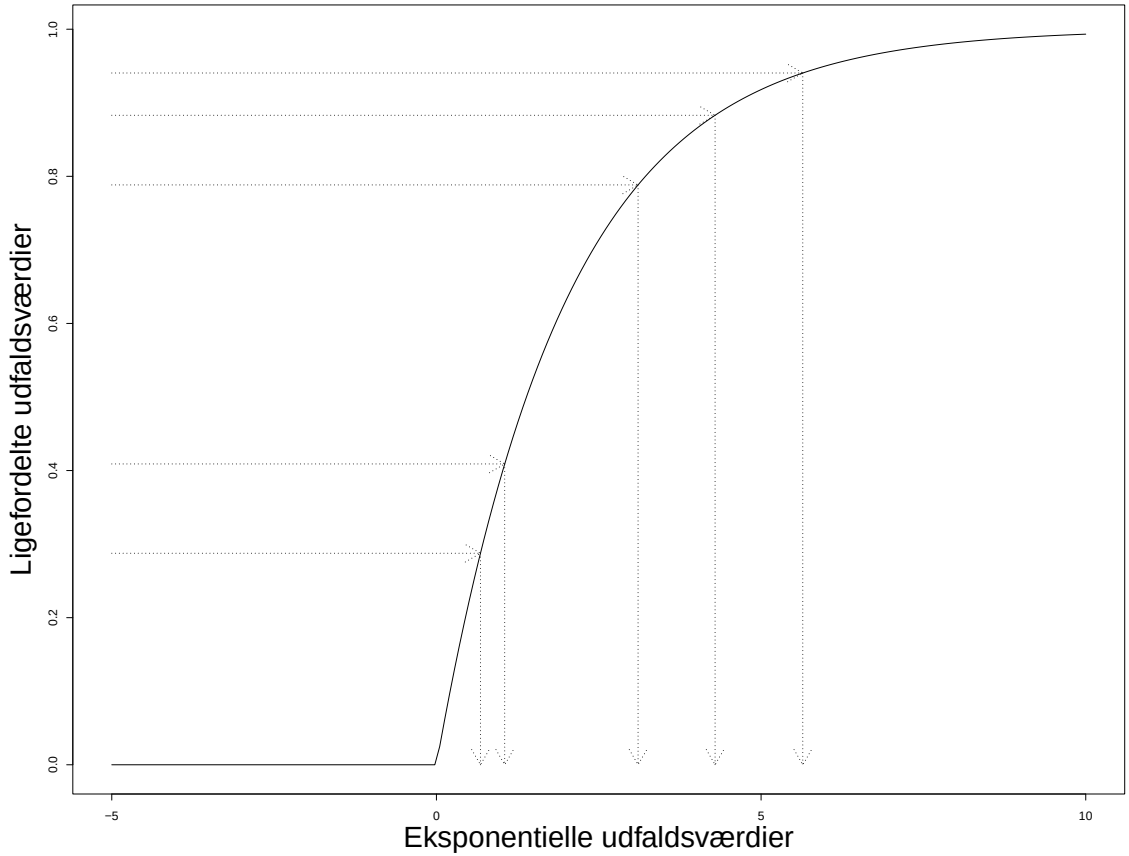
- (Pseudo)-tilfældige tal genereret af en computer.
- En *tilfældighedsgenerator* er en algoritme, der kan generere en talfølge af tilsyneladende tilfældige tal.
- Algoritmen kræver en begyndelsesværdi kaldet et *seed*.
- Man kan simulere fra (næsten) alle fordelinger igennem den uniforme fordeling ved at benytte følgende resultat:

Sætning 2.51: Alle fordelinger kan "fremskaffes" fra den uniforme fordeling

Hvis $U \sim \text{Uniform}(0,1)$ og F er fordelingsfunktionen for en given sandsynlighedsfordeling, så vil $F^{-1}(U)$ følge fordelingen givet ved F.

