Analiza podataka

Bojan Basrak, PMF–MO Zagreb

Financijski praktikum

27. ožujka 2017.

Inicijalni korak u analizi i modeliranju u aktuarstvu i financijskoj matematici je opisna analiza jednodimenzionalnih podataka npr.

- visina šteta,
- log povrata

$$X_t = \log S_t/S_{t-1}\,,$$

- relativnih povrata

$$X'_t = \frac{S_t - S_{t-1}}{S_{t-1}}\,,$$

uočimo $S_t = (1 + X'_t)S_{t-1}$, kao u CRR modelu npr. Za male povrate $X_t \approx X'_t$.

Neparametarske metode

Empirijska funkcija distribucije

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(\infty, x]}(X_i)\,,$$

jasno càdlàg, mon. raste, $\hat{F}_n(-\infty) = 0,\ \hat{F}_n(+\infty) = 1.$
Ako su $X_{(1)} \leq \cdots \leq X_{(n)}$ **uređajne statistike** uzorka (i nema jednakih među njima)

$$\hat{F}_n(X_{(k)}) = \frac{k}{n},\ k = 1, \ldots, n\,.$$

Za njd uzorak $X_1, \ldots, X_n \sim F$ vrijedi

Teorem (Glivenko–Cantelli)

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{gs} 0\,.$$

Generalizirani inverz

Za u $u \in (0,1)$

$$F^{\leftarrow}(u) = \inf\{x : F(x) \geq u\}\,,$$

vrijednost

$$q_u = F^{\leftarrow}(u)\,,$$

zovemo u–kvantilom razdiobe F. Jasno ako je F bijekcija kao gore $F^{\leftarrow} = F^{-1}$.

Ako je $Y = aX + b$, $a > 0, b \in \mathbb{R}$ lako se vidi

$$F_Y^{\leftarrow}(u) = aF_X^{\leftarrow}(u) + b\,.$$

Funkcija $\hat{F}_n$ skače u točkama $X_{(k)}$ za $1/n$, tako da funkcija $\hat{F}_n^{\leftarrow}$ skače u točkama k/n za $X_{(k)} - X_{(k-1)}$, tj

$$\hat{F}_n^{\leftarrow}(t) = \begin{cases} X_{(k)} & t \in (\frac{k-1}{n}, \frac{k}{n}], \ k \leq n-1\,, \\ X_{(n)} & t \in (\frac{n-1}{n}, 1). \end{cases} \quad . \tag{1}$$

Korolar

Za sve t u kojima je $F^{\leftarrow}$ neprekidna vrijedi

$$\hat{F}_n^{\leftarrow}(t) \xrightarrow{gs} F^{\leftarrow}(t)\,.$$

Graf kvantila

qq-plot

je skup točaka

$$\left\{ \left(F^{\leftarrow} \left(\frac{k}{n+1} \right), X_{(k)} \right) : k = 1, \ldots, n \right\},$$

a omogućuje usporedbu dvije razdiobe.

Kažemo da je gornji rep razdiobe F_2 teži od gornjeg repa razdiobe F_1 ako

$$\lim_{u \to 1} F_2^{\leftarrow}(u) / F_1^{\leftarrow}(u) = +\infty$$

U upravljanju rizicima jedan je od repova važniji (jer predstavlja gubitke), mi ćemo pretpostaviti da je to gornji rep.

Neka je

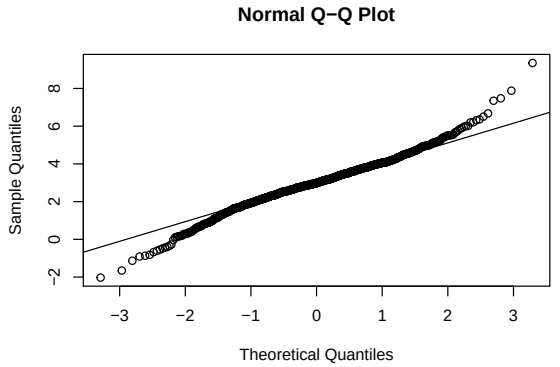
$Y =$ gubitak u nekom portfelju ,

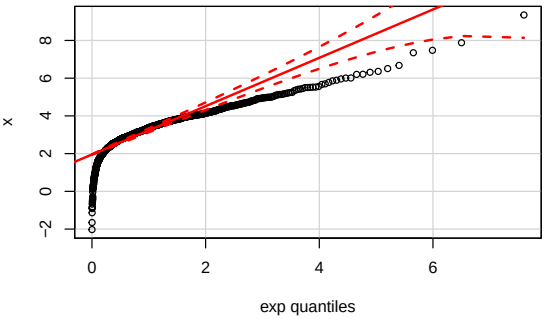
definiramo

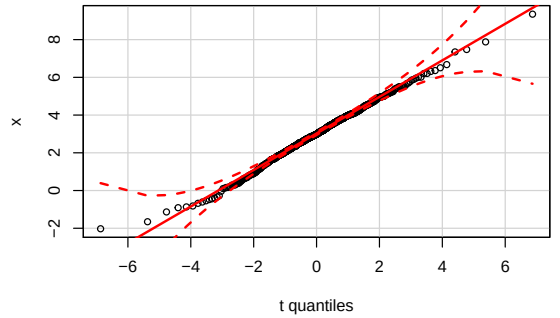
$$x_l = \inf\{x : F_Y(x) > 0\}$$
$$x_r = \sup\{x : F_Y(x) < 1\}$$

