

Kursus 02403

Introduktion til Statistik

Forelæsning 10: Statistik ved hjælp af simulering

Jan Kloppenborg Møller

DTU Informatik

Bygning 305 - rum 219

Danmarks Tekniske Universitet

2800 Lyngby – Danmark

e-mail: jkm@imm.dtu.dk

Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlophobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat

Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlophobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat

Motivation

- Table 8.1 har et "hul":
- Små stikprøver som IKKE kommer fra en normalfordeling??
- I gl. dage: non-parametriske tests, e.g. Kapitel 14.
- Mere almindeligt nu: Simuleringsbaseret:
 - Konfidensintervaller er meget nemmere at opnå
 - De er meget nemmere at anvende i mere komplicerede situationer
 - De "forgrover" ikke informationen i samme udstrækning
 - De afspejler i højere grad dagens virkelighed - de anvendes simpelt hen nu i rigtig mange sammenhænge
- 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:

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

Eksempel 1

En virksomhed producerer rektangulære plader. Længden af pladerne (i meter), X , antages at kunne beskrives med en normalfordeling $N(2, 0.1^2)$ og bredden af pladerne (i meter), Y , antages at kunne beskrives med en normalfordeling $N(3, 0.2^2)$. 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 1, løsning i R

- Kode:

- `k=10000`
- `X=rnorm(k,2,0.1)`
- `Y=rnorm(k,3,0.2)`
- `A=X*Y`
- `mean(A)`
- `sd(A)`
- `sum(abs(A-6)>0.1)/k`

- Resultat:

```
> mean(A)
[1] 5.999061
> sd(A)
[1] 0.5030009
> sum(abs(A-
6)>0.1)/k
[1] 0.8462
```


Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlafhobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat

Fejlophobningslove

- Skal kunne finde: $\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$
- Vi kender allerede:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2, \text{ hvis } f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i$$

- Ny regel for ikke-lineære funktioner:

$$\sigma_{f(X_1, \dots, X_n)}^2 \approx \sum_{i=1}^n \left(\frac{\partial f}{\partial X_i} \right)^2 \sigma_i^2$$

Fejlophobningslove

Eller ved simulering:

Simuler k udfald af samtlige n målinger som

$$N(X_i, \sigma_i^2): X_i^{(j)}, j = 1 \dots, k$$

Beregn spredningen direkte som den observerede spredning af de k værdier for f :

$$\sigma_{f(X_1, \dots, X_n)} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (f_j - \bar{f})^2}$$
$$f_j = f(X_1^{(j)}, \dots, X_n^{(j)})$$

Eksempel 1, 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 $A = 2.05 \times 2.99 = 6.13$ fundet ved den ikke-lineære fejlophobningslov?

Eksempel 1, fortsat

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

$$\begin{aligned}\text{Var}(XY) &= E[(XY)^2] - [E(XY)]^2 \\&= E(X^2)E(Y^2) - E(X)^2E(Y)^2 \\&= [\text{Var}(X) + E(X)^2] [\text{Var}(Y) + E(Y)^2] - E(X)^2E(Y)^2 \\&= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)E(Y)^2 + \text{Var}(Y)E(X)^2 \\&= 0.1^2 \times 0.2^2 + 0.1^2 \times 3^2 + 0.2^2 \times 2^2 \\&= 0.0004 + 0.09 + 0.16 \\&= 0.2504\end{aligned}$$

Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlophobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat

Konfidensintervaller ved hjælp af simulering: bootstrapping

- What to do med en lille stikprøve, som IKKE er normalfordelt?
- To mulige løsninger
 - 1 Find/identificer/antag en anden og mere rigtig fordeling for populationen ("systemet")
 - 2 Undlad at antage nogen fordeling overhovedet
- Bootstrapping findes i to versioner:
 - 1 Parametrisk bootstrap: Simuler gentagne stikprøver fra den antagede fordeling.
 - 2 Ikke-parametrisk bootstrap: Simuler gentagne stikprøver direkte fra data.

Ikke-parametrisk bootstrap for one-sample situationen

Data: $x_1, \dots, x_n$.

$100(1 - \alpha)\%$ -konfidensintervallet for μ :

Simuler k stikprøver af størrelse n ved at udtage tilfældigt
blandt de tilgængelige data

(med tilbagelægning - stort k , e.g. $k > 1000$)

Beregn gennemsnittet i hver af de k stikprøver: $\bar{x}_1^*, \dots, \bar{x}_k^*$

Beregn $100\alpha/2\%$ - og $100(1 - \alpha/2)\%$ fraktilerne for disse

Intervallet er: $[\text{fraktil}_{100\alpha/2\%}, \text{fraktil}_{100(1-\alpha/2)\%}]$

Eksempel 2, one-sample

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		

Eksempel 2, løsning i R

- Dataindlæsning:
`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`
- R-Metode 1:
`k=10000`
`mysamples = replicate(k, sample(dif, replace = TRUE))`
`mymeans = apply(mysamples, 2, mean)`
`quantile(mymeans,c(0.025,0.975))`
- R-Metode 2: (Installer først pakken "bootstrap")
`library(bootstrap)`
`quantile(bootstrap(dif,k,mean)$thetastar,c(0.025,0.975))`

Two-sample situationen

Data: $x_1, \dots, x_{n_1}$ og $y_1, \dots, y_{n_2}$

100(1 - α)%-konfidensintervallet for $\mu_1 - \mu_2$:

Simuler k sæt af 2 stikprøver af størrelse n_1 og n_2 ved at udtage tilfældigt blandt de tilgængelige data (med tilbagelægning - stort k , e.g. $k > 1000$)

Beregn forskellen i gennemsnittene for hver af de k stikprøvepar:

$$\bar{x}_1^* - \bar{y}_1^*, \dots, \bar{x}_k^* - \bar{y}_k^*$$

Beregn 100 α /2%- og 100(1 - α /2)% fraktilerne for disse
Intervallet er: $[\text{fraktil}_{100\alpha/2\%}, \text{fraktil}_{100(1-\alpha/2)\%}]$

Eksempel 3

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 3, løsning i R

- Dataindlæsning:

```
x=c(9,10,12,6,10,8,6,20,12)
```

```
y=c(14,15,19,12,13,13,16,14,9,12)
```

- Bootstrapping i R:

```
k=10000
```

```
xsamples = replicate(k, sample (x, replace = TRUE))
```

```
ysamples = replicate(k, sample (y, replace = TRUE))
```

```
mymeandifs = apply(xsamples, 2, mean)-apply(ysamples, 2,  
mean)
```

```
quantile(mymeandifs,c(0.025,0.975))
```

Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlophobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat

Hypotesetest ved hjælp af bootstrap konfidensintervaller

- Sammenhæng mellem hypotese og konfidensinterval:

$H_0 : \theta = \theta_0$ accepteres $\Leftrightarrow \theta_0$ ligger i konfidensintervallet for θ

- F.eks. ensidet hypotese-test vha. bootstrap:

$H_0 : \theta = \theta_0$ mod $H_1 : \theta > \theta_0$ accepteres $\Leftrightarrow$

$\theta_0 > 100\alpha\%$ -fraktilen for bootstrapværdierne for θ

One-sample setup, Eksempel 2, fortsat

Vi fortsætter cigaretforbrugseksemplet.

Man vil nu gerne påvise, at cigaretforbruget er faldet efter fødslen:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ mod } H_0 : \mu_1 - \mu_2 > 0$$

P-værdien findes i R som: `sum(mymeans<0)/k`

Hypotesetest ved hjælp af permutationstest

Vi har nu stikprøverne: $x_1, \dots, x_{n_1}$ og $y_1, \dots, y_{n_2}$

$$\hat{\mu}_1 = \bar{x} \text{ og } \hat{\mu}_2 = \bar{y}$$

Et permutationstest for hypotesen $\mu_1 = \mu_2$ er defineret ved:

Simuler k sæt af 2 stikprøver af størrelse n_1 og n_2
ved at permutere de tilgængelige data
(stort k , e.g. $k > 1000$)

Beregn forskellen i gennemsnittene for hver af de k stikprøvepar:

$$\bar{x}_1^* - \bar{y}_1^*, \dots, \bar{x}_k^* - \bar{y}_k^*$$

Find P-værdien ud fra positionen af $\bar{x} - \bar{y}$ i denne fordeling
(2-sidet eller 1-sidet - på sædvanlig vis)

Two-sample setup, Eksempel 3, fortsat

Vi fortsætter eksemplet med tænderne.

Vi ønsker at udføre et tosidet test for om $\mu_1 = \mu_2$.

Følgende R-kode gennemfører beregningerne:

```
x=c(9,10,12,6,10,8,6,20,12)
```

```
y=c(14,15,19,12,13,13,16,14,9,12)
```

```
k=100000
```

```
perms = replicate(k,sample(c(x,y)))
```

```
mymeandifs = apply(perms[1:9,], 2, mean)-apply(perms[10:19,], 2,  
mean)
```

```
sum(abs(mymeandifs)>abs(mean(x)-mean(y)))/k
```

Oversigt

- 1 Introduktion til simulering
 - Eksempel 1
- 2 Fejlophobningslove
 - Eksempel 1, fortsat
- 3 Konfidensintervaller ved hjælp af simulering: bootstrapping
 - Eksempel 2, one-sample
 - Two-sample situationen
 - Eksempel 3
- 4 Hypotesetest ved hjælp af simulering
 - Vha. bootstrap konfidensintervaller
 - One-sample setup, Eksempel 2, fortsat
 - Hypotesetest ved hjælp af permutationstest
 - Two-sample setup, Eksempel 3, fortsat