STK1100 våren 2018

Bootstrapping, Stokastisk simulering og Monte Carlo integrasjon

Svarer til notatet «Bootstrap og simulering»

Ørnulf Borgan
Matematisk institutt
Universitetet i Oslo

Estimering

Vi anta at $x_1, x_2, ..., x_n$ er observerte verdier av uavhengige og identisk (u.i.f.) fordelte stokastiske variabler $X_1, X_2, ..., X_n$ og at X_i-ene har en fordeling som avhenger av en parameter θ

Vi vil estimere verdien til θ på grunnlag av observasjonene våre

Til det bruker vi en estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$

På grunnlag av de observerte x_i-ene får vi estimatet $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$

Standardfeil

Når vi rapporterer resultatet av en undersøkelse, bør vi ikke nøye oss med å oppgi et estimatet.
Vi bør også si noe om hvor presist estimatet er. Det er da vanlig å oppgi (et estimat for) standardfeilen

Standardavviket $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ til en estimator $\hat{\theta}$ blir vanligvis kalt standardfeilen til estimatoren

Ofte vil $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ avhenge av en eller flere ukjente parmetere. Hvis vi estimerer disse, får vi den estimerte standardfeilen $s_{\hat{\theta}}$

Bootstrap

For enkle situasjoner kan vi finne et uttrykk for standardfeilen $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ til en estimator

Men hvis estimatoren og/eller fordelingen til X_i-ene er komplisert, kan det være vanskeligere å finne et slikt uttrykk

Da kan vi bruke stokastisk simulering til å finne et estimat $s_{\hat{\theta}}$ for standardfeilen

Vi ser først på en metode som kalles parametrisk bootstrap

Anta at X_i-ene har tetthet/punktsannsynlighet $f(x;\theta)$

Ut fra de observerte x_i-ene får vi estimatet $\hat{\theta}$

For $b = 1, 2, ..., B$ gjør vi nå følgende:

- Genererer et bootstrap-utvalg $x_1^*, x_2^*, ..., x_n^*$ fra tettheten/punktsannsynligheten $f(x;\hat{\theta})$
- Beregner estimatet $\hat{\theta}_b^*$ ut fra bootstrap-utvalget (på samme måte som $\hat{\theta}$ ble beregnet ut fra de opprinnelige observasjonene)

Bootstrap estimatet for standardfeilen er da

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\theta}^*\right)^2}$$

Eksempel (jf. oppgave 8.38)

Vi har observasjonene (antall skritt per sekund):

0.95	0.85	0.92	0.95	0.93	0.86	1.00	0.92	0.85	0.81
0.78	0.93	0.93	1.05	0.93	1.06	1.06	0.96	0.81	0.96

Vi vil anta at disse er observasjoner av $X_1, ..., X_{20}$ som er uavhengige og $N(\mu, \sigma^2)$ -fordelte, og vi lar $Y_1 < ... < Y_{20}$ være X_i-ene gitt i stigende rekkefølge

Vi ser på tre estimatorer for μ :

- Gjennomsnittet $\hat{\mu}_1 = \frac{1}{20}\sum_{i=1}^{20} X_i$
- Empirisk median $\hat{\mu}_2 = (Y_{10} + Y_{11})/2$
- 10% trimmet gjennomsnitt $\hat{\mu}_3 = \frac{1}{16}\sum_{j=3}^{18} Y_j$

Estimatene blir:

$\hat{\mu}_1 = 0.926$ $\hat{\mu}_2 = 0.930$ $\hat{\mu}_3 = 0.925$

Vi kan bestemme standardfeilene ved bootstrap:

```
x=[0.95 0.85 0.92 0.95 0.93 0.86 1.00 0.92 0.85 0.81 0.78 0.93 0.93 1.05 0.93
1.06 1.06 0.96 0.81 0.96]
n=length(x); m=mean(x); s=std(x);
B=1000;
meanvec=zeros(1,B); medianvec=zeros(1,B); trmeanvec=zeros(1,B);
for b=1:B
   xstar=normrnd(m,s,1,n);
   meanvec(b)=mean(xstar);
   medianvec(b)=median(xstar);
   trmeanvec(b)=trimmean(xstar,20);
end
std(meanvec)
std(medianvec)
std(trmeanvec)
```

Vi finner standardfeilene (basert på én kjøring)