Eksempel: Eksponentialfordelingen med $\lambda = 0.5$

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



I praksis i R

Mange fordelinger er gjort klar til simulering, for eksempel:

`rbinom`	Binomialfordelingen
`rpois`	Poissonfordelingen
`rhyper`	Den hypergeometriske fordeling
`rnorm`	Normalfordelingen
`rlnorm`	Lognormalfordelingen
`rexp`	Eksponentialfordelingen
`runif`	Den uniforme fordeling (ligefordelingen)
`rt`	t-fordelingen
`rchisq`	χ^2-fordelingen
`rf`	F-fordelingen

Eksempel: Areal af plader

En virksomhed producerer rektangulære plader.

Længden af pladerne (i meter), X, antages at kunne beskrives ved normalfordelingen $N(2, 0.01^2)$, medens bredden af pladerne (i meter), Y, antages at kunne beskrives ved normalfordelingen $N(3, 0.02^2)$. Man kan antage, at pladernes længder og bredder er uafhængige.

Man er interesseret i arealet, A, som er givet ved $A = XY$.

- Hvad er middelarealet?
- Hvad er spredningen i arealet fra plade til plade?
- Hvor ofte har sådanne plader et areal, der afviger mere end 0.1 m^2 fra de angivne 6 m^2?
- Sandsynligheder for andre hændelser.
- Generelt: Hvad er fordelingen for den stokastiske variabel A?

Eksempel: Areal af plader – løsning ved simulation

```
k = 10000                  # Antal simulationer
X = rnorm(k, 2, 0.01) # Længde
Y = rnorm(k, 3, 0.02) # Bredde
A = X*Y                    # Areal

mean(A)                     # Stikprøvegennemsnit
```

[1] 6

```
var(A)                      # Stikprøvevariansen
```

[1] 0.002458

```
mean(abs(A - 6) > 0.1) # P(|A - 6| > 0.1)
```

[1] 0.0439

Dagsorden

1. Introduktion til simulation
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrapping
   - Introduktion til bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
4. Ikke-parametrisk bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\mathrm{Var}(f(X_1,\ldots,X_n)) = \sigma^2_{f(X_1,\ldots,X_n)}$$

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\mathrm{Var}(f(X_1,\ldots,X_n)) = \sigma^2_{f(X_1,\ldots,X_n)}$$

Lineærkombination af uafhængige variable:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \quad \text{når} \quad f(X_1,\ldots,X_n) = \sum_{i=1}^{n} a_i X_i$$

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\text{Var}(f(X_1,\ldots,X_n)) = \sigma^2_{f(X_1,\ldots,X_n)}$$

Lineærkombination af uafhængige variable:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} a_i^2\sigma_i^2 \quad \text{når} \quad f(X_1,\ldots,X_n) = \sum_{i=1}^{n} a_i X_i$$

Metode 4.3: For ikke-lineære funktioner af uafhængige variable $X_1,\ldots,X_n$:

$$\sigma^2_{f(X_1,\ldots,X_n)} \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}(\mathbf{x})\right)^2 \sigma_i^2,$$

hvor $\mathbf{x}$ er et punkt (som regel $\mathbf{x} = (\mu_1, \mu_2, \ldots, \mu_n)$) og $\sigma_i = \text{Var}(X_i)$.

Eksempel: Areal af plader (fortsat)

Vi brugte simulation i den første del af eksemplet.

Nu er vi givet to konkrete målinger for X og Y, $x = 2.00$ m og $y = 3.00$ m: Hvad er variansen af $A = XY$ beregnet med fejlophobningsloven?

