# Kursus 02323: Introducerende Statistik

# Forelæsning 7: Simuleringsbaseret statistik


Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk


Forår 2021

# Kapitel 4: Statistik ved simulering

## Simulering:

- Træk tilfældige værdier og beregn statistik mange gange
- Fejlforplantning (error propagation rules)
  *(F.eks. igennem ikke-lineær funktion)*
- Bootstrapping af konfidensintervaller:
  - Parametrisk *(Simuler mange udfald af stokastisk var.)*
  - Ikke-parametrisk *(Træk direkte fra data)*

## Specifikke bootstrap setups: (4 versioner af konfidensintervaller)

- Èn gruppe/stikprøve og to grupper/stikprøver data
- Parametrisk vs. ikke-parametrisk

# Chapter 4: Statistics by simulation

## Simulation:

- Draw random values and calculate the statistic many times
- Error propagation rules
  *(e.g. through a non-linear function)*
- Bootstrapping of confidence intervals:
  - Parametric *(Simulate many outcomes of random var.)*
  - Non-parametric *(Draw values directly from data)*

## Specific bootstrap set ups: (4 versions of confidence intervals)

- One-sample and Two-sample data
- Parametric vs. non-parametric

# Oversigt

1. Introduktion til simulation
   - Hvad er simulering egentlig?
2. Fejlophobningslove
3. Parametrisk bootstrap
   - Introduction to bootstrap
   - One-sample konfidensinterval for $\mu$
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller for en vilkårlig fordeling
4. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for $\mu$
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller

# Motivation

- Mange relevant statistikker ("computed features") har komplicerede samplingfordelinger:
  - Medianen
  - Fraktiler generelt, dvs. f.eks. også IQR$= Q_3 - Q_1$
  - Enhver ikke-lineær funktion af en eller flere input variable
  - ...
- Populations (og stikprøve) fordelingen kan være ikke-normal (komplicerer den statistiske teori)
- MEN: Nogle gange kan vi ikke være helt sikre på om det er godt nok - simulering kan hjælpe til at verificere!
- Kræver: Brug af computer - R er et super værktøj til dette!

# Anvendelser

*Stokastisk* simulering anvendes mange steder:

- Trafiksimulering
- Kø simulering, f.eks. call-center
- Agent baseret simulering, f.eks. evakuering og markeder
- ...

Generelt, kan bruges til at modellere komplekse stokastiske processer ved generere tilfældige udfald