$s_{\hat{\mu}_1} = 0.0180$ $s_{\hat{\mu}_2} = 0.0216$ $s_{\hat{\mu}_3} = 0.0186$

Parametrisk bootstrap forusetter at den modellen vi bruker for dataene gir en rimelig god beskrivelse av fordelingen til X_i-ene

For ikke-parametrisk bootstrap gjør vi ingen forutsetninger om fordelingen til X_i-ene

Da trekker vi bootstrap-utvalget fra den empiriske fordelingsfunksjonen

$$\hat{F}(x) = \frac{1}{n}\{\text{antall } x_i \le x\}$$

Merk at $\hat{F}$ gir sannsynlighet 1/*n* til hver x_i

Trekking fra $\hat{F}$ svarer derfor til trekning fra $x_1, x_2, ..., x_n$ med tilbakelegging

Diskrete fordelinger og empirisk fordeling

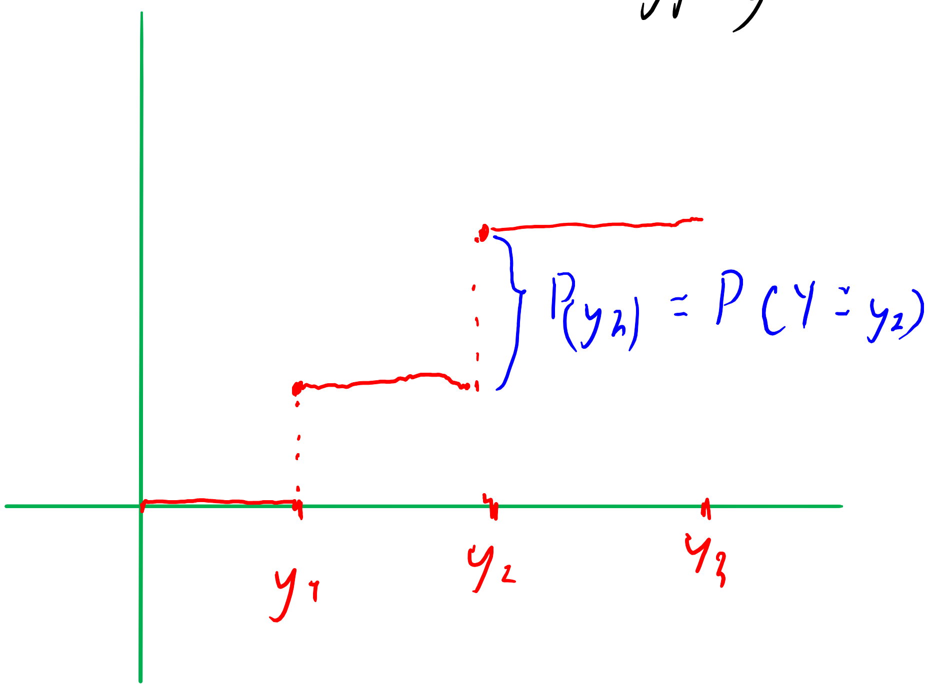
Y er stok. variabel med mulige verdier $y_1 < y_2 < \dots < y_k$

Punktsannsynligheter

$$p(y_j) = P(Y = y_j)$$

Kumulativ fordeling

$$F(y) = P(Y \leq y) = \sum_{y_j \leq y} p(y_j)$$



Kap 3!