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \mathrm{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \mathrm{Var}(Y) = 0.02^2.$$

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \mathrm{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \mathrm{Var}(Y) = 0.02^2.$$

Funktionen og dens partielt afledte er:

$$f(x,y) = xy, \ \frac{\partial f}{\partial x} = y, \ \frac{\partial f}{\partial y} = x.$$

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \text{Var}(Y) = 0.02^2.$$

Funktionen og dens partielt afledte er:

$$f(x,y) = xy, \ \frac{\partial f}{\partial x} = y, \ \frac{\partial f}{\partial y} = x.$$

Så resultatet bliver:

$$\begin{aligned} Var(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2\sigma_1^2 + x^2\sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025. \end{aligned}$$

Fejlophobning – ved simulation

Metode 4.4: Fejlophobning ved simulation

Antag at vi har (faktiske) målinger $x_1, \ldots, x_n$ med kendte/antagede (estimerede) varianser $\sigma_1^2, \ldots, \sigma_n^2$.

1. Simulér k udfaldsværdier af alle n målinger fra de antagne fordelinger, f.eks. $X_i^{(j)} \sim N(x_i, \sigma_i^2)$, $j = 1, \ldots, k$, $i = 1, \ldots, n$.
2. Udregn standardafvigelsen som den observerede standardafvigelse af de k simulerede værdier af $f(X_1^{(j)}, \ldots, X_n^{(j)})$:

$$s^{\text{sim}}_{f(X_1,\ldots,X_n)} = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(f_j - \bar{f})^2}$$

hvor

$$f_j = f(X_1^{(j)}, \ldots, X_n^{(j)}) \quad \text{og} \quad \bar{f} = \frac{1}{k}\sum_{j=1}^{k} f_j.$$

Eksempel: Areal af plader (fortsat)

Faktisk kan vi i dette eksempel udlede variansen for A teoretisk:

$$\begin{aligned}
\mathrm{Var}(XY) &= \mathrm{E}\left[(XY)^2\right] - [\mathrm{E}(XY)]^2 \\
&= \mathrm{E}(X^2)\mathrm{E}(Y^2) - \mathrm{E}(X)^2\mathrm{E}(Y)^2 \\
&= \left[\mathrm{Var}(X) + \mathrm{E}(X)^2\right]\left[\mathrm{Var}(Y) + \mathrm{E}(Y)^2\right] - \mathrm{E}(X)^2\mathrm{E}(Y)^2 \\
&= \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathrm{E}(Y)^2 + \mathrm{Var}(Y)\mathrm{E}(X)^2 \\
&= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\
&= 0.00000004 + 0.0009 + 0.0016 \\
&= 0.00250004.
\end{aligned}$$

Eksempel: Areal af plader (fortsat)

Tre forskellige tilgange:

1. Simulation
2. Den approksimative *fejlophobningslov*
3. Teoretisk udledning

Eksempel: Areal af plader (fortsat)

Tre forskellige tilgange:

1. Simulation
2. Den approksimative *fejlophobningslov*
3. Teoretisk udledning

Simulationstilgangen har nogle vigtige fordele:

1. Nem måde at beregne andre størrelser end blot standardafvigelsen (de teoretiske udledninger kan være meget komplicerede sammenlignet med variansen).
2. Nem måde at bruge andre fordelinger end normalfordelingen, hvis vi tror, at det bedre beskriver virkeligheden.
3. Afhænger ikke af en lineær tilnærmelse af den underliggende ikke-lineære funktion (i modsætning til fejlophobningsloven).

Dagsorden

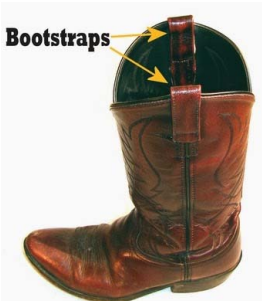
1. Introduktion til simulation
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrapping
   - Introduktion til bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
4. Ikke-parametrisk bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Bootstrapping

Bootstrap = Støvlestrop

Bootstrapping findes i to versioner:

1. Parametrisk bootstrap: Simulér gentagne stikprøver fra den antagede (og estimerede) fordeling.
2. Ikke-parametrisk bootstrap: Simulér gentagne stikprøver direkte fra data.