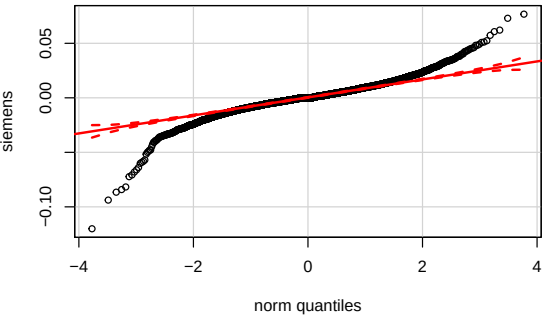
tipično pretpostavljamo da je $x_r = \infty$.

```r
x<-3+rt(1000,5)

par(mfrow=c(2,2))

qqnorm(x)
qqline(x)
library(car)
qqPlot(x, dist="exp")
qqPlot(x, dist="t", df=5)

library(evir)
data(siemens)
# ts.plot(siemens)
qqPlot(siemens,dist="norm")
par(mfrow=c(1,1))
```

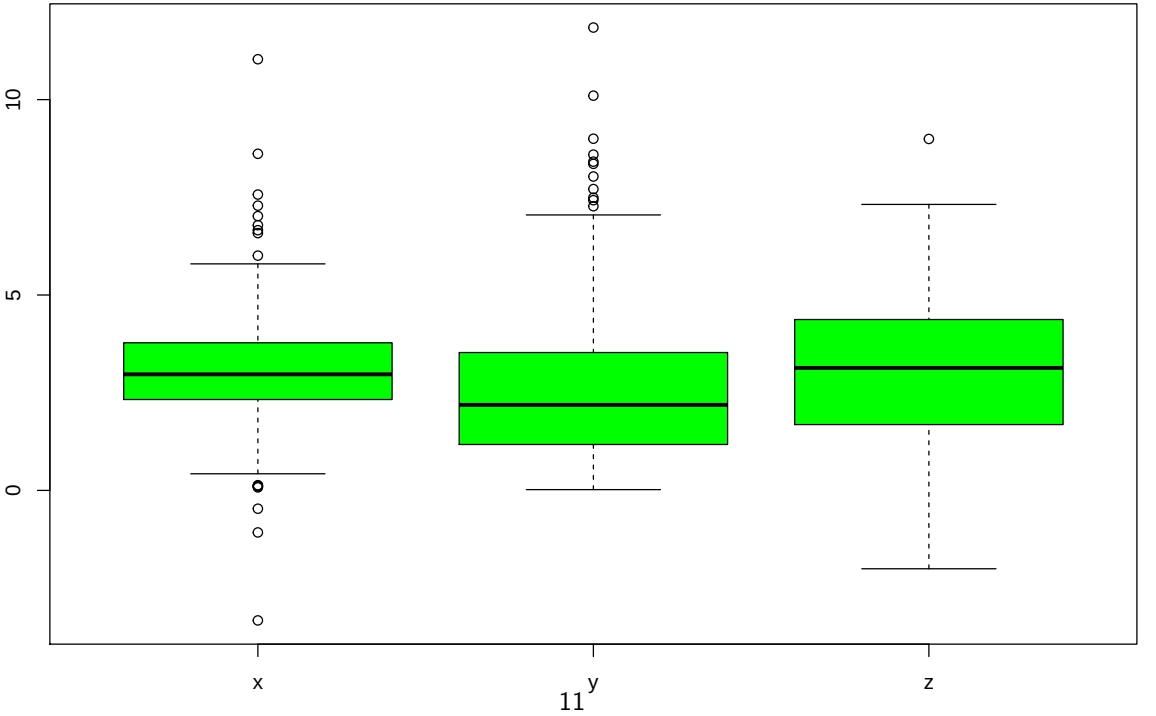

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−2
0
2
4
6
8


x
exp quantiles
0
2
4
6
−2
0
2
4
6
8


x
t quantiles
−6
−4
−2
0
2
4
6
−2
0
2
4
6
8


siemens
norm quantiles
−4
−2
0
2
4
−0.10
0.00
0.05

```r
x<-3+rt(200,3)
y<-rchisq(200,3)
z<-rnorm(200,3,2)
boxplot(cbind(x,y,z),col="green")
```



```r
library(moments)
kurtosis(cbind(x,y,z))
```

```
##        x        y        z
## 7.857922 5.202722 2.700421
```

```r
skewness(cbind(x,y,z))
```

```
##           x           y           z
##  0.74161598  1.40959328 -0.08766888
```

Očekivani manjak

expected shortfall / mean excess loss

uz uvjet $EY < \infty$, je funkcija

$$y \mapsto e_F(u) = E(Y - y | Y > y)\,,$$

za $y \in (-\infty, x_r)$, $Y \sim F$, no vrijedi i

$$e_F(y) = \frac{1}{\overline{F}(y)} \int_y^\infty \overline{F}(t)dt\,.$$

Primjer

$Y \sim \text{Exp}(\lambda)$, tada je

$$e_F(u) = \frac{1}{\lambda}, \quad \forall u > 0\,.$$

Katkad se razdiobe sa svojstvom $e_F(y) \nearrow +\infty$, $y \to \infty$ zovu razdiobama teškog repa, a razdiobe za koje je $e_F(y)$ ograničena zovu se razdiobama lakog repa.

Primjer

razdioba	$e_F(y)$
Exp(λ)	$1/\lambda$
$\Gamma(\alpha, \beta)$	$\beta^{-1}(1 + \frac{\alpha-1}{\beta y} + o(\frac{1}{u})$
normalna	$\frac{1}{y}(1 + o(1))$
lognormalna	$\frac{\sigma^2 y}{\log y - \mu}(1 + o(1))$
Pareto(α) $\alpha > 1$	$\frac{\kappa+y}{\alpha-1}$

U praksi $e_F(y)$ možemo neparametarski procijeniti td zamijenimo F sa $\hat{F}_n$ tj. sa

$$\hat{e}_n(y) = e_{\hat{F}_n}(y) = \frac{\frac{1}{n}\sum_1^n (X_i - y)_+}{1 - \hat{F}_n(y)}\,.$$

Iz jakog zakona velikih brojeva slijedi

Propozicija

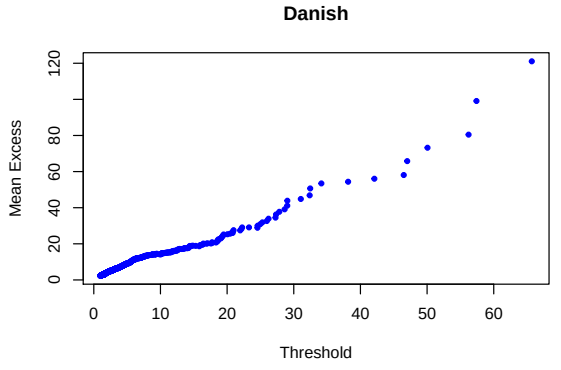
Za X_i njd td $x_r = \infty$, $EX_1 < \infty$, za sve $y > 0$

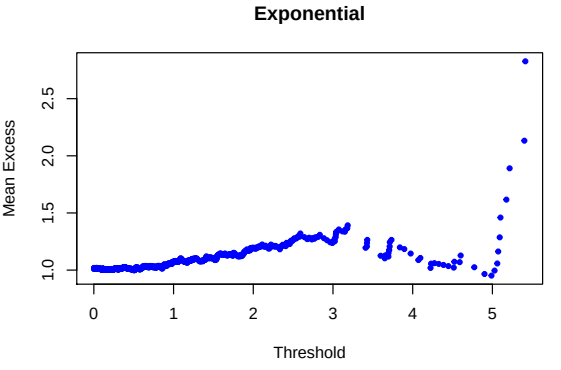
$$\hat{e}_n(y) \xrightarrow{gs} e_F(y)\,.$$

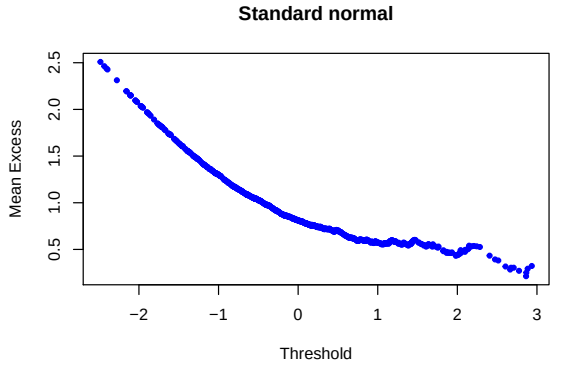
Graf očekivanog manjka (mean excess plot) je skup točaka

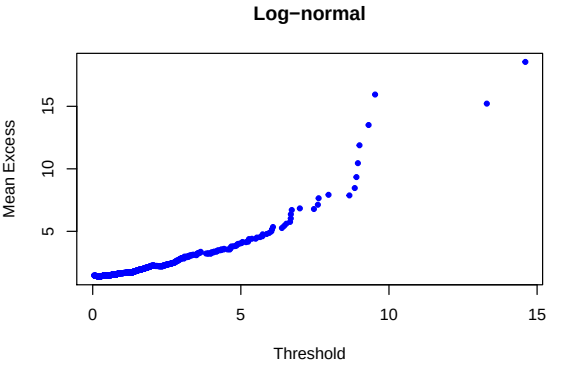
$$\left\{\left(X_{(k)}, \hat{e}_n(X_{(k)})\right) : k = 1, \ldots, n-1\right\}$$

```r
par(mfrow=c(2,2))

data(danish)
meplot(danish,main="Danish",col='blue', pch=20)
z<-rexp(1000)
meplot(z,main="Exponential",col='blue', pch=20)
z<-rnorm(1000)
meplot(z,main="Standard normal",col='blue', pch=20)
z<-exp(rnorm(1000))
meplot(z,main="Log-normal",col='blue', pch=20)
```

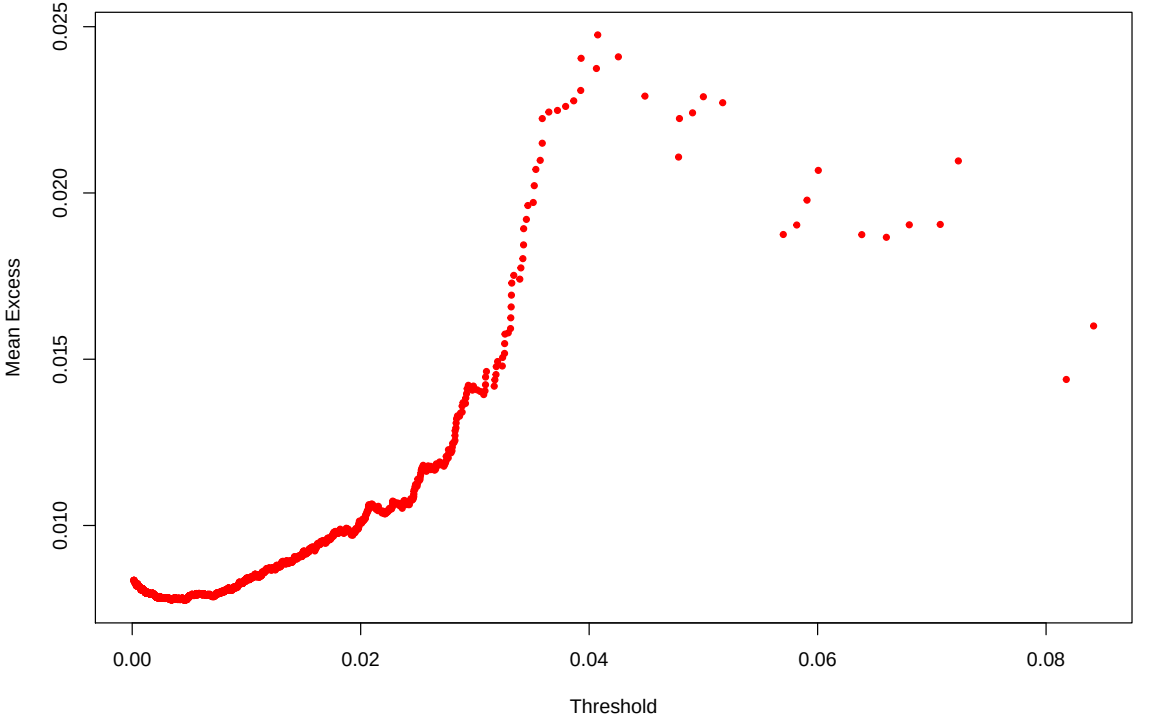

Danish
Mean Excess
Threshold
0
20
40
60
80
120
0
10
20
30
40
50
60


Exponential
Mean Excess
Threshold
1.0
1.5
2.0
2.5
0
1
2
3
4
5


Standard normal
Mean Excess
Threshold
0.5
1.0
1.5
2.0
2.5
−2
−1
0
1
2
3


Log−normal
Mean Excess
Threshold
5
10
15
0
5
10
15

```
par(mfrow=c(1,1))
meplot(-siemens[siemens<0],col='red', pch=20)
```



Plug–in procjenitelji

Procjenitelj $\hat{e}_n$ možemo smatrati *plug-in* procjeniteljem, jer je dobiven ubacivanjem e.f.d. na mjesto nepoznate razdiobe F, slično možemo procjenivati i razne druge numeričke karakteristike od F.

Primjer

i) procjenitelj očekivanja

$$\int_{\mathbb{R}} x d\hat{F}_n(x) = \frac{1}{n}\sum_{1}^{n} X_i = \overline{X}_n\,.$$

ii) procjenitelj varijance

$$\int_{\mathbb{R}} x^2 d\hat{F}_n(x) - \left(\int_{\mathbb{R}} x d\hat{F}_n(x)\right)^2$$
$$= \ldots = \frac{1}{n}\sum_{1}^{n}(X_i - \overline{X}_n)^2 = \hat{\sigma}^2\,.$$

