02402: Introduktion til Statistik

Forelæsning 7: Simulationsbaseret statistik

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overview

1. Introduktion til simulation – hvad er det egentlig?
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrap
   - Introduktion til bootstrap
   - One-sample konfidensinterval for vilkårlig parameter
   - Two-sample konfidensinterval for en vilkårlig fordeling
4. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for vilkårlig feature
   - Two-sample konfidensinterval

Overview

Motivation

- Mange relevante beregningsstørrelser ("computed features") har komplicerede fordelinger. Det kunne f.eks. dreje sig om:
  - Medianen
  - Fraktiler generelt, eller måske IQR $= Q_3 - Q_1$
  - Variationskoefficienten (coefficient of variation)
  - Enhver ikke-lineær funktion af en eller flere inputvariable
  - (Spredningen)
- Yderligere er data måske ikke-normalfordelt, hvilket gør det svært selv for middelværdien.
- Vi kan håbe på CLTs "magi".

Motivation

- Mange relevante beregningsstørrelser ("computed features") har komplicerede fordelinger. Det kunne f.eks. dreje sig om:
  - Medianen
  - Fraktiler generelt, eller måske IQR $= Q_3 - Q_1$
  - Variationskoefficienten (coefficient of variation)
  - Enhver ikke-lineær funktion af en eller flere inputvariable
  - (Spredningen)
- Yderligere er data måske ikke-normalfordelt, hvilket gør det svært selv for middelværdien.
- Vi kan håbe på CLTs "magi".
- **Men**: Vi kan aldrig være helt sikre på om CLT er god nok i en given situation – simulering kan gøre os mere sikre!

Motivation

- Mange relevante beregningsstørrelser ("computed features") har komplicerede fordelinger. Det kunne f.eks. dreje sig om:
  - Medianen
  - Fraktiler generelt, eller måske IQR $= Q_3 - Q_1$
  - Variationskoefficienten (coefficient of variation)
  - Enhver ikke-lineær funktion af en eller flere inputvariable
  - (Spredningen)
- Yderligere er data måske ikke-normalfordelt, hvilket gør det svært selv for middelværdien.
- Vi kan håbe på CLTs "magi".
- **Men**: Vi kan aldrig være helt sikre på om CLT er god nok i en given situation – simulering kan gøre os mere sikre!
- **Kræver**: Brug af computer med simuleringsværktøj – R er et super værktøj til dette!

Hvad er simulation egentlig?

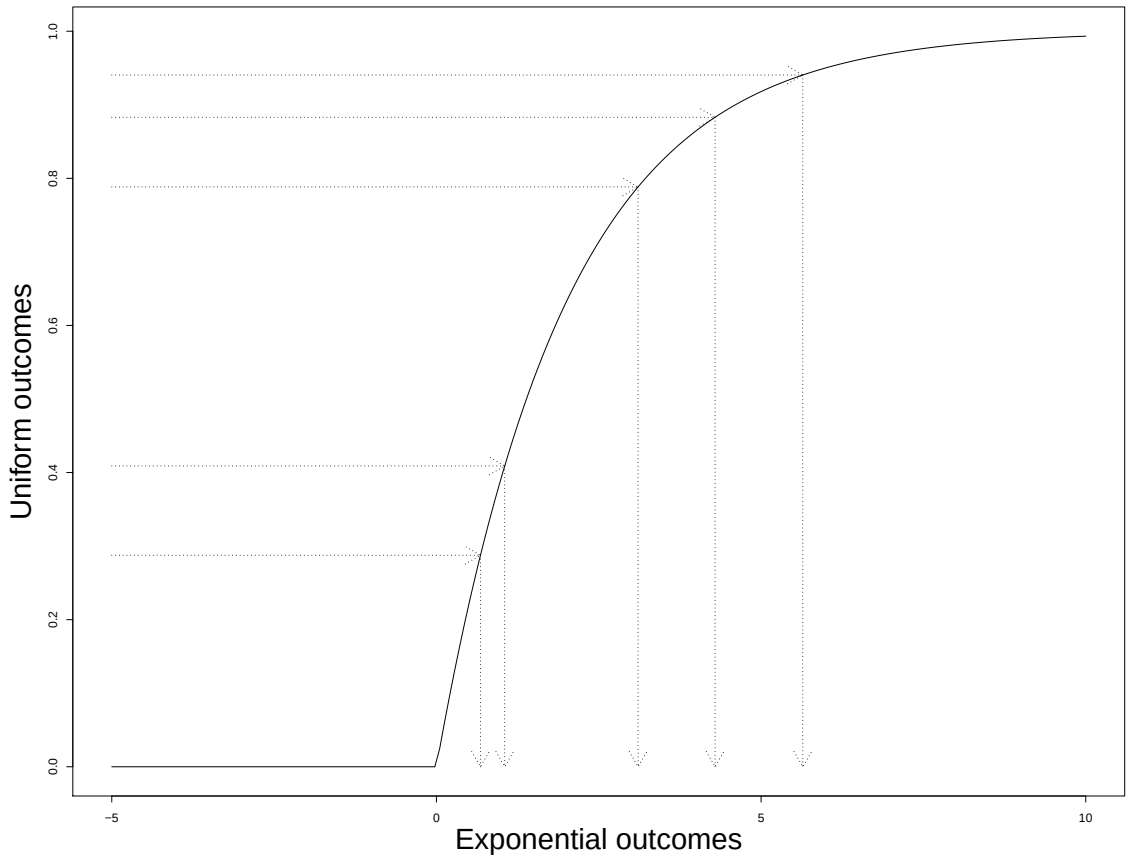
- (Pseudo)-tilfældige tal genereret af en computer.
- En *tilfældighedsgenerator* er en algoritme, der kan generere x_{i+1} ud fra x_i.
- Talfølgen heraf ser tilfældig ud.
- Kræver en "start" kaldet et *seed*.
- Faktisk simuleres den uniforme fordeling, hvorefter følgende resultat benyttes:

Theorem 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 med en normalfordeling $N(2, 0.01^2)$. Bredden af pladerne (i meter), Y, antages at kunne beskrives med en normalfordeling $N(3, 0.02^2)$. Man er interesseret i arealet, som jo så 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?
- (Sandsynligheden for andre hændelser?..)
- Generelt: Hvad er fordelingen for den stokastiske variabel A?

Eksempel: Areal af plader – løsning ved simulation

```
k = 10000 # Number of simulations
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y

mean(A)
```

```
[1] 6
```

```
var(A)
```

```
[1] 0.002458
```

```
mean(abs(A - 6) > 0.1)
```

```
[1] 0.0439
```

Overview

Fejlophobningslove (propagation of error)

Har brug for at finde:

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

Fejlophobningslove (propagation of error)

Har brug for at finde:

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

Vi kender allerede:

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

Fejlophobningslove (propagation of error)

Har brug for at finde:

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

Vi kender allerede:

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

Method 4.3: For ikke-lineære funktioner, hvis $X_1,\ldots,X_n$ uafhængige:

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

Eksempel: Areal af plader (fortsat)

Vi brugte simulation i den første del af eksemplet.

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

Eksempel: Areal af plader (fortsat)

Varianserne er:

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

Eksempel: Areal af plader (fortsat)

Varianserne er:

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

Funktionen og de 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:

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

Funktionen og de 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

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

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

hvor

$$f_j = f(X_1^{(j)}, \ldots, X_n^{(j)})$$

Eksempel: Areal af plader (fortsat)

Faktisk kan vi i dette eksempel finde variansen 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. Simulationstilgang.
2. Teoretisk udledning.
3. Den analytiske, men approksimative, *error propagation* metode

Eksempel: Areal af plader (fortsat)

Tre forskellige tilgange :

1. Simulationstilgang.
2. Teoretisk udledning.
3. Den analytiske, men approksimative, *error propagation* metode

Simulationstilgangen har nogle vigtigt 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 beskriver virkeligheden mere korrekt.
3. Afhænger ikke af en lineær approksimation (som error propagation) til den underliggende ikke-lineære funktion

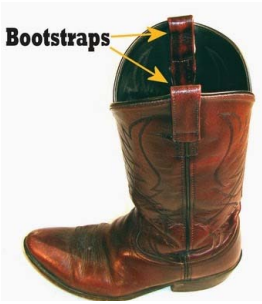
Overview

Bootstrapping

Bootstrap = Støvlestrop

Bootstrapping findes i to versioner:

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

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

Fra data estimerer vi

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

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

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

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

Fra data estimerer vi

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

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:

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

Fra data estimerer vi

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

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for μ?