https://en.wikipedia.org/wiki/Bootstrapping#Etymology

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer middelværdien og intensiteten ud fra data:

$$\hat{\mu} = \bar{x} = 26.08 \text{ og dermed: } \hat{\lambda} = 1/26.08 = 0.03834356.$$

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer middelværdien og intensiteten ud fra data:

$$\hat{\mu} = \bar{x} = 26.08 \text{ og dermed: } \hat{\lambda} = 1/26.08 = 0.03834356.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer middelværdien og intensiteten ud fra data:

$$\hat{\mu} = \bar{x} = 26.08 \text{ og dermed: } \hat{\lambda} = 1/26.08 = 0.03834356.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for μ?

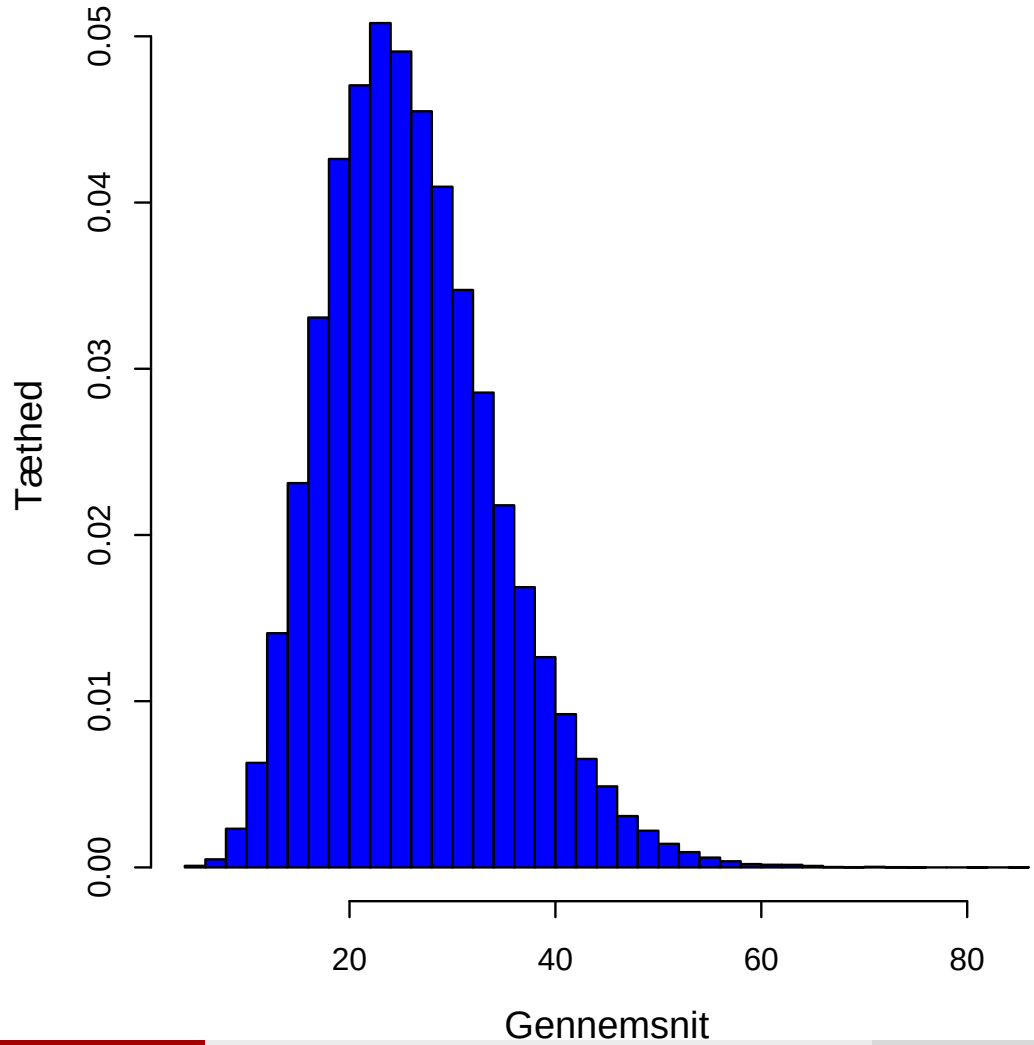
Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```r
# Antal simulationer
k <- 100000