Ili za usporedbu s normalnom razdiobom

iii) koeficijent nagnutosti (skewness)

$$\kappa = \frac{E(X-\mu)^3}{\sigma^3}\,,\ \text{procjenjujemo sa } \hat{\kappa} = \frac{\frac{1}{n}\sum_1^n (X_i - \overline{X}_n)^3}{\hat{\sigma}^3}\,,$$

iv) koeficijent spljoštenosti (kurtosis)

$$\gamma = \frac{E(X-\mu)^4}{\sigma^4}\,,\ \text{procjenjujemo sa } \hat{\gamma} = \frac{\frac{1}{n}\sum_1^n (X_i - \overline{X}_n)^4}{\hat{\sigma}^4}\,,$$

za normalne razdiobe $\gamma = 3$, što nam omogućuje grubu usporedbu težine repova između podataka i norm. razdiobe.

v) kvantili, možemo ih procijeniti sa

$$\hat{q}_u = \hat{q}_{n,u} = \hat{F}_n^{\leftarrow}(u) = \inf\{x : \hat{F}_n(x) \geq u\}$$

Ako je T statistika koja procjenjuje θ_F num. karakteristiku nepoznate razdiobe F često se pokazuje da je T asimptotski normalna za njd podatke. Tada

$$T \pm z_{\alpha/2} s.e.(T)$$

daje približno $(1-\alpha)100\%$ interval pouzdanosti za θ_F, gdje je $s.e.(T)$ standardna greška procjenitelja (koju često ne znamo također).

U procjeni očekivanja npr. po c.g.t.

$$s.e.(\overline{X}_n) = \frac{\sigma}{n} \approx \frac{\hat{\sigma}}{n}$$

Asimptotski je normalan i procjenitelj kvantila $\hat{q}_u$ a st.gr. mu je

$$\sqrt{\frac{u(1-u)}{n f_X(q_u)^2}}$$

ako X_i imaju gustoće $f_X(q_u) > 0$ u q_u. Izvedite st. grešku za uzorački medijan, no uočite da je $f_X(q_u) > 0$ nepoznato!!

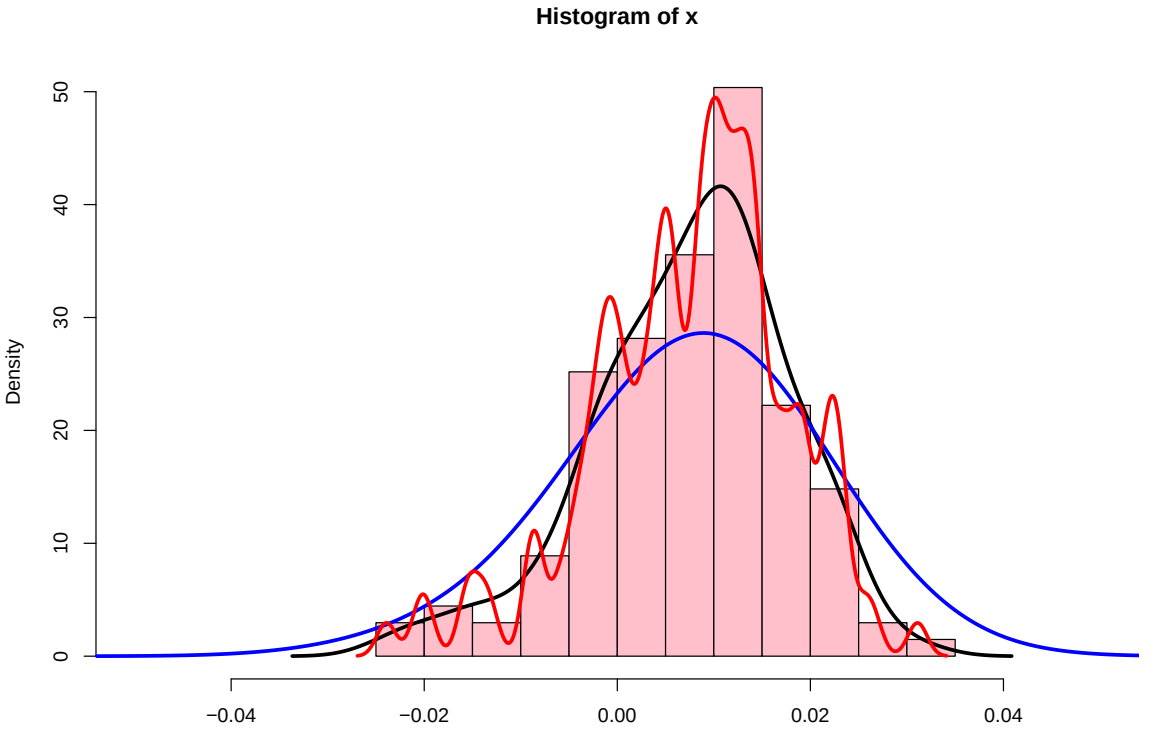
Za usporedbu uzorka s teorijskom razdiobom možemo koristiti **histogram** ili zaglađeni histogram, kojim zapravo procjenjujemo nepoznatu fukciju gustoće uzorka

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{1}^{n} k\left(\frac{X_i - x}{h}\right),$$

gdje je k neka funkcija gustoće simetrična oko 0 (npr. standardna normalna), a h je tzv. parametar zaglađivanja odn. *bandwidth.* Ponekad je korisno uzeti k koja nije nužno nenegativna no $\int_{\mathbb{R}} k(x)dx = 1$.

Parametar h nije lako odabrati, kao ni broj razreda u izradi histograma, no tipično je $\asymp n^{-1/5}$. Optimalni h se izvodi iz formule za rastav greške na varijancu i pristranost.

```r
library(tseries)
data(USeconomic)
x<-diff(USeconomic[,2])
hist(x,prob=T,xlim=c(-0.05,0.05),col="pink", xlab="")
lines(density(x),lwd=3)
lines(density(x,bw=0.01),lwd=3,col=4)
lines(density(x,bw=0.001),lwd=3,col=2)
```


Histogram of x
Density
0
10
20
30
40
50
−0.04
−0.02
0.00
0.02
0.04

Rastav greške

bias & variance

Ako je $T = g(X_1, \ldots, X_n)$ procjenitelj za neku num. karakteristiku razdiobe F, recimo θ_F, njegova je srednjekvadratna greška

$$\begin{aligned} E(T - \theta_F)^2 &= E(T - ET)^2 + (ET - \theta_F)^2 \\ &= \text{variance} \quad + \text{bias}^2 \end{aligned}$$

Prvi od gornjih članova je varijanca procjenitelja T, a drugi je kvadrat pristranosti. Tipično složeniji odn. manje zaglađeni procjenitelji imaju veću varijancu a manju pristranost (i obrnuto), optimalne T je teško odrediti u praksi.

Primjer

$$T_h = \hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{1}^{n} k\left(\frac{X_i - x}{h}\right) ,$$

Za $h \to 0$ raste varijanca, a smanjuje se pristranost. Za $h \to \infty$ raste pristranost, no smanjuje se varijanca.

Nakon procjene gustoće možemo dati približni interval pouzdanosti za q_u t.d. stavimo za standardnu grešku

$$\sqrt{\frac{u(1-u)}{n\hat{f}_n(\hat{q}_u)^2}}$$

Alternativno bismo mogli definirati novi procjenitelj za funkciju distribucije

$$\tilde{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(t)dt\,,$$

i procjenjivati q_u sa

$$\tilde{q}_u = \tilde{F}_n^{\leftarrow}(u) = \inf\{x : \tilde{F}_n(x) \geq u\,.\}$$

Procjenitelji ne moraju biti asimptotski normalni, ako slučajno znamo razdiobu G slučajne varijable $T - \theta_F$ opet možemo dati interval pouzdanosti $(1-\alpha)100\%$ kao

$$(T - q^G_{1-\alpha/2}, T - q^G_{\alpha/2})$$

Bootstrap

Koristimo ga za procjenu

- nepoznate razdiobe procjenitelja
- varijance i pristranosti procjenitelja

Ako je $T = T(\hat{F}_n) = t(X_1, \ldots, X_n)$ plug–in procjenitelj za $\theta_F = T(F)$ i vrijedi

$$T - \theta \overset{d}{\approx} N(\text{bias}(T), \text{var}(T))$$

možemo dobiti interval pouzdanosti za θ ako procjenimo $\text{bias}(T), \text{var}(T)$

Primjer (plug–in procjenitelji kao funkcije uzorka)

i) procjenitelj očekivanja

$$T(F) = \int x dF(x),$$
$$T(\hat{F}_n) = \int x d\hat{F}_n(x) = \frac{1}{n}(X_1 + \cdots + X_n).$$

ii) procjenitelj kvantila

$$T(F) = F^{\leftarrow}(u)\,,$$
$$T(\hat{F}_n) = \hat{F}_n^{\leftarrow}(u) = X_{(\lceil nu \rceil)}\,.$$

gdje zadnja nejednakost vrijedi za $u \in (0,1)$ i uzorke bez izjednačenih vrijednosti.

Intervali pouzdanosti

i) ako je

$$T - \theta \sim AN(0, se_T^2) \tag{2}$$

interval pouzd. $(1-\alpha)100\%$ za θ je

$$(T - z_{\alpha/2} se_T, T + z_{\alpha/2} se_T)$$

npr. $T = \overline{X}_n$ $se_T = \sigma/\sqrt{n}$, no za mnoge procjenitelje je se_T nepoznata.

ii) ako je

$$T - \theta \sim G$$

tada je interval pouzd. $(1-\alpha)100\%$ za θ

$$(T - q^G_{1-\alpha/2}, T - q^G_{\alpha/2})$$

gdje su q^G_u kvantili razdiobe G, koja je opet uglavnom nepoznata.

Ako je razdioba $\hat{F}_n \approx F$, za očekivati je da

$$T - \theta = T(\hat{F}_n) - T(F) \sim G \overset{d}{\approx} T(F_n^*) - T(\hat{F}_n) \sim G_n^* \,,$$

gdje je F_n^* empirijska razdioba n.j.d. uzorka

$$X_1^*, \ldots, X_n^*$$

s razdiobom $\hat{F}_n$, no ta razdioba se po volji točno može procjeniti Monte Carlo metodom, tj. uzimajući veliki broj uzoraka iz $\hat{F}_n$ razdiobe.

Za dani uzorak $X_1, \ldots, X_n$ i statistiku

$T = t(X_1, \ldots, X_n)$

Algoritam (bootstrap)

> $i = 1$
> repeat
> > sample $X^*_{i,1}, \ldots, X^*_{i,n} \overset{njd}{\sim} \hat{F}_n$
> > $T^*_i = t(X^*_{i,1}, \ldots, X^*_{i,n})$
> > $i = i + 1$
> until $i > B$

Bootstrap uzorci su odnosu na $\hat{F}_n$ isto kao i originalni u odn. na F. Često je ppravdano pretpostaviti da je

$$\text{razdioba od } T - \theta \approx \text{ razdioba od } T^*_i - T\,.$$

No razdiobu na desnoj strani možemo ako je B veliki proizvoljno dobro procijeniti.

$$X_i, \ldots, X_n \sim F\,,$$
$$T = t(X_i, \ldots, X_n) = T(\hat{F}_n)$$

razdioba od $T - \theta$ je nepoznata.

$$X^*_{i,1}, \ldots, X^*_{i,n} \sim \hat{F}_n\,, \qquad i = 1, \ldots, B$$
$$T^*_i = t(X^*_{i,1}, \ldots, X^*_{i,n}), \quad T = T(\hat{F}_n)$$

no razdioba od $T^*_i - T$ je "poznata".

Ako je za dani uzorak

$$T_1^* - T \overset{d}{\approx} T - \theta$$

intervali pouzdanosti uvedeni gore mogu se procijeniti, naime tada (uz asimpt. nepristranost i normalnost)

$$se_T \approx \sqrt{\operatorname{var} T} \approx \sqrt{\operatorname{var} T_1^*} \approx se_{boot}$$

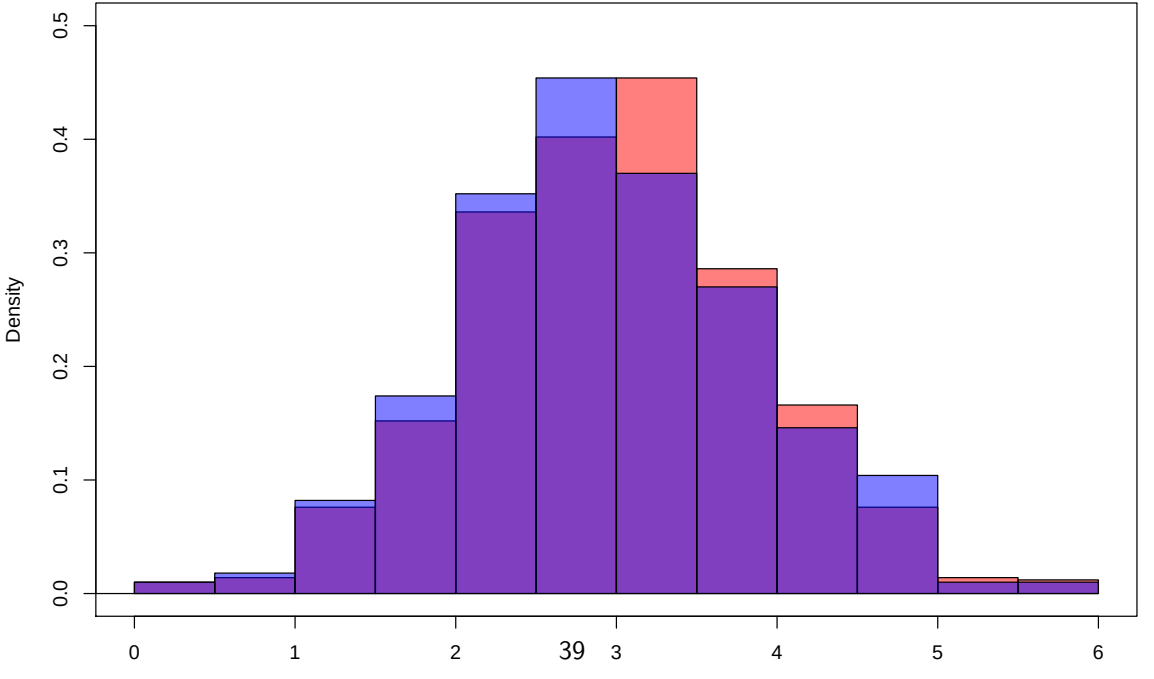
no $\operatorname{var} T_1^*$ uvjetno na uzorak znamo procijeniti (v. v_{boot} dolje).