# Hvad er simulering egentlig?

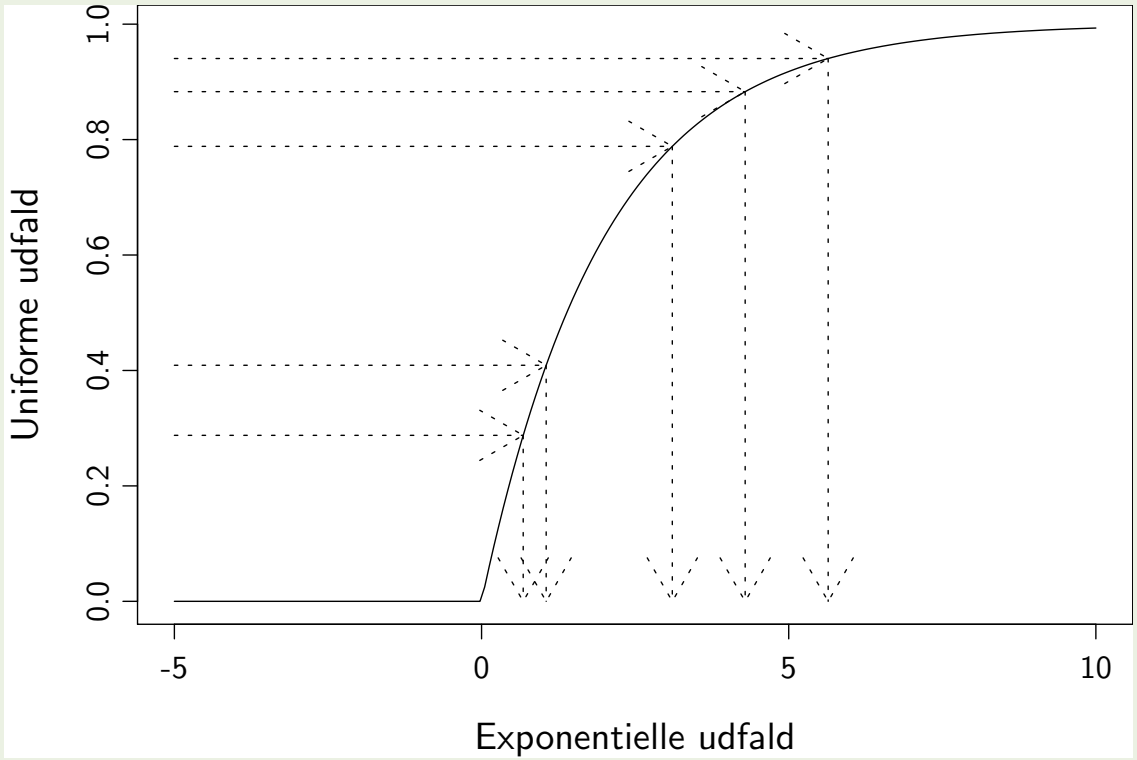
- (Pseudo)tilfældige tal genereret af en computer
- En tilfældighedsgenerator er en algoritme der kan generere $x_{i+1}$ ud fra $x_i$
- En sekvens af tal "ser tilfældige ud"
- Kræver en "start" - kaldet "seed" (Bruger typisk uret i computeren)
- Kapitel 2.6: Grundlæggende simuleres den uniforme fordeling, og så bruges:

Remark 2.51:

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

# Eksempel: Exponentialfordelingen med $\lambda = 0.5$

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



# Øvelse: Simuler på papir

## Tegn på papir (el. pc)

- Tegn eksponentiel fordelingsfunktion (altså cdf: $F(x)$)
- Træk uniform fordelte værdier og "put dem igennem" $F^{-1}(x)$

Gentag med normal fordelingsfunktion

# I praksis i R

De forskellige fordelinger er gjort klar til simulering:

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

# Eksempel: Simuler 100 binomialfordelte værdier

```r
## Simuler fra nogle fordelinger

## Sæt et seed for at få samme udfald hver gang
set.seed(123)

## 100 realisationer fra binomialfordelingen:
## Antal successer fra 25 trækninger med 0.2 sandsynlighed for succes
rbinom(n=100, size=25, prob=0.2)

## Exponential fordelte
hist(rexp(n=100, rate=2), prob=TRUE)
## Plot the theoretical pdf
xseq <- seq(0,10,len=1000)
lines(xseq, dexp(xseq, rate=2))
```

# Eksempel: Areal af plader

## En virksomhed producerer rektangulære plader

- Bredden $X$ af pladerne (i meter) antages at kunne beskrives med en normalfordeling $N(2, 0.01^2)$ og længden $Y$ af pladerne (i meter) antages at kunne beskrives med en normalfordeling $N(3, 0.02^2)$
- Man er interesseret i arealet, som jo så givet ved

$$A = XY$$

## Spørgsmål som kan stilles er f.eks.:

- Hvor ofte sådanne plader har et areal, der afviger mere end $0.1\text{m}^2$ fra de $6\text{m}^2$?
- Sandsynligheden for andre mulige hændelser?
- Generelt: Hvad er *sandsynlighedsfordelingen* for $A$?

# Eksempel: Areal af plader

- Hvor ofte sådanne plader har et areal, der afviger mere end $0.1\text{m}^2$ fra de $6\text{m}^2$?