# Simulerer 10 observationer med den estimerede intensitet k gange
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Udregner gennemsnittet af de 10 simulerede observationer k gange
sim_means <- apply(sim_samples, 2, mean)

# Finder relevante fraktiler i fordelingen af de k simulerede gennemsnit
quantile(sim_means, c(0.025, 0.975))

##  2.5% 97.5%
## 12.59 44.63
```

Eksempel: Histogram

```r
# Histogram over simulerede gennemsnit
hist(sim_means, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Gennemsnit", ylab = "Tæthed")
```



Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer meadianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer meadianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center:

$$32.6,\ 1.6,\ 42.1,\ 29.2,\ 53.4,\ 79.3,\ 2.3,\ 4.7,\ 13.6,\ 2.0.$$

Vi estimerer meadianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for medianen?

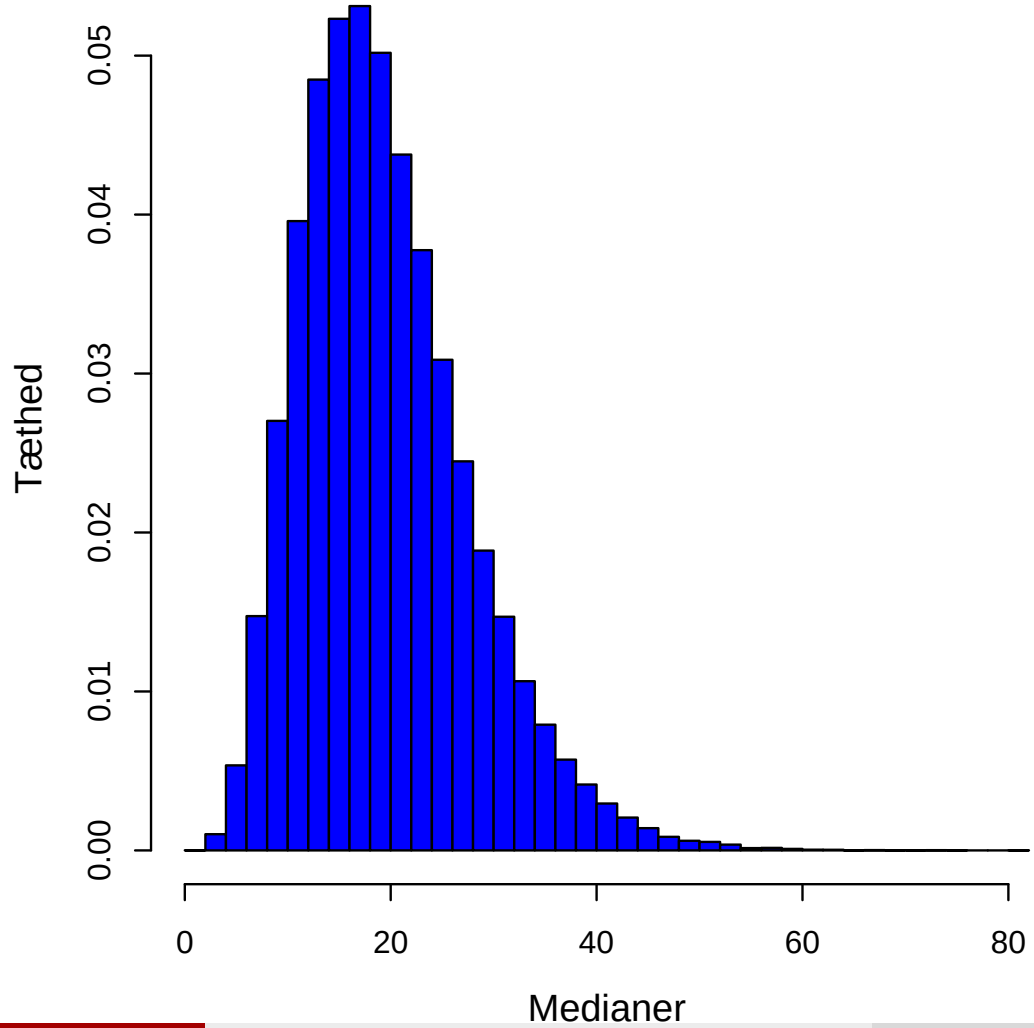
Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
# Antal simulationer
k <- 100000

# Simulerer 10 observationer med den estimerede intensitet k gange
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Udregner medianen af de 10 simulerede observationer k gange
sim_medians <- apply(sim_samples, 2, median)

# Finder relevante fraktiler i fordelingen af de k simulerede medianer
quantile(sim_medians, c(0.025, 0.975))

##   2.5%  97.5%
##  7.038 38.465
```