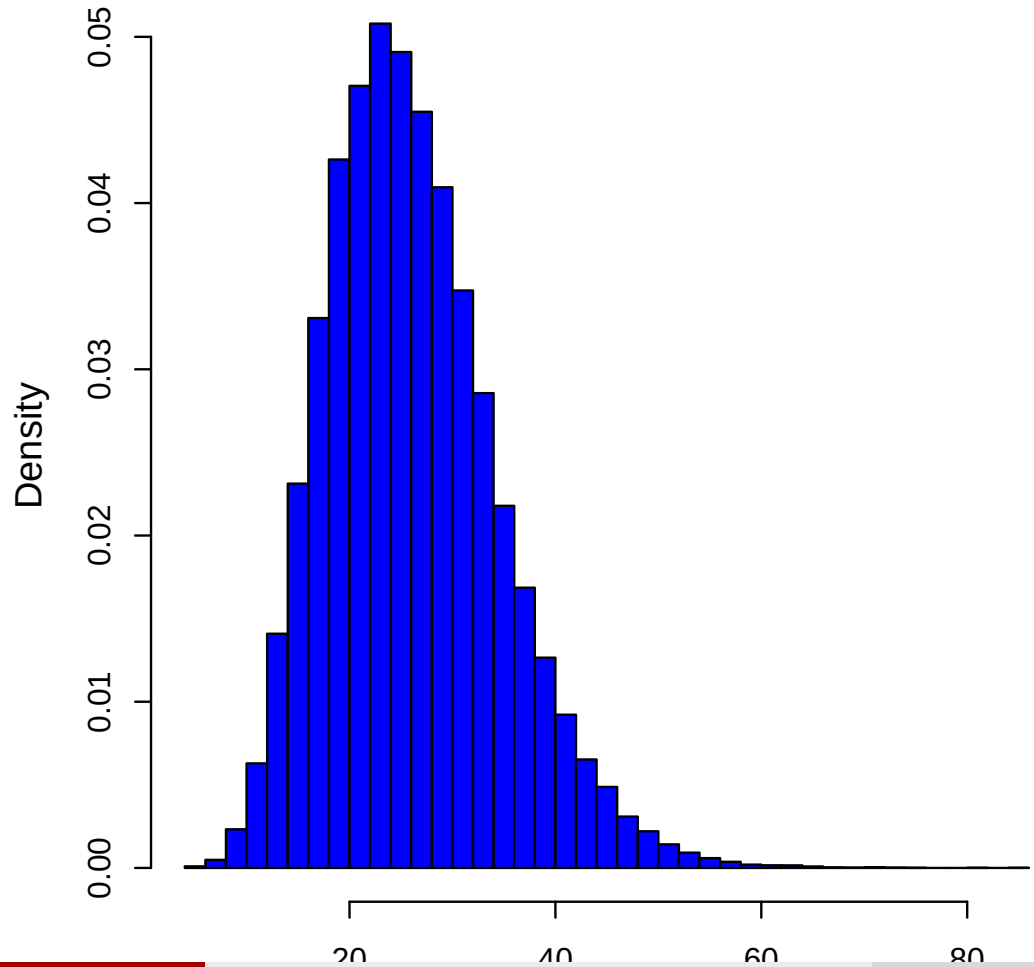
Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Compute the mean of the 10 simulated observations k times
sim_means <- apply(sim_samples, 2, mean)

# Find relevant quantiles of the k simulated means
quantile(sim_means, c(0.025, 0.975))

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

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
# Make histogram of simulated means
hist(sim_means, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Simulated means")
```



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:

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

Fra data estimerer vi:

$$\text{Median } = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

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

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

Fra data estimerer vi:

$$\text{Median } = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Vores 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:

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

Fra data estimerer vi:

$$\text{Median } = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Vores fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for medianen?

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

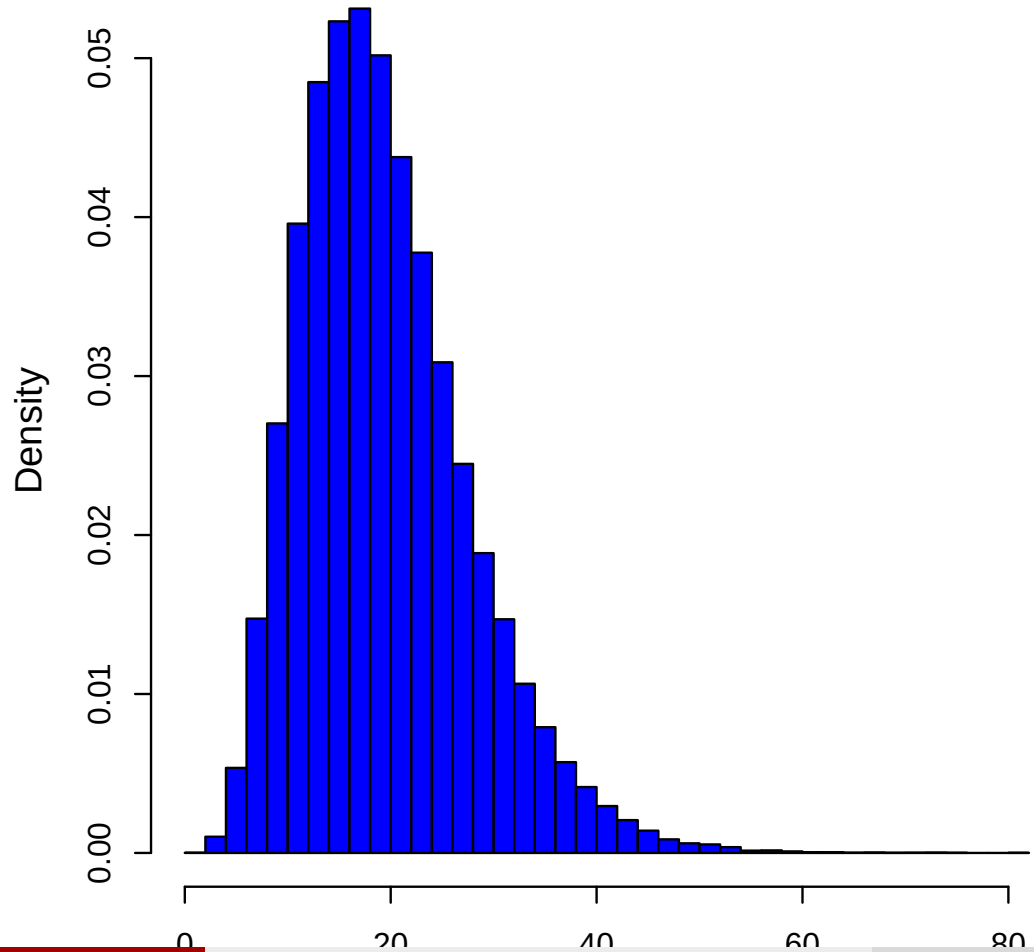
Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```r
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

# Compute the median of the 10 simulated observations k times
sim_medians <- apply(sim_samples, 2, median)

# Find relevant quantiles of the k simulated medians
quantile(sim_medians, c(0.025, 0.975))
```

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

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

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



Konfidensinterval for en vilkårlig feature (inkl. μ)

Method 4.7: Konfidensinterval for en vilkårlig feature θ ved parametrisk bootstrap

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, kald disse $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find $100(\alpha/2)\%$ og $100(1-\alpha/2)\%$ fraktilerne i $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $100(1-\alpha)\%$ 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% CI for Q_3 for en normalfordeling

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

# Define a Q3-function
Q3 <- function(x){ quantile(x, 0.75)}

# Set number of simulations
k <- 100000

# Simulate k samples of n = 10 normals with the 'right' mean and variance
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))

# Compute the Q3 of the n = 10 simulated observations k times
simQ3s <- apply(sim_samples, 2, Q3)

# Find the two relevant quantiles of the k simulated Q3s
quantile(simQ3s, c(0.005, 0.995))

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

Two-sample konfidensinterval for vilkårlig feature-sammenligning $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$)

Method 4.10: Two-sample konfidensinterval for vilkårlig feature-sammenligning $\theta_1 - \theta_2$ ved parametrisk bootstrap

Antag at vi har faktiske observationer $x_1, \ldots, x_n$, og at disse kommer fra sandsynlighedsfordelinger f_1 og f_2.

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 featuresne i hver af de k stikprøver: $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find $100(\alpha/2)\%$ og $100(1-\alpha/2)\%$ fraktilerne i disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så vi får et $100(1-\alpha)\%$ konfidensinterval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

Eksempel: Konfidensinterval for forskellen mellem to exponentielle middelværdier

```r
# Day 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)

# Day 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 to exponentielle middelværdier

```r
# Set number of simulations:
k <- 100000

# Simulate k samples of each n1 = 10 and n2 = 12 exponentials
# with the 'right' means

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

# Compute the difference between the simulated means k times
sim_dif_means <- apply(simX_samples, 2, mean) -
  apply(simY_samples, 2, mean)

# Find the relevant quantiles of the k simulated differences of means:
quantile(sim_dif_means, c(0.025, 0.975))

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

Parametrisk bootstrap – et overblik

Vi antager en eller anden fordeling!