Løsning ved simulering:

```
## Sæt et seed for at få samme udfald hver gang
set.seed(345)
## Antal simuleringer
k = 10000
## Simuler længderne af siderne
x = rnorm(k, 2, 0.01)
y = rnorm(k, 3, 0.02)
## Beregn arealet for hver simulering
area = x*y

## Beregn statistikker
## (stikprøve)Gennemsnit
mean(area)
## (stikprøve)spredning
sd(area)
## Fraktion af arealer afviger mere end 0.1
sum(abs(area-6) > 0.1) / k
```

# Fejlophobning

## Fejlophobning (error propagation)

- Vi har $n$ stokastiske variabler $X_1, X_2, \ldots, X_n$ og kender deres varianser $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$
- Og vil finde variansen af en ikke-lineær funktion af dem $\sigma^2_{f(X_1,\ldots,X_n)} = \mathrm{Var}\big(f(X_1,\ldots,X_n)\big)$

## 3 måder at løse på for ikke-lineær funktion $f()$:

- Simulation (Method 4.4)
- Lineær approksimation (Method 4.3)
- Teoretisk udledning

# Fejlophobning - ved simulering

## Method 4.4: Error propagation by simulation

Assume we have actual measurements $x_1, \ldots, x_n$ with known/assumed error variances $\sigma_1^2, \ldots, \sigma_n^2$.

1. Simulate $k$ outcomes of all $n$ measurements from assumed error distributions, e.g. $N(x_i, \sigma_i^2)$: $X_i^{(j)}, j = 1, \ldots, k$
2. Calculate the standard deviation directly as the observed standard deviation of the $k$ simulated values of $f$

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

where

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

# Eksempel: Varians af areal (Simulering)

Beregn variansen af arealet med simulering:

```
## Beregn variansen af arealet med simulering
## Antal simuleringer
k = 10000
## Simuler længderne af siderne
x = rnorm(k, 2, 0.01)
y = rnorm(k, 3, 0.02)
## Beregn arealet for hver simulering
area = x*y
## Beregn variansen af de simulerede arealer
var(area)
```

# Fejlophobning - ved lineær approksimation

Vi kender allerede regneregel for lineære funktioner (Theorem 2.56)

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

Method 4.3: for ikke-lineære funktioner

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

# Spørgsmål om hvilken metode til beregning af varians

(socrative.com - ROOM:PBAC)

Hvordan beregnes: $\text{Var}(2^2X - Y)$?

A: Som lineær funktion B: Som ikke-lineær funktion C: Ved ikke

Svar: A: Som *lineær funktion*

(Regneregel: Theorem 2.56 mean and variance of linear combinations)

Hvordan beregnes: $\text{Var}(X + Y^{-1})$?

A: Som lineær funktion B: Som *ikke-lineær* funktion C: Ved ikke

Svar: B: Som ikke-lineær funktion

(Brug simulation (Method 4.4) eller approximation (Method 4.3))

# Eksempel: Varians af areal (ikke-lineære fejlophobningslov)

- Vi har allerede brugt eksemplet med areal af en plade
- Nu spørger vi: Hvad er "fejlen" på $A = 2.00 \times 3.00 = 6.00$ m$^2$ fundet ved den ikke-lineære fejlophobningslov?
- Bredden $X \sim N(2, 0.01^2)$, længden $Y \sim N(3, 0.02^2)$

# Eksempel: Varians af areal (ikke-lineære fejlophobningslov)

Varianserne er

$$\sigma_X^2 = \mathrm{Var}(X) = 0.01^2 \quad \text{og} \quad \sigma_Y^2 = \mathrm{Var}(Y) = 0.02^2$$

Funktionen og de partielt afledte er

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

Så resultatet bliver

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

Method 4.3:

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

# Eksempel: Varians af areal (Teoretisk udledning)

Faktisk kan man finde variansen for $A = XY$ teoretisk

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

# Et summary

Igen: 3 måder til *beregning af varians af ikke-lineær funktion* *(husk, det er rent teoretisk, dvs. ingen data)*:

1. Simuleringsbaseret
2. Den analytiske, men approksimative, fejlophobningslov
3. Teoretisk udledning

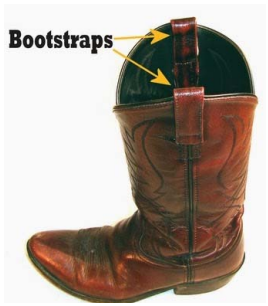
Simulering har en række fordele:

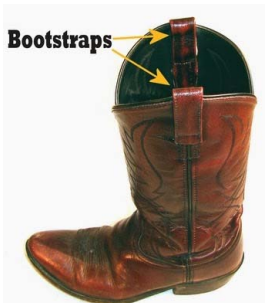
1. Simpel måde at beregne andre størrelser (kan være mere komplicerede, f.eks. udtryk sværere at differentiere)
2. Simpel måde at bruge andre fordelinger end normalfordelingen
3. Afhænger ikke af en lineær approksimation (som error propagation) til den underliggende ikke-lineære funktion

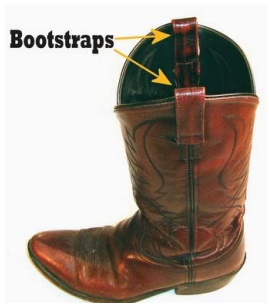
# Bootstrapping

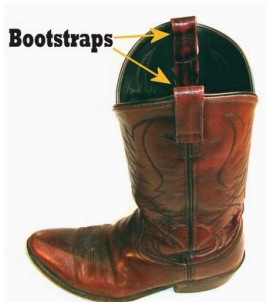
## Bootstrapping findes i to versioner:

1. Parametrisk bootstrap: Simuler gentagne samples fra en antagede (og estimerede) fordeling
2. Ikke-parametrisk bootstrap: Simuler gentagne samples direkte fra data









# Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

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

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

Vores fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling

Vi estimerer fra data (eksponential fordelingen har en parameter *raten*)

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

Hvad er konfidensintervallet for $\mu$?

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

# Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
## Beregn konfidensinterval for middelværdien med simulering

## Sæt seed hvis sammen resultat ønskes
set.seed(758)

## Første gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean1 <- mean(simSample)

## Anden gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean2 <- mean(simSample)

## Tredje gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean3 <- mean(simSample)

## Gør det 100000 gange!
```

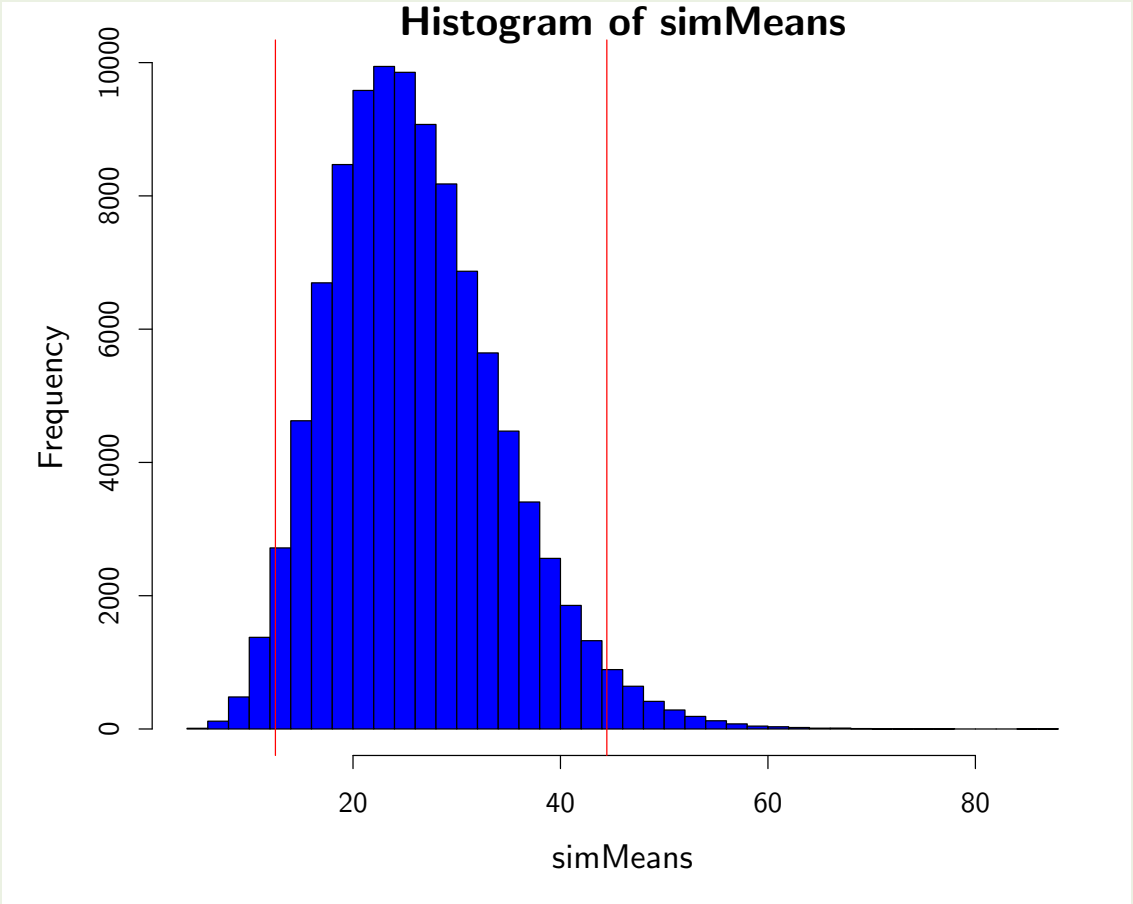
Alright, det må kunne gøres smartere!!

# Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
## Beregn konfidensinterval for middelværdien med simulering
## Set the number of simulations:
k <- 100000
set.seed(321)

## 1. Simulate 10 exponentials k times and keep in a matrix:
simSamples <- replicate(k, rexp(10, 1/26.08))

## 2. Compute the mean of the 10 simulated observations k times:
simMeans <- apply(simSamples, 2, mean)

## 3. Find the two relevant quantiles of the k simulated means:
quantile(simMeans, c(0.025, 0.975))

##  2.5% 97.5%
## 12.52 44.47
```