Eksempel: Histogram

```
# Make histogram of simulated medians
hist(sim_medians, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Medianer", ylab = "Tæthed")
```



Konfidensinterval for en vilkårlig stikprøvefunktion (inkl. μ)

Metode 4.7: Konfidensinterval for en vilkårlig stikprøvefunktion θ ved parametrisk bootstrapping

Antag at vi har faktiske observationer $x_1, \ldots, x_n$, og at disse kommer fra en sandsynlighedsfordeling (med tæthed) f.

1. Simulér k stikprøver af n observationer fra den antagede fordeling f, hvor middelværdien er lig $\bar{x}$. [a]
2. Udregn estimatet $\hat{\theta}$ for hver af de k stikprøver, $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find $\alpha/2$- og $(1-\alpha/2)$-fraktilerne i $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $(1-\alpha)$-konfidensinterval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

[a] Andre parametre/størrelser i fordelingen skal også matche data bedst muligt. Nogle fordelinger har mere end en parameter, f.eks. har log-normalfordelingen to parametre. Mere generelt bør man anvende den såkaldte *maximum likelihood* tilgang.

Eksempel: 99% KI for Q_3 i en normalfordeling

```
# Indlæser data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)

# Definerer en Q3-funktion
Q3 <- function(x){quantile(x, 0.75)}

# Antal simulationer
k <- 100000

# Simulerer 10 observationer med estimerede parametre k gange
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))

# Udregner Q3 af de 10 simulerede observationer k gange
simQ3s <- apply(sim_samples, 2, Q3)

# Finder relevante fraktiler i fordelingen af de k simulerede Q3
quantile(simQ3s, c(0.005, 0.995))

##  0.5% 99.5%
## 172.8 198.0
```

Konfidensinterval for en vilkårlig stikprøvefunktion (sammenligning) $\theta_1 - \theta_2$ (inkl. $\mu_1 - \mu_2$) fra to stikprøver

Metode 4.10: Konfidensinterval for en vilkårlig sammenligning $\theta_1 - \theta_2$ baseret på to stikprøver ved parametrisk bootstrapping:

Antag at vi har faktiske observationer $x_1, \ldots, x_{n_1}$ og $y_1, \ldots, y_{n_2}$, og at disse kommer fra sandsynlighedsfordelinger f_1 og f_2. (Fordelingerne antages uafhængige)

1. Simulér k grupper af 2 stikprøver med hhv. n_1 og n_2 observationer fra de antagede fordelinger, hvor middelværdierne er hhv. $\hat{\mu}_1 = \bar{x}$ og $\hat{\mu}_2 = \bar{y}$.
2. Udregn forskellen mellem stikprøvefunktionerne i hver af de k stikprøver: $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find $\alpha/2$- og $(1-\alpha/2)$-fraktilerne i disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $(1-\alpha)$-konfidensinterval: $\left[q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right]$.

