Kursus 02402/02323 Introducerende Statistik

Forelæsning 7: Simuleringsbaseret statistik

Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Bygning 324, Rum 220
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: perbb@dtu.dk

DTU Compute
Department of Applied Mathematics and Computer Science

Oversigt

1. Introduktion til simulation
   - Hvad er simulering egentlig?
   - Eksempel, Areal af plader
   - Fejlophobningslove
2. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller for en vilkårlig fordeling
3. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller

DTU Compute
Department of Applied Mathematics and Computer Science

Motivation

- Mange relevant beregningsstørrelser ("computed features") har komplicerede samplingfordelinger:
  - Et trimmed gennemsnit
  - Medianen
  - Fraktiler generelt, dvs. f.eks. også IQR$= Q_3 - Q_1$
  - Variationkoefficienten
  - Enhver ikke-lineær function af en eller flere input variable
  - (Spredningen)
- Data/populations fordelingen kan være ikke-normal, hvilket komplicerer den statistiske teori for selv en simpel gennemsnitsberegning
- Vi kan HÅBE på the magic of CLT (Central Limit Theorem)
- MEN men: Vi kan aldrig være helt sikre på om det er godt nok - simulering kan gøre os mere sikre!
- Kræver: Brug af computer - R er et super værktøj til dette!

DTU Compute
Department of Applied Mathematics and Computer Science

Hvad er simulering egentlig?

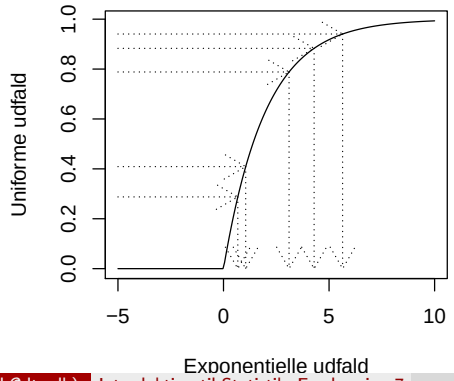
- (Pseudo)tilfældige tal genereret af en computer
- En tilfældighedsgenerator er en algoritme der kan generere x_{i+1} ud fra x_i
- En sekvens af tal "ser tilfældige ud"
- Kræver en "start" - kaldet "seed" .(Bruger typisk uret i computeren)
- Grundlæggende simuleres den uniforme fordeling, og så bruges:

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

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel: Exponentialfordelingen med $\lambda = 0.5$:

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



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`	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 $N(2, 0.01^2)$ og 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 sådanne plader har et areal, der afviger mere end $0.1m^2$ fra de $6m^2$?
- Sandsynligheden for andre mulige hændelser?
- Generelt: Hvad er sandsynlighedsfordelingen for A?

Eksempel, Løsning ved simulering

```
set.seed(345)
k = 10000 # Number of simulations
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y
```

```
mean(A)

## [1] 5.9995

sd(A)

## [1] 0.049575

mean(abs(A-6)>0.1)

## [1] 0.0439
```

Fejlophobningslove

Har brug for at finde:

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

Vi kender allerede:

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

Method 4.6: for ikke-lineære funktioner:

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

DTU Compute
Department of Applied Mathematics and Computer Science

Fejlophobning - ved simulering

Method 4.7: 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, fortsat

Vi har allerede brugt simulerings-metoden i første del af eksemplet. To konkrete målinger for X og Y, er givet: $x = 2.05m$ og $y = 2.99m$. Hvad er "fejlen" på $A = 2.00 \times 3.00 = 6.00$ fundet ved den ikke-lineære fejlophobningslov?

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel, fortsat

Varianserne er:

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

Funktionen og de afledede er:

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

Så resultatet bliver:

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

Eksempel 1, fortsat

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

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

DTU Compute
Department of Applied Mathematics and Computer Science

Areal-eksempel – et summary

Tre forskellige approaches:

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

The simulation approach has a number of crucial advantages:

1. It offers a simple tool to compute many other quantities than just the standard deviation (the theoretical derivations of such other quantities could be much more complicated than what was shown for the variance here)
2. It offers a simple tool to use any other distribution than the normal, if we believe such better reflect reality.
3. It does not rely on any linear approximations of the true non-linear relations.

Bootstrapping

Bootstrapping findes i to versioner:

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

DTU Compute
Department of Applied Mathematics and Computer Science

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 fra data:

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

Vores fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling

Hvad er konfidensintervallet for μ?

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

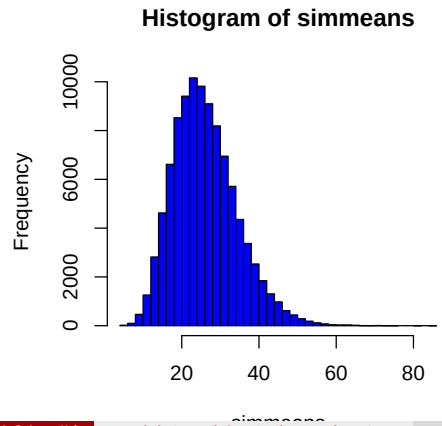
Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
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%
## 12.587 44.627
```


Example: Konfidensinterval for middelværdien i en eksponentialfordeling

```
hist(simmeans, col="blue", nclass=30)
```



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 fra data:

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

Vores fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling

Hvad er konfidensintervallet for medianen?

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

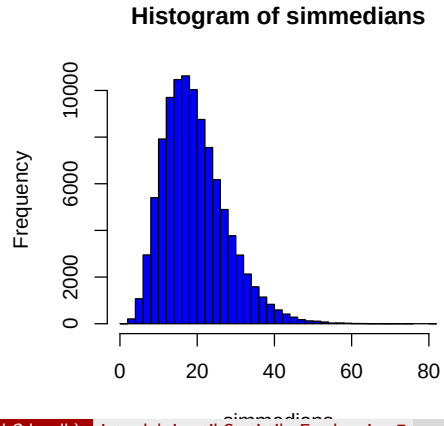
Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
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%
##  7.038 38.465
```

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
hist(simmedians, col="blue", nclass=30)
```



DTU Compute
Department of Applied Mathematics and Computer Science

Konfidensinterval for en vilkårlig beregningsstørrelse

Method 4.10: Confidence interval for any feature θ 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

```
## Read in the 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 the number of simulations:
k <- 100000
## 1. Simulate k samples of n=10 normals with the right mean and variance:
set.seed(9876.543)
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%
## 172.82 198.00
```


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

Method 4.13: Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ and assume that they stem from some probability distributions with density f_1 and f_2.

1. Simulate k sets of 2 samples of n_1 and n_2 observations from the assumed distributions setting the means [a] to $\hat{\mu}_1 = \bar{x}$ and $\hat{\mu}_2 = \bar{y}$, 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: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

Eksempel: 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)
n1 <- length(x)
n2 <- length(y)
```


Eksempel: Konfidensinterval for the forskellen mellem to exponentielle middelværdier

```
## Set the number of simulations:
k <- 100000
## 1. Simulate k samples of each n1=10 and n2=12
## exponentials with the right means:
set.seed(9876.543)
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%
## -40.735  14.117
```


Parametrisk bootstrap - et overblik

Vi antager en eller anden fordeling!

To konfidensinterval-metodeboxe blev givet:

	One-sample	Two-sample
For any feature	Method 4.10	Method 4.13

Ikke-parametrisk bootstrap - et overblik

Vi antager IKKE noget om nogen fordelinger!

To konfidensinterval-metodeboxe bliver givet:

	One-sample	Two-sample
For any feature	Method 4.18	Method 4.20

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!

```
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)
dif <- x1-x2
dif

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

mean(dif)

## [1] 5.2727
```


Eksempel: Kvinders cigaretforbrug - bootstrapping

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

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


Eksempel: Kvinders cigaretforbrug - de ikke-parametrisk bootstrap resultater:

```
k = 100000

simsamples = replicate(k, sample(dif, replace = TRUE))
simmeans = apply(simsamples, 2, mean)
quantile(simmeans, c(0.025,0.975))

##   2.5%  97.5%
## 1.3636 9.8182
```

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

Method 4.18: Confidence interval for any feature θ 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

```
k = 100000
simsamples = replicate(k, sample(dif, replace = TRUE))
simmedians = apply(simsamples, 2, median)
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 $\mu_1 - \mu_2$

```
## 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)

k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmeandifs <- apply(simxsamples, 2, mean)-
                          apply(simysamples, 2, mean)
quantile(simmeandifs, c(0.025,0.975))

##     2.5%    97.5%
## -6.23333 -0.14444
```

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

Method 4.20: Two-sample confidence interval for $\theta_1 - \theta_2$ 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_1 and n_2 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: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

DTU Compute
Department of Applied Mathematics and Computer Science

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

```
k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmediandifs <- apply(simxsamples, 2, median)-
                         apply(simysamples, 2, median)
quantile(simmediandifs, c(0.005,0.995))

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

DTU Compute
Department of Applied Mathematics and Computer Science

Bootstrapping - et overblik

Vi har fået 4 ikke så 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!

DTU Compute
Department of Applied Mathematics and Computer Science

Oversigt

1. Introduktion til simulation
   - Hvad er simulering egentlig?
   - Eksempel, Areal af plader
   - Fejlophobningslove
2. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller for en vilkårlig fordeling
3. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller

DTU Compute
Department of Applied Mathematics and Computer Science