# Example: Konfidensinterval for middelværdien i en eksponentialfordeling

```
hist(simMeans, col="blue", nclass=30)
abline(v=quantile(simMeans, c(0.025, 0.975)), col="red")
```



# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

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

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

## Vores fordelingsantagelse

Ventetiderne kommer fra en eksponentialfordeling

## Vi estimerer fra data (som før)

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

## Hvad er konfidensintervallet for medianen?

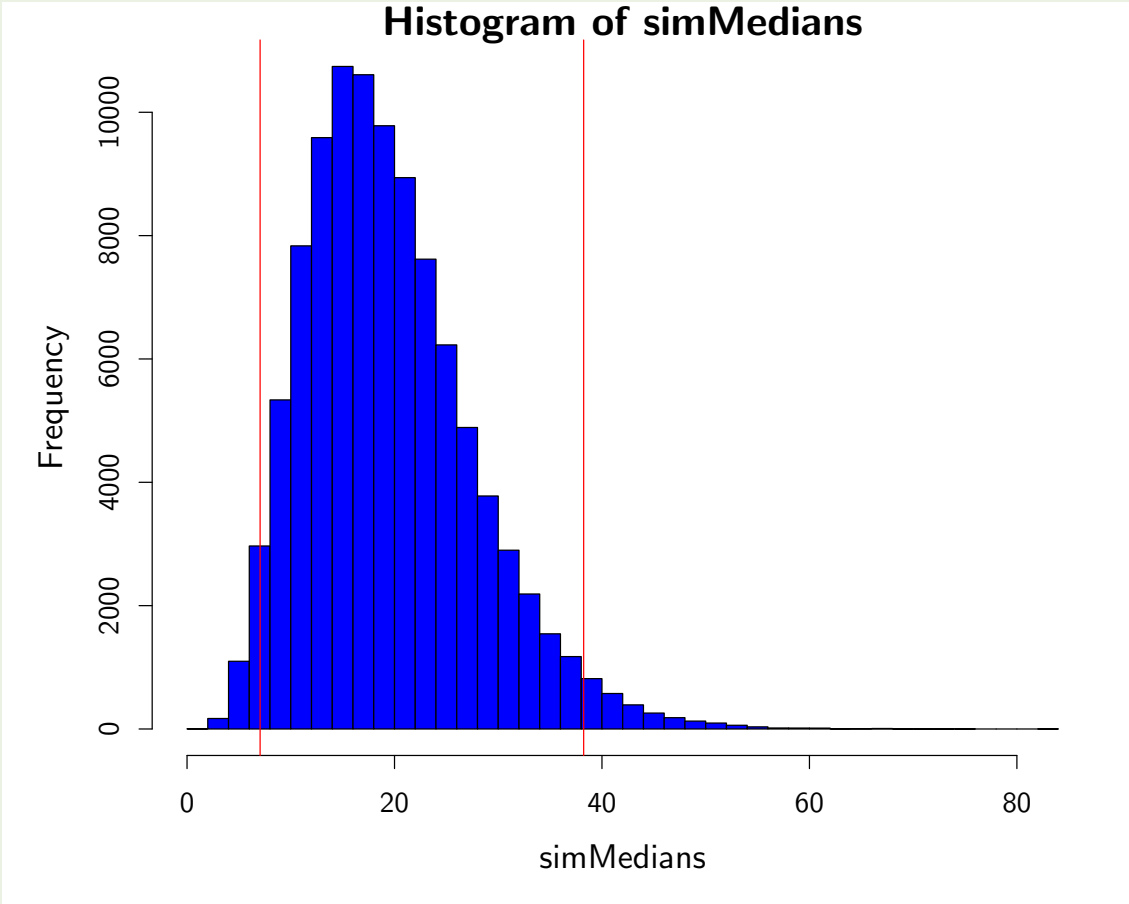
Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
## Beregn konfidensinterval for medianen med parametrisk bootstrapping

## Set the number of simulations:
k <- 100000
set.seed(543)

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

## 2. Compute the median of the n=10 simulated observations k times:
simMedians <- apply(simSamples, 2, median)

## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simMedians, c(0.025, 0.975))

##   2.5%  97.5%
##  7.045 38.235
```

# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
hist(simMedians, col="blue", nclass=30)
abline(v=quantile(simMedians, c(0.025, 0.975)), col="red")
```



# Konfidensinterval for en vilkårlig beregningsstørrelse

## Method 4.7: Confidence interval for any feature $\theta$ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and assume that they stem from some probability distribution with density $f$.

1. Simulate $k$ samples of $n$ observations from the assumed distribution $f$ where the mean[a] is set to $\bar{x}$
2. Calculate the statistic $\hat{\theta}$ in each of the $k$ samples $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval:
$$\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$$

[a] And otherwise chosen to match the data as good as possible: Some distributions have more than just a single mean related parameter, e.g. the normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data. Even more generally the approach would be to match the chosen distribution to the data by the so-called *maximum likelihood* approach

# Et andet eksempel: 99% konfidensinterval for $Q_3$ for en normalfordeling

```
## Konfidensinterval for den øvre kvartil (Q_3) i en normalfordeling
## Read in the heights data:
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Set the number of simulations:
k <- 100000
set.seed(543)
## 1. Simulate k samples of n=10 normals with the right mean and variance:
simSamples <- replicate(k, rnorm(n, mean(x), sd(x)))
## 2. Compute the Q3 of the n=10 simulated observations k times:
simQ3s <- apply(simSamples, 2, Q3)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simQ3s, c(0.005, 0.995))

##  0.5% 99.5%
## 172.9 198.0
```

# Two-sample konfidensinterval for en vilkårlig feature sammenligning $\theta_X - \theta_Y$ (inkl. $\mu_X - \mu_Y$)

## Method 4.10: Two-sample confidence interval for any feature comparison $\theta_X - \theta_Y$ by parametric bootstrap

Assume we have actual observations $x_X, \ldots, x_n$ and $y_X, \ldots, y_n$ and assume that they stem from some probability distributions with density $f_X$ and $f_Y$.

1. Simulate $k$ sets of 2 samples of $n_X$ and $n_Y$ observations from the assumed distributions setting the means [a] to $\hat{\mu}_X = \bar{x}$ and $\hat{\mu}_Y = \bar{y}$, respectively
2. Calculate the difference between the features in each of the $k$ samples $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

[a] As before

# Eksempel: Konfidensinterval for the forskellen mellem to exponentielle middelværdier

```
## Konfidensinterval for the forskellen mellem to exponentielle middelværdier
## 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)
## 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)
## Keep sample sizes
n1 <- length(x)
n2 <- length(y)
```

# Eksempel: Konfidensinterval for the forskellen mellem to exponentielle middelværdier

```
## Konfidensinterval for the forskellen mellem to exponentielle middelværdier
## Set the number of simulations:
k <- 100000
set.seed(987)
## 1. Simulate k samples of each n1=10 and n2=12
## exponentials with the right means:
simxSamples <- replicate(k, rexp(n1, 1/mean(x)))
simySamples <- replicate(k, rexp(n2, 1/mean(y)))
## 2. Compute the difference between the simulated
## means k times:
simDifMeans <- apply(simxSamples, 2, mean) -
                    apply(simySamples, 2, mean)
## 3. Find the two relevant quantiles of the
## k simulated differences of means:
quantile(simDifMeans, c(0.025, 0.975))

##   2.5%  97.5%
## -40.53  13.94
```

# Parametrisk bootstrap - et overblik

**Vi antager en fordeling!**

Der er parametre i en fordeling, derfor *parametrisk bootstrap*

**To konfidensinterval-metodeboxe blev givet:**

| | One-sample | Two-sample |
|---|---|---|
| For any feature | Method 4.7 | Method 4.10 |

# Ikke-parametrisk bootstrap - et overblik

**Vi antager IKKE noget om nogen fordelinger!**

Altså der er ingen parametre, derfor *ikke-parametrisk bootstrap*

**To konfidensinterval-metodeboxe bliver givet:**

| | One-sample | Two-sample |
|---|---|---|
| For any feature | Method 4.15 | Method 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 før og efter! Er der sket nogen ændring i gennemsnitsforbruget!

# Eksempel: Kvinders cigaretforbrug

Et parret $t$-test setup, MEN med tydeligvis ikke-normale data!

```
## Parret test af middelværdiforskel med ikke-parametrisk bootstrapping
## Input the two cigaret use samples
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)
## Calculate the difference
dif <- x1 - x2
dif

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

## And the sample mean
mean(dif)

## [1] 5.273
```

# Eksempel: Kvinders cigaretforbrug - bootstrapping

```
##################################
## Resample several times
sample(dif, replace = TRUE)

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

sample(dif, replace = TRUE)

##  [1]  5  9  6 -1 -2 -2 -2  0 13  9  6

sample(dif, replace = TRUE)

##  [1]  7 -1  0 22  5 -1 -1 -4  5 -4 -2

sample(dif, replace = TRUE)

##  [1] -1 -4  6 22  9  5 -2  9 -2  9  9
```

# Eksempel: Kvinders cigaretforbrug - de ikke-parametriske bootstrap resultater:

```
## Resample calculate mean statistic many time, and find 95% confidence interval
k = 100000
simSamples = replicate(k, sample(dif, replace = TRUE))
## Take the mean for every resample
simMeans = apply(simSamples, 2, mean)
## Take the two quantiles to get the confidence interval
quantile(simMeans, c(0.025,0.975))

##  2.5% 97.5%
## 1.273 9.818
```

# One-sample konfidensinterval for en vilkårlig feature $\theta$ (inkl. $\mu$)

**Method 4.15: Confidence interval for any feature $\theta$ by non-parametric bootstrap**

Assume we have actual observations $x_1, \ldots, x_n$.

1. Simulate $k$ samples of size $n$ by randomly sampling among the available data (with replacement)
2. Calculate the statistic $\hat{\theta}$ in each of the $k$ samples $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval:
$$\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$$

# Eksempel: Kvinders cigaretforbrug

Lad os finde 95% konfidensintervallet for ændringen af median cigaretforbruget

```
## Konfidensintervallet for ændringen af median cigaretforbruget
## Simulate many times
k = 100000
simsamples = replicate(k, sample(dif, replace = TRUE))
## Take the median for each resample
simMedians = apply(simsamples, 2, median)
## Take the two quantiles to get the confidence interval
quantile(simMedians, c(0.025,0.975))

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

# Eksempel: Tandsundhed og flaskebrug

I et studie ville man undersøge, 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.

| flaske | alder | flaske | alder | flaske | alder |
|---|---|---|---|---|---|
| nej | 9 | nej | 10 | ja | 16 |
| ja | 14 | nej | 8 | ja | 14 |
| ja | 15 | nej | 6 | ja | 9 |
| nej | 10 | ja | 12 | nej | 12 |
| nej | 12 | ja | 13 | ja | 12 |
| nej | 6 | nej | 20 | | |
| ja | 19 | ja | 13 | | |

Find konfidensintervallet for forskellen!

# Eksempel: Tandsundhed og flaskebrug - et 95% konfidensinterval for $\mu_X - \mu_Y$

```
## Tandsundhed og flaskebrug. Konfidensinterval for forskel i middelværdi
## Reading in no group:
x <- c(9,10,12,6,10,8,6,20,12)
## Reading in yes group:
y <- c(14,15,19,12,13,13,16,14,9,12)

## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Take the mean for each time and subtract
simmeandifs <- apply(simxSamples, 2, mean) -
                          apply(simySamples, 2, mean)
## Take the two quantiles to get the confidence interval
quantile(simmeandifs, c(0.025,0.975))

##    2.5%   97.5%
## -6.2000 -0.1444
```

# Two-sample konfidensinterval for $\theta_X - \theta_Y$ (inkl. $\mu_X - \mu_Y$) med ikke-parametrisk bootstrap

**Method Method 4.17: Two-sample confidence interval for $\theta_X - \theta_Y$ by non-parametric bootstrap**

Assume we have actual observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

1. Simulate $k$ sets of 2 samples of $n_X$ and $n_Y$ observations from the respective groups (with replacement)
2. Calculate the difference between the features in each of the $k$ samples $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval:
$$\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$$

# Eksempel: Tandsundhed og flaskebrug - et 99% confidence interval for median-forskellen

```
## Tandsundhed og flaskebrug. Konfidensinterval for forskel i median
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Take the difference in medians
simmedianDiffs <- apply(simxSamples, 2, median)-
                        apply(simySamples, 2, median)
## Take the two quantiles to get the confidence interval
quantile(simmedianDiffs, c(0.005,0.995))

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

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3Diffs <- apply(simxSamples, 2, Q3) - apply(simySamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3Diffs, c(0.005, 0.995))
```

## Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

## Svar D

Resampling af x og y, men ingen fordelingsantagelse.
Derfor to stikprøver og non-parametric.

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3 <- apply(simxSamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3, c(0.005, 0.995))
```

## Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

## Svar C

Kun resampling af x, og ingen fordelingsantagelse.
Derfor kun een stikprøve og non-parametric.

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, rnorm(length(x), mean(x), sd(x)))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3 <- apply(simxSamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3, c(0.005, 0.995))
```

## Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

## Svar A

Kun resampling af x og fordelingsantagelse (normalfordeling).
Derfor kun een stikprøve og parametric.

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, rnorm(length(x), mean(x), sd(x)))
simySamples <- replicate(k, rnorm(length(y), mean(y), sd(y)))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3Diffs <- apply(simxSamples, 2, Q3) - apply(simySamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3Diffs, c(0.005, 0.995))
```

## Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

## Svar B

Resampling af x og y, samt fordelingsantagelse (normal). Derfor to stikprøver og parametric.

# Bootstrapping - et overblik

## Vi har fået 4 ret ens forskellige metode-bokse:

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

## Bemærk:

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

## Hypotesetest også muligt

Vi kan udføre hypotese test ved at kigge på konfidensintervallerne!