Slično, kvantile razdiobe G, možemo zamijeniti kvantilima razdiobe

$$\hat{G}^*(x) = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}_{(\infty,x]}(T_i^* - T)$$

```r
library(boot)
nboot <- 10000 # Number of simulations
alpha <- .05 # alpha level
n <- 1000 # sample size
bootThetaQuantile <- function(x,i) {
quantile(x[i], probs=.5)  ### moze i drugi kvantil
}
bootThetaMean <- function(x,i) {
mean(x[i])
}

raw <- rnorm(n,3, 1) # raw data
bootsample <- replicate(1, raw[sample(1:length(raw), n, replace=
hist(raw, col=rgb(1,0,0,0.5),xlim=c(0,6),prob=T,ylim=c(0,0.5),ma
hist(bootsample, col=rgb(0,0,1,0.5),prob=T, add=T)
box()
```

```
##
## Attaching package: 'boot'

## The following object is masked from 'package:car':
##
##     logit
```



Bootstrap varijance

$$v_{boot} = \frac{1}{B}\sum_{i=1}^{B}\left(T_i^* - \frac{1}{B}\sum_{i=1}^{B}T_i^*\right)^2$$

Bootstrap pristranosti

$$\text{bias}_{boot} = \frac{1}{B}\sum_{i=1}^{B}(T_i^* - T)$$

Bootstrap intervali pouzdanosti

i) ako vrijedi asimptotska normalnost (i nepristranost) za $T-\theta$, $(1-\alpha)100\%$ interval pouzdanosti možemo procijeniti sa

$$T \pm z_{\alpha/2}\sqrt{v_{boot}}$$

ii) ako to i ne vrijedi možemo napraviti *pivotalni bootstrap* interval pouzdanosti tako da stavimo

$$(T - (q^{T*}_{1-\alpha/2} - T), T - (q^{T*}_{\alpha/2} - T))$$

gdje je q^{T*}_u u –kvantil uzorka T^*_i–ova.

Dodatni bootstrap intervali pouzdanosti

iii) gornji interval se oslanja na približnu jednakost razdiobi $T-\theta$ i T_i^*-T, no i kad ona ne vrijedi, *studentizirani pivotalni bootstrap* interval pouzdanosti $(1-\alpha)100\%$ možemo konstruirati ako procijenimo standardnu grešku za T $s.e.(T)$ t.d. prvo transformiramo

$$\tau_i^* = \frac{T_i^* - T}{s.e.(T_i^*)}$$

gdje $s.e.(T_i^*)$ izračunamo po formuli (ako znamo) ili ga možemo procijeniti npr. *2nd level bootstrap* procedurom i damo interval

$$(T - q_{1-\alpha/2}^{\tau*}\sqrt{v_{boot}}, T - q_{\alpha/2}^{\tau*}\sqrt{v_{boot}})$$

gdje je $q_u^{\tau*}$ u–kvantil razdiobe τ_i^*–ova.

iv) manje precizan no vrlo jednostavan je direktni *percentilni bootstrap* interval pouzdanosti $(1-\alpha)100\%$ možemo konstruirati kao

$$(q^{T*}_{\alpha/2}, q^{T*}_{1-\alpha/2})$$

gdje je q^{T*}_u u–kvantil razdiobe T^*_i–ova (v. npr. Wasserman, 2006)

v) Postoje i *biased corrected adjusted (BCa)* bootstrap intervali istog tipa

$$(q^{T*}_{\alpha'}, q^{T*}_{\alpha''})$$

gdje su α', α'' pomno odabrani umjesto $\alpha/2, 1-\alpha/2$ da poboljšaju svojstva intervala u iv).

```r
theta.boot.median <- boot(raw, bootThetaQuantile, R=nboot)
boot.ci(theta.boot.median, conf=(1-alpha),type=c("norm","basic",
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = theta.boot.median, conf = (1 - alpha), typ
##     "basic", "perc", "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 2.946,  3.106 )   ( 2.961,  3.103 )
##
## Level     Percentile            BCa
## 95%   ( 2.933,  3.076 )   ( 2.932,  3.075 )
## Calculations and Intervals on Original Scale
```

```r
theta.boot.mean <- boot(raw, bootThetaMean, R=nboot)
boot.ci(theta.boot.mean, conf=(1-alpha),type=c("norm","basic","p
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = theta.boot.mean, conf = (1 - alpha), type
##     "basic", "perc", "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 2.939,  3.051 )   ( 2.938,  3.051 )
##
## Level     Percentile            BCa
## 95%   ( 2.940,  3.053 )   ( 2.939,  3.051 )
## Calculations and Intervals on Original Scale
```

Bootstrap se oslanja na teoreme koji pokazuju da za njd podatke uz dodatne uvjete na glatkoću funkcionala $T = T(F)$ (u odn. na nepoznatu razdiobu!!??) vrijedi npr. za $n \to \infty$

$$\sup_u \left| P_{\hat{F}_n} \left(\sqrt{n}(T_1^* - T) \leq u \right) - P_F \left(\sqrt{n}(T - \theta) \leq u \right) \right| \xrightarrow{gs} 0$$

Bootstrap procedura je opravdana samo za statistike za koje vrijede uvjeti ovakvih teorema, iako to uključuje većinu primjera koje će nas zanimati, treba znati da bootstrap nije uvijek opravdan i ne možemo ga slijepo primjenjivati!!