To konfidensinterval-metodebokse blev givet:

	One-sample	Two-sample
For en vilkårlig feature	Metode 4.7	Metode 4.10

Overview

Ikke-parametrisk bootstrap – et overblik

Vi antager *ikke* noget om fordelinger!

To konfidensinterval-metodebokse gives:

	One-sample	Two-sample
For en vilkårlig feature	Method 4.15	Method 4.17

Eksempel: Kvinders cigaretforbrug

I et studie undersøgtes 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

Et parret t-test setup, *men* med tydeligvis ikke-normalfordelte data!

```
# 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)

# Compute differences
dif <- x1-x2
dif

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

# Compute average difference
mean(dif)

## [1] 5.273
```

Eksempel: Kvinders cigaretforbrug – 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 – ikke-parametriske bootstrap-resultater

Lad os finde 95% konfidensintervallet for *middelændringen* i cigaretforbrug.

```
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
```

One-sample konfidensinterval for en vilkårlig feature θ (inkl. μ)

Method 4.15: Confidence interval for any feature θ by non-parametric bootstrap

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 $100(\alpha/2)\%$ og $100(1-\alpha/2)\%$ fraktilerne for disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så får vi et $100(1-\alpha)\%$ konfidensinterrval: $\left[q^*_{\alpha/2},\, q^*_{1-\alpha/2}\right]$

Eksempel: Kvinders cigaretforbrug

Lad os finde 95% konfidensintervallet for *medianændringen* i cigaretforbrug 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 havde fået mælk fra flaske som barn, 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:

bottle	age	bottle	age	bottle	age
no	9	no	10	yes	16
yes	14	no	8	yes	14
yes	15	no	6	yes	9
no	10	yes	12	no	12
no	12	yes	13	yes	12
no	6	no	20		
yes	19	yes	13		

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

```
# Reading in 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% CI for mean difference by non-parametric bootstrap
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
```

Two-sample konfidensinterval for $\theta_1 - \theta_2$ (inkl. $\mu_1 - \mu_2$) ved ikke-parametrisk bootstrap

Method 4.17: Two-sample confidence interval for $\theta_1 - \theta_2$ by non-parametric bootstrap

Antag at vi har observationer $x_1, \ldots, x_n$ og $y_1, \ldots, y_n$.

1. Lav k gange tilfældig trækning af 2 grupper af n_1 og n_2 observationer fra de respektive stikprøver (med tilbagelægning).
2. Udregn forskellen mellem estimaterne i hver af de k stikprøver: $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ fraktilerne i disse, $q^*_{\alpha/2}$ og $q^*_{1-\alpha/2}$, så får vi et $100(1-\alpha)\%$ konfidensinterval: $\left[q^*_{\alpha/2},\ q^*_{1-\alpha/2}\right]$

Eksempel: Tandsundhed og spædbørns brug af flaske – et 99% konfidendinterval 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 metodebokse

1. Med eller uden fordeling (parametrisk eller ikke-parametrisk)
2. For one- eller two-sample analyse (en eller to grupper)

Bootstrapping – et overblik

Vi har set 4 ikke så forskellige metodebokse

1. Med eller uden fordeling (parametrisk eller ikke-parametrisk)
2. For one- eller two-sample analyse (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige beregningsstørrelser* (other features). Dvs: Disse metoder kan også anvendes for andre analyser end for middelværdier!

Bootstrapping – et overblik

Vi har set 4 ikke så forskellige metodebokse

1. Med eller uden fordeling (parametrisk eller ikke-parametrisk)
2. For one- eller two-sample analyse (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige beregningsstørrelser* (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!

Overview

1. Introduktion til simulation – hvad er det egentlig?
   - Eksempel: Areal af plader
2. Fejlophobningslove
3. Parametrisk bootstrap
   - Introduktion til bootstrap
   - One-sample konfidensinterval for vilkårlig parameter
   - Two-sample konfidensinterval for en vilkårlig fordeling
4. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for vilkårlig feature
   - Two-sample konfidensinterval