Eksempel: Konfidensinterval for forskellen mellem middelværdierne i to eksponentialfordelinger

```r
# Dag 1 data
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3,  2.3 , 4.7, 13.6, 2.0)
n1 <- length(x)

# Dag 2 data
y <- c(9.6, 22.2, 52.5, 12.6, 33.0, 15.2, 76.6, 36.3, 110.2,
       18.0, 62.4, 10.3)
n2 <- length(y)
```

Eksempel: Konfidensinterval for forskellen mellem middelværdierne i to eksponentialfordelinger

```r
# Antal simulationer
k <- 100000

# Simulerer k par af stikprøver med hhv. n1 = 10 and n2 = 12 observationer

# fra eksponentialfordelinger med de estimerede intensiteter.

simX_samples <- replicate(k, rexp(n1, 1/mean(x)))
simY_samples <- replicate(k, rexp(n2, 1/mean(y)))

# Udregner forskellen mellem de simulerede middelværdier k gange
sim_dif_means <- apply(simX_samples, 2, mean) -
  apply(simY_samples, 2, mean)

# Finder relevante fraktiler i fordelingen af de k simulerede forskelle
quantile(sim_dif_means, c(0.025, 0.975))

##   2.5%  97.5%
## -40.74  14.12
```

Parametrisk bootstrapping: Et overblik

Vi antager en eller anden fordeling!

To metoder med konfidensintervaller bliver givet:

	Med en SP	Med to SP'er
Vilkårlig stikprøvefunktion	Metode 4.7	Metode 4.10

Dagsorden

1. Introduktion til simulation
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrapping
   - Introduktion til bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
4. Ikke-parametrisk bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Ikke-parametrisk bootstrapping: Et overblik

Vi antager *ikke* noget om fordelinger!

To metoder med konfidensintervaller bliver givet:

	Med en SP	Med to SP'er
Vilkårlig stikprøvefunktion	Metode 4.15	Metode 4.17

Eksempel: Kvinders cigaretforbrug

I et studie undersøgte man kvinders cigaretforbrug før og efter fødsel. Man fik følgende observationer af antal cigaretter pr. dag:

før	efter	før	efter
8	5	13	15
24	11	15	19
7	0	11	12
20	15	22	0
6	0	15	6
20	20		

Sammenlign middelværdierne før og efter! Er der sket nogen ændring i gennemsnitsforbruget?

Eksempel: Kvinders cigaretforbrug

En parret test, *men* data er tydeligvis ikke normalfordelt!

```
# Data
x1 <-  c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <-  c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)

# Udregner forskellene
dif <- x1-x2
dif

##  [1]  3 13  7  5  6  0 -2 -4 -1 22  9

# Udregner gennemsnitsforskellen
mean(dif)

## [1] 5.273
```

Eksempel: Kvinders cigaretforbrug – Ikke-parametrisk bootstrapping

```
t(replicate(5, sample(dif, replace = TRUE)))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]   -2    0    9   22    0   -1    0   -2    0     3     0
## [2,]   13    3   -2   -1   -2    7   13   -4   -2    -1     5
## [3,]    9   -4    5   -4    5    3   -4   13    3     0    22
## [4,]   -1   22   -2   -1   13    6   -4    0    0    -1    22
## [5,]    9   -2   13    6    9   22    0   -1    7     7    -1
```

Eksempel: Kvinders cigaretforbrug – Resultater

Lad os finde et 95%-konfidensinterval for *middelændringen* i cigaretforbruget.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_means = apply(sim_samples, 2, mean)
quantile(sim_means, c(0.025,0.975))