Empirisk fordeling

La nå $x_1, \dots, x_n$ fra kumulativ fordeling $F(x)$

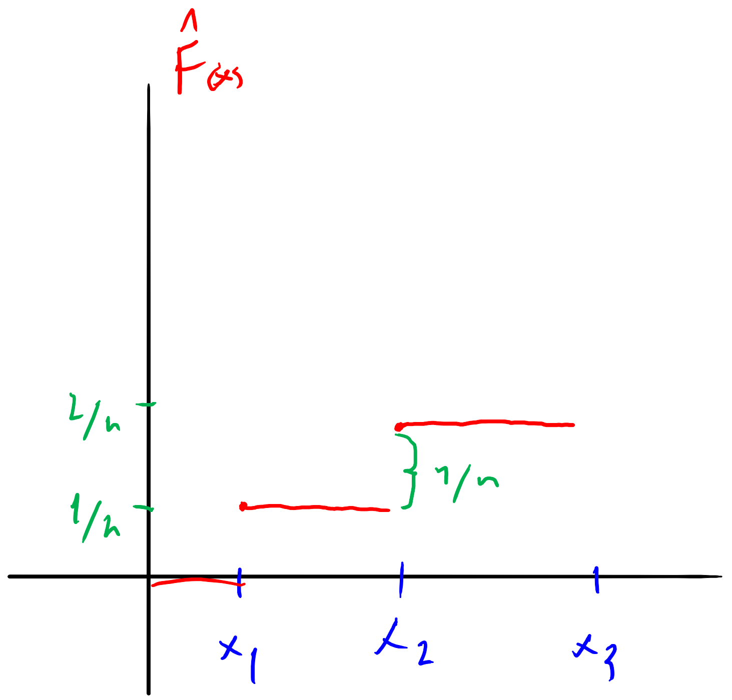
Har at

$$F(x) = P(X_i \le x)$$

Estimator for $F(x)$ er den empiriske kumulative fordelingen

$$\hat{F}(x) = \frac{\text{antall } x_i \le x}{n}$$

Observerer $x_1, x_2, \dots, x_n$



La nå x^* ha fordeling $\hat{F}(x)$

Da har vi

$$P(x^* = x_1) = 1/n$$

x^* bli svare til en bootstrap observasjon

Framgangsmåten for ikke-parametrisk bootstrap er dermed som følger:

For $b = 1, 2, ..., B$ gjør vi følgende:

- Trekk *n* verdier med tilbakelegging fra $x_1, x_2, ..., x_n$
  Kall de valgte verdiene $x_1^*, x_2^*, ..., x_n^*$
- Beregner estimatet $\hat{\theta}_b^*$ ut fra bootstrap-utvalget

Bootstrap estimatet for standardfeilen er

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\theta}^*\right)^2}$$

For eksemplet har vi kommandoene:

```
x=[0.95 0.85 0.92 0.95 0.93 0.86 1.00 0.92 0.85 0.81 0.78 0.93 0.93 1.05 0.93
1.06 1.06 0.96 0.81 0.96]
n=length(x);
B=1000;
meanvec=zeros(1,B); medianvec=zeros(1,B); trmeanvec=zeros(1,B);
for b=1:B
  xstar=randsample(x,n,true);
  meanvec(b)=mean(xstar);
  medianvec(b)=median(xstar);
  trmeanvec(b)=trimmean(xstar,20);
end
std(meanvec)
std(medianvec)
std(trmeanvec)
```

Vi finner standardfeilene (basert på én kjøring)

$s_{\hat{\mu}_1} = 0.0175 \qquad s_{\hat{\mu}_2} = 0.0150 \qquad s_{\hat{\mu}_3} = 0.0197$

Simulering av tilfeldige tall på [0,1]

Datamaskiner kan generere en følge av tall, såkalt «pseudotilfeldige» tilfeldige tall, som for (nesten) alle praktiske formål ligner på tilfeldige tall på intervallet [0,1]

Formelt svarer et tilfeldig tall på [0,1] til en stokastisk variabel *U* som er uniformt fordelt på [0,1]

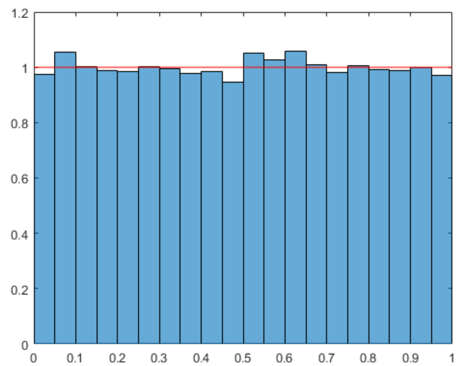
Hvis *U* er uniformt fordelt på [0,1] , har vi at

$$f_U(u) = \begin{cases} 1 & \text{for } 0 \le u \le 1 \\ 0 & \text{ellers} \end{cases}$$

$$F_U(u) = \begin{cases} 0 & u < 0 \\ u & 0 \le u \le 1 \\ 1 & u > 1 \end{cases}$$

MATLAB:

```
u=rand(1,10000);
histogram(u, 'Normalization','pdf')
hold on
plot([0,1],[1,1],'r')
```



Hvordan kan vi ut fra tilfeldige tall på [0,1] generere en kontinuerlig fordelt stokastisk variabel X som har en gitt fordeling?

Vi skal se på to metoder (det fins flere):

- Inversjonsmetoden
- Forkastningsmetoden

Inversjonsmetoden kan vi bruke når vi har et eksplisitt uttrykk for den inverse av den kumulative fordelingen til X

Forkastningsmetoden kan vi bruke også når vi ikke har et uttrykk for den inverse av den kumulative fordelingen

Inversjonsmetoden

Vi vil generere en kontinuerlig stokastisk variabel X som har kumulativ fordelingsfunksjon $F(x)$. Her er $F(x)$ en strengt voksende kumulativ fordelingsfunksjon.

La $U \sim \text{uniform}[0,1]$ og sett $X = F^{-1}(U)$

Da er den kumulative fordelingen til X gitt ved

$$F_X(x) = P(X \le x) = P(F^{-1}(U) \le x)$$
$$= P(U \le F(x)) = F(x)$$

så X har kumulativ fordeling $F(x)$

Eksempel: eksponentialfordelingen

Vi vil at X skal ha kumulativ fordeling

$$F(x) = \begin{cases} 0 & \text{for } x \le 0 \\ 1 - e^{-\lambda x} & \text{for } x > 0 \end{cases}$$

Den inverse funksjonen er (for $u > 0$)

$$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$$

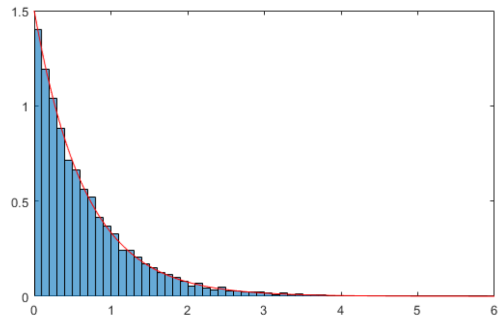
Så hvis U er uniformt fordelt på [0,1], så er

$$X = -\frac{1}{\lambda}\ln(1-U)$$

eksponentialfordelt med parameter λ

MATLAB:

```
u=rand(1,10000);
lambda=1.5;
x=-log(1-u)/lambda;
histogram(x,'Normalization','pdf')
hold on
xp=0:0.01:6;
plot(xp,exppdf(xp,1/lambda),'r')
```



Eksempel: Cauchy fordelingen

Standard Cauchy fordelingen har tetthet

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Den kumulative fordelingen er

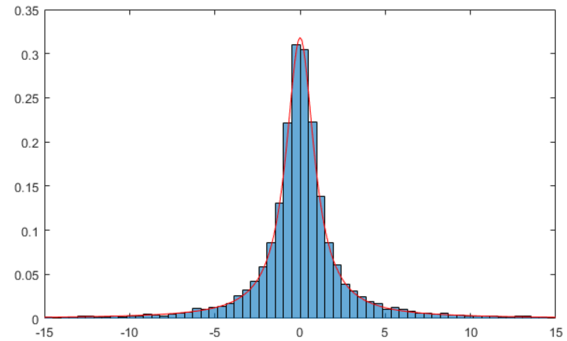
$$F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$$

Den inverse av den kumulative fordelingen er

$$F^{-1}(u) = \tan\left[\pi(u - 1/2)\right]$$

MATLAB:

```
u=rand(1,10000);
x=tan(pi*(u-1/2));
histogram(x,'BinLimits',[-15,15],'Normalization','pdf')
hold on
xp=-15:0.1:15;
plot(xp,1./(pi.*(1+xp.^2)),'r')
```



Forkastningsmetoden

Vi vil generere en kontinuerlig stokastisk variabel X som har sannsynlighetstetthet $f(x)$, men vi har ikke noe analytisk uttrykk for den kumulative fordelingen $F(x)$

Et alternativ er da forkastningsmetoden

Vi trenger da en forslagsfordeling med tetthet $g(x)$ og kumulativ fordeling $G(x)$ slik at vi lett kan generere $Y \sim G$

Vi må velge forslagsfordelingen slik at

$$\frac{f(x)}{g(x)} \le c$$

for alle x, der $c \ge 1$

Vi genererer nå X ved følgende algoritme:

1. Generer $Y \sim G$
2. Generer $U \sim \text{uniform}[0,1]$
3. Hvis $U \le \frac{f(Y)}{cg(Y)}$, sett $X = Y$. Ellers gå til trinn 1

Da har X tettheten $f(x)$ og sannsynligheten for å akseptere en generert Y er $1/c$ (bevis i oppgave 5.85)

Det er vanlig å velge c slik at

$$c = \max_x \frac{f(x)}{g(x)}$$

Eksempel: Beta fordelingen

Vi vil generere X som har tetthet (et spesialtilfelle av betafordelingen)

$$f(x) = \begin{cases} 20x(1-x)^3 & \text{for } 0 \le x \le 1 \\ 0 & \text{ellers} \end{cases}$$

Vi kan her la forslagsfordelingen være uniform:

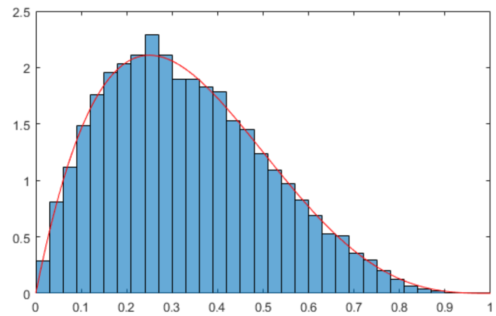
$$g(y) = \begin{cases} 1 & \text{for } 0 \le y \le 1 \\ 0 & \text{ellers} \end{cases}$$

Vi velger da

$$c = \max_x \frac{f(x)}{g(x)} = \max_x \{20x(1-x)^3\} = \frac{135}{64}$$

MATLAB:

```
n = 10000;
x = zeros(1,n);
c = 135/64;
for i=1:n
  y = rand(1);
  u = rand(1);
  while (u > 20*y*(1-y)^3/c)
    y = rand(1);
    u = rand(1);
  end
  x(i) = y;
end
histogram(x,'Normalization','pdf')
hold on
xp=-0:0.01:1;
plot(xp,20.*xp.*(1-xp).^3,'r')
```



Monte Carlo integrasjon

Vi er interessert i å bestemme integralet

$$\theta = \int_{-\infty}^{\infty} \cos(x) e^{-x^2/2} \, dx$$

Merk at vi kan skrive

$$\theta = \int_{-\infty}^{\infty} \sqrt{2\pi} \cos(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx \quad = \int_{-\infty}^{\infty} \sqrt{2\pi} \cos(x) f(x) \, dx$$

der $f(x)$ er standardnormaltettheten

Altså er

$$\theta = E\left\{ \sqrt{2\pi} \cos(X) \right\}$$

der $X \sim f$

Vi kan estimere θ med

$$\hat{\theta} = \bar{Y} = \frac{1}{M} \sum_{i=1}^{M} Y_i \qquad \text{der} \qquad Y_i = \sqrt{2\pi} \cos(X_i)$$

og $X_1, X_2, \ldots, X_M$ er u.i.f med tetthet $f(x)$

MATLAB:

```
M = 10000;
x = normrnd(0,1,1,M);
y = sqrt(2*pi)*cos(x);
thetahat = mean(y)
```

Vi kan lett bestemme en feilmargin for estimatet $\hat{\theta}$

Sett $S = \sqrt{\frac{1}{M-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

Da er $\frac{\bar{Y}-\theta}{S/\sqrt{M}}$ tilnærmet standardnormalfordelt

Et 100(1-α)% konfidensintervall for θ er gitt ved

$$\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{M}}$$

MATLAB:

```
alfa=0.01;
s=std(y);
feilmargin=norminv(1-alfa/2)*s/sqrt(M)
```


Generelt ser vi på et integral av formen

$$\theta = \int_{-\infty}^{\infty} g(x)dx$$

Det kan vi skrive

$$\theta = \int_{-\infty}^{\infty} \frac{g(x)}{f(x)} f(x)dx$$

der $f(x)$ er en tetthet og $f(x) > 0$ hvis $g(x) > 0$

Da er

$$\theta = E\{g(X)/f(X)\}$$

der $X \sim f$

Vi kan estimere θ med

$$\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M} Y_i \qquad \text{der} \qquad Y_i = g(X_i)/f(X_i)$$

og $X_1, X_2, \ldots, X_M$ er u.i.f med tetthet $f(x)$

Dette kalles Monte Carlo integrasjon