Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 7: Simuleringsbaseret statistik

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

1. Introduktion til simulation
   - Hvad er simulering egentlig?
2. Fejlophobningslove
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller for en vilkårlig fordeling
4. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for en vilkårlig størrelse
   - One-sample konfidensinterval for μ
   - Two-sample konfidensintervaller
5. Første del af kurset

Oversigt

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!

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

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

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.

Eksempel 5, 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?

Fejlophobning - ved simulering

Method 4.4: Error propagation by simulation

Assume that $X_1, \ldots, X_n$ follow som distribution e.g. $N(\mu_i, \sigma_i)$

1. Simulate k outcomes of all n random variables from the assumed distributions.
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

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

Oversigt

Fejlophobningslove

Antag at X_i er stokastiske variable med $E(X_i) = \mu_i$ og $V(X_i) = \sigma_i^2$ og $Cov(X_i, X_j) = \sigma_{ij}$

Har brug for at finde:

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

(Generalisering af) Method 4.3: for ikke-lineære funktioner:

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

Hvor de afledte af f evalueres i $\mu_1, ..., \mu_n$.
Læg mærke til at hvis $X_1, ..., X_n$ uafhængige fås (Method 4.3)

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

Fejlophobning ved Taylorudvikling

Lad $f(X_1, \ldots, X_n)$ være en ikke-lineær funktion af de stokastiske variable $X_1, ..., X_n$.
Vi tager nu første ordens Taylor udviklingen omkring
$\mu = [E(X_1), ..., E(X_n)]^T$

$$\begin{aligned} f(x_1, \ldots, x_n) =& f(\mu) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\bigg|_{\mathbf{x}=\mu} (x_i - \mu_i) + HOT \\ \approx& f(\mu) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\bigg|_{\mathbf{x}=\mu} (x_i - \mu_i) \end{aligned}$$

Dermed har vi

$$f(X_1, \ldots, X_n) \approx f(\mu) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\bigg|_{\mathbf{x}=\mu} (X_i - \mu_i)$$

Fejlophobning ved Taylorudvikling - Fortsat

Tag nu forventningsværdien

$$\begin{aligned} E[f(X_1, \ldots, X_n)] \approx & E[f(\mu)] + \sum_{i=1}^{n} \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}=\mu} E[X_i - \mu_i] \\ = & f(\boldsymbol{\mu}) \end{aligned}$$

Dermed har vi

$$f(\mathbf{X}) - E[f(\mathbf{X})] \approx \sum_{i=1}^{n} \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}=\mu} (X_i - \mu_i)$$

Vi kan finde variansen af $f(\mathbf{X})$, $V[f(\mathbf{X})] = E[(f(\mathbf{X}) - E[f(\mathbf{X})])^2]$

Fejlophobning ved Taylorudvikling - Fortsat

Vi kan nu tage variansen af $f(\mathbf{X})$, $V[f(\mathbf{X})] = E[(f(\mathbf{X}) - E[f(\mathbf{X})])^2]$

$$
\begin{aligned}
V[f(\mathbf{X})] \approx & E\left[\left(\sum_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}=\boldsymbol{\mu}} (X_i - \mu_i)\right)^2\right] \\
= & \sum_{i=1}^{n} \left(\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}=\mu}\right)^2 E\left[(X_i - \mu_i)^2\right] + \\
& \sum_{i \neq j} \left(\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\right)\Bigg|_{\mathbf{x}=\mu} E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\
= & \sum_{i=1}^{n} \left(\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}=\mu}\right)^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} \left(\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\right)\Bigg|_{\mathbf{x}=\mu} \sigma_{ij}
\end{aligned}
$$

Varians af BMI

Body Mass Index er defineret ved

$$BMI = \frac{V}{H^2}$$

hvor V er en persons vægt (kg), mens H er personens højde (m). Antag nu at vi kender en populations middelhøjde og middelvægt (μ_V, og μ_H), samt varians- kovariansmatricen for vægt og højde (σ_V, σ_H og σ_{VH}).

Varians af BMI

Find (en approksimation til) middel BMI og varians for denne

$$E[BMI] \approx \frac{\mu_V}{\mu_H^2}$$

$$\begin{aligned} V[BMI] &\approx \left(\frac{\partial BMI}{\partial V}\right)^2 \sigma_V^2 + \left(\frac{\partial BMI}{\partial H}\right)^2 \sigma_H^2 + 2\left(\frac{\partial BMI}{\partial H}\frac{\partial BMI}{\partial V}\right)\sigma_{VH} \\ &= \left(\frac{1}{\mu_H^2}\right)^2 \sigma_V^2 + \left(-2\frac{\mu_V}{\mu_H^3}\right)^2 \sigma_H^2 + 2\left(-2\frac{\mu_V}{\mu_H^3}\frac{1}{\mu_H^2}\right)\sigma_{VH} \\ &= \frac{\sigma_V^2}{\mu_H^4} + \frac{4\mu_V^2\sigma_H^2}{\mu_H^6} - \frac{4\mu_V\sigma_{VH}}{\mu_H^5} \\ &= \frac{1}{\mu_H^4}\left(\sigma_V^2 + \frac{4\mu_V^2\sigma_H^2}{\mu_H^2} - \frac{4\mu_V\sigma_{VH}}{\mu_H}\right) \end{aligned}$$

Eksempel 1, 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:

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

Eksempel 1, fortsat

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

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

Areal-eksempel – et summary

Tre forskellige approaches:

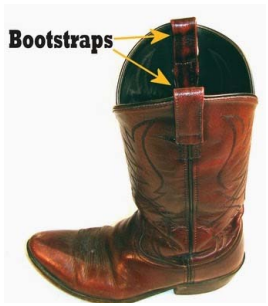
1. Simuleringsbaseret
2. Teoretisk udledning
3. Den analytiske, men approksimative, error propagation metode

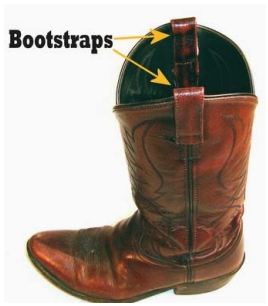
Oversigt

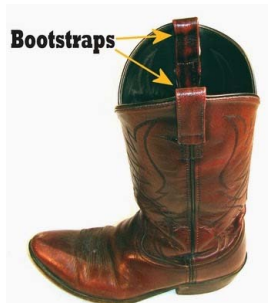
Bootstrapping

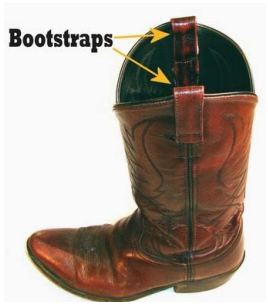
Bootstrapping findes i to versioner:

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









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: $\hat{\lambda} = 1/26.08 = 0.03834356$

Antagelse:

$X_i \sim Exp(\lambda)$

Hvad er konfidensintervallet for μ?

Lille stikprøve og ikke normalfordelt, dvs. vores antagelser holder ikke.

Konfidensinterval for en vilkårlig beregningsstørrelse

Method 4.7: 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 in some cases more quantities e.g. mean and variance (normal distribution)

Og fodnoten...

And otherwise chosen to match the data as good as possible:

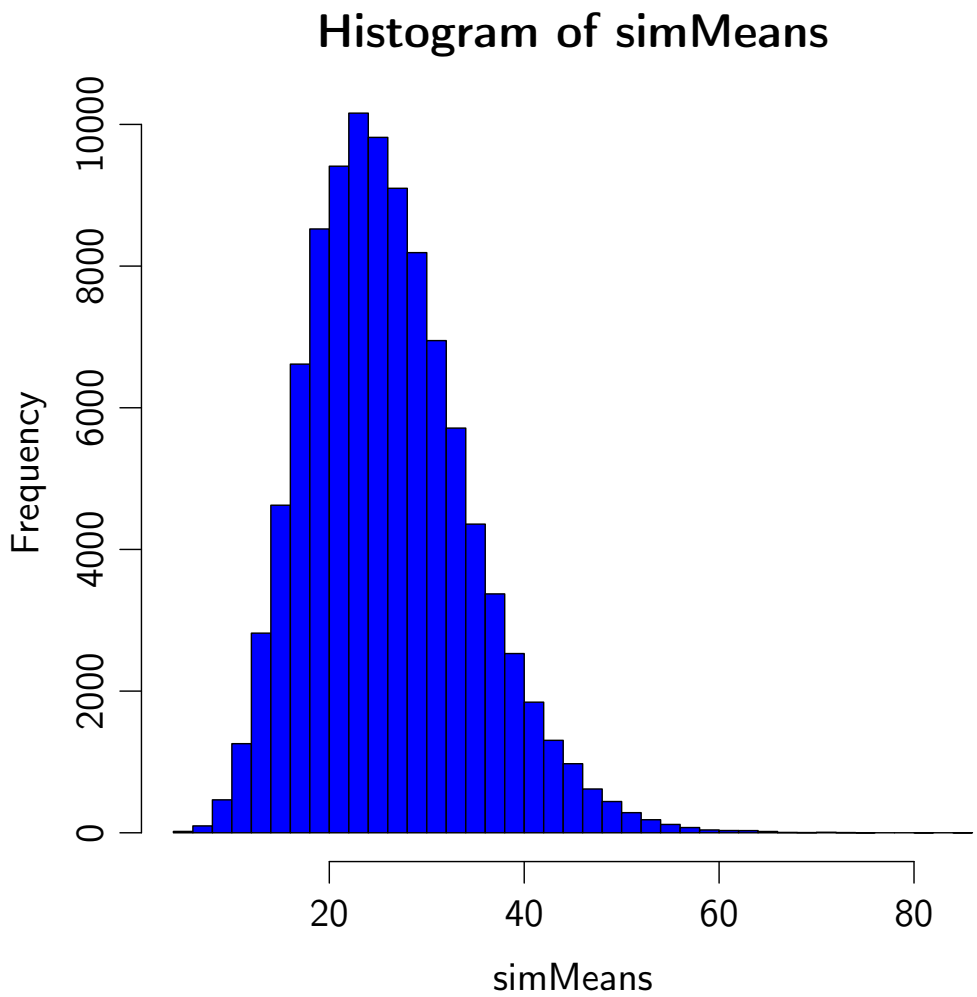
- Some distributions have more than just a single mean related parameter.
- The normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data.
- Even more generally the approach would be to match the chosen distribution to the data by the so-called *maximum likelihood* approach

Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
##################################
## Beregn konfidensinterval for middelværdien med simulering
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876)
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



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:

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

Antagelse:

$X_i \sim Exp(\lambda)$

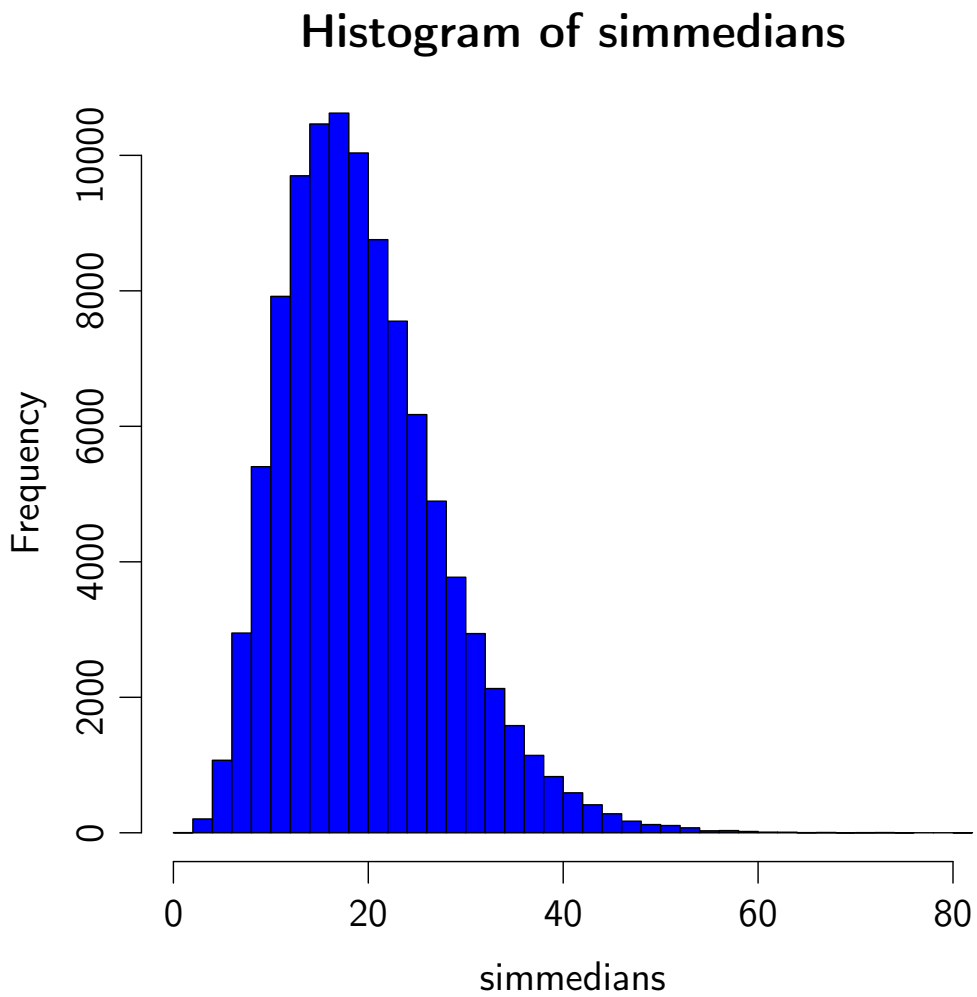
Hvad er konfidensintervallet for **medianen**?

Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
## Beregn konfidensinterval for middelværdien med parametrisk bootst
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876)
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



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)
## Set the number of simulations:
k <- 100000
## 1. Simulate k samples of n=10 normals with the right mean and var
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, quantile, prob = 0.75)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simQ3s, c(0.005, 0.995))

##    0.5%   99.5%
## 172.88 198.03
```

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

Method 4.10: Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ 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

```
## Konfidensinterval for the forskellen mellem to exponentielle midd
## 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)
```

Parametrisk bootstrap - et overblik

Vi antager en eller anden fordeling!

To konfidensinterval-metodeboxe blev givet:

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

Parret / ikke parret situationer behandles/bestemmes som for den sædvanlige t-test.

Oversigt

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 4.15	Method 4.17

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

Method 4.15: 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

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 bootstrapp
## 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] 5.2727
```

Eksempel: Kvinders cigaretforbrug - bootstrapping

```
## Resample from the dif sample
t(replicate(5, sample(dif, replace = TRUE)))
```

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

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

```
## Resample many time
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%
## 1.3636 9.8182
```

Eksempel: Kvinders cigaretforbrug

Lad os finde 95% 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!

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

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

Assume we have actual observations $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$.

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: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

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!

Oversigt

Nøgletal (summary statistics)

Vi anvender en række *nøgletal* (eller statistikker) for at opsummere og beskrive data (og stokastiske variable)

- **Gennemsnit:** tyngdepunkt eller centrering
- **Median:** tyngdepunkt eller centrering
- **Varians:** variation
- **Spredning:** variation (samme enhed som data)
- **Variations koefficient:** variationen i data (enhedsløs)
- **Kovarians:** samvariation mellem datasæt
- **Korrelation:** samvariation mellem datasæt (enhedsløs)
- **Fraktiler:** siger noget om fordelingen af data

Grafisk Fremstilling

- Histogram
- Empirisk kumulativ tæthedsfunktion
- Boxplot
- Scatterplot
- qqplot

Stokastiske variable

- Tæthedsfunktioner ($f(x)$)
- Sandynligheder ($\sum f(x)$ eller $\int f(x)dx$)
- Middelværdi (e.g. $\mu = \int x f(x)dx$)
- Varians (e.g. $\sigma^2 = \int (x-\mu)^2 f(x)dx$)
- Kovarians og korrelation
- Konkrete fordelinger (e.g. Binomial, Normal, log-normal,...)

Funktioner af stokastiske variable

- Simulation (e.g. X og $Y \sim N(\mu, \sigma^2)$, $P(X \cdot Y > c)$)
- Error propagation (Taylor omkring $E(X_i)$ plus middelværdi og varians/kovarians regneregler)
- $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_i \sim N(\mu_2, \sigma_2^2)$
  - Fordeling af gennemsnit (normalfordeling)
  - Fordeling af varians estimator (χ^2-fordeling)
  - Fordeling af standardliseret gennemsnit (t-fordeling)

Normal fordelingen

En standard normal fordeling:

$$Z \sim N(0, 1^2)$$

En normalfordeling med middelværdi 0 og varians 1.

Standardisering:

En vilkårlig normal fordelt variabel $X \sim N(\mu, \sigma^2)$ kan standardiseres ved at beregne

$$Z = \frac{X - \mu}{\sigma}$$

Fordeling for gennemsnit af normalfordelinger (Theorem 3.2)

(Stikprøve-) fordelingen/ The (sampling) distribution for $\bar{X}$

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Fordeling af varians estimator

Hvis $X_1, \ldots, X_n$ er i.i.d. $N(\mu, \sigma^2)$, og lad $\overline{X}$, S^2 være hhv. gennemsnit og empirisk varians. Så gælder der at

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

og det følger at

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

t-fordelingen som stikprøvefordeling

Lad $X_1, \ldots, X_n$ være i.i.d. $\sim N(\mu, \sigma^2)$ så følger

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{1}$$

en t-fordeling med $n - 1$ frihedsgrader.

Hypotesetest og konfidensintervaller

$X_i \sim N(\mu, \sigma^2)$:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{observeret - hypotese}{std(obs)}$$

Under H_0

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t$$

E.g.

$$p.value = 2P(T > |t_{\text{obs}}|)$$

Oversigt

1. Introduktion til simulation
   - Hvad er simulering egentlig?
2. Fejlophobningslove
3. Parametric bootstrap
   - Introduction to bootstrap
   - One-sample konfidensinterval for μ
   - One-sample konfidensinterval for en vilkårlig størrelse
   - Two-sample konfidensintervaller for en vilkårlig fordeling
4. Ikke-parametrisk bootstrap
   - One-sample konfidensinterval for en vilkårlig størrelse
   - One-sample konfidensinterval for μ
   - Two-sample konfidensintervaller
5. Første del af kurset