##  2.5% 97.5%
## 1.364 9.818
```

Konfidensinterval for en vilkårlig stikprøvefunktion θ (inkl. μ) fra en stikprøve

Metode 4.15: Konfidensinterval for en vilkårlig stikprøvefunktion θ ved ikke-parametrisk bootstrapping

Antag at vi har observeret $x_1, \ldots, x_n$.

1. Simulér k stikprøver af størrelse n ved tilfældig trækning (med tilbagelægning) fra de observerede/tilgængelige data.
2. Udregn estimatet $\hat{\theta}$ for hver af de k stikprøver: $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find $\alpha/2$- og $(1-\alpha/2)$-fraktilerne for disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $(1-\alpha)$-konfidensinterval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$.

Eksempel: Kvinders cigaretforbrug

Lad os finde et 95%-konfidensinterval for *medianændringen* i cigaretforbruget i eksemplet fra før.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_medians = apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025,0.975))

##  2.5% 97.5%
##    -1     9
```

Eksempel: Tandsundhed og spædbørns brug af flaske

I et studie undersøgtes det om børn, der som spæde havde fået mælk fra flaske, havde dårligere eller bedre tænder end dem, der ikke havde fået mælk fra flaske. Fra 19 tilfældigt udvalgte børn registrerede man, hvornår de havde haft deres første tilfælde af karies:

Flaske	Alder	Flaske	Alder	Flaske	Alder
N	9	N	10	J	16
J	14	N	8	J	14
J	15	N	6	J	9
N	10	J	12	N	12
N	12	J	13	J	12
N	6	N	20		
J	19	J	13		

Eksempel: Tandsundhed og spædbørns brug af flaske – 95%-konfidensinterval for $\mu_1 - \mu_2$

```
# Indlæser data
x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
y <- c(14,15,19,12,13,13,16,14,9,12)

# 95% KI: gns. forskel ved ikke-parametrisk bootstrapping
k <- 100000
simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_mean_difs <- apply(simx_samples, 2, mean)-
                          apply(simy_samples, 2, mean)
quantile(sim_mean_difs, c(0.025,0.975))

##     2.5%    97.5%
## -6.2111 -0.1111
```

Konfidensinterval for $\theta_1 - \theta_2$ (inkl. $\mu_1 - \mu_2$) ved ikke-parametrisk bootstrapping fra to stikprøver

Metode 4.17: Konfidensinterval for $\theta_1 - \theta_2$ ved ikke-parametrisk bootstrapping fra to stikprøver

Antag at vi har observationer $x_1, \ldots, x_{n_1}$ og $y_1, \ldots, y_{n_2}$.

1. Udtag k par bootstrap-stikprøver med hhv. n_1 og n_2 observationer fra de respektive stikprøver (ved tilfæld trækning med tilbagelægning).
2. Udregn forskellen mellem estimaterne i hver af de k par bootstrap-stikprøver:
   $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find $\alpha/2$- og $(1-\alpha/2)-$fraktilerne i disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $(1-\alpha)$-konfidensinterval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$.

Eksempel: Tandsundhed og spædbørns brug af flaske – Et 99%-konfidensinterval for median-forskellen

```
k <- 100000
simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_median_difs <- apply(simx_samples, 2, median)-
                        apply(simy_samples, 2, median)
quantile(sim_median_difs, c(0.005,0.995))

##  0.5% 99.5%
##    -8     0
```

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

1. Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
2. Analyser med en eller to stikprøver (en eller to grupper)

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

1. Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
2. Analyser med en eller to stikprøver (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige stikprøvefunktioner* (other features). dvs. disse metoder kan også anvendes for andre analyser end for middelværdier!

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

1. Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
2. Analyser med en eller to stikprøver (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige stikprøvefunktioner* (other features). dvs. disse metoder kan også anvendes for andre analyser end for middelværdier!

Hypotesetest også muligt

Vi kan udføre hypotesetest ved at kigge på konfidensintervallerne!

Dagsorden

1. Introduktion til simulation
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrapping
   - Introduktion til bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
4. Ikke-parametrisk bootstrapping
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
   - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver