Kodutöö näidilahendus

1. Ülesanne

Hindame Tallinna börsi tootlust kasutades geomeetrilist keskmist.

```
# Defineerime funktsiooni, mis arvutab vaatluste geomeetrilise
    keskmise:
GeomKeskmine=function(x){
  exp(mean(log(x)))
}

# Sisestame vaatlused ehk tootluste protsendid:
tootlused=c(47.2, 72.6, -23.9, 38.22, 11.38, -7.66, 19.06, 19.63,
    15.49, -6.38, 10.05, 5)

# Leiame keskmise tootluse:
gkeskm=GeomKeskmine(100+tootlused)-100
gkeskm
```

Ehk keskmine tootlus (geomeetriline keskmine) on 14,1

Kui täpne on saadud hinnang? Leiame hinnangu standardvea kasutades bootstrap-meetodit:

```
k=100000
BShinnangud=rep(NA,k) # Bootstrap hinnangud salvestame siia

for(i in 1:k){
  # Võtame tagasipanekuga valimi mahuga 12
  BSvalim=sample(tootlused, length(tootlused), replace=TRUE)
  # Leiame bootstrap-valimi põhjal hinnangu:
  BShinnangud[i]=GeomKeskmine(100+BSvalim)-100
}

# Bootstrap-meetodil leitud standardviga:
sd(BShinnangud)

# Ehk tootluse hinnang ja tema standardviga bootstrap-meetodil:
paste(round(gkeskm,1), "\u00B1", round(sd(BShinnangud),1))
```

Tallinna börsi keskmiseks tootluseks saame seega 14.1 ± 7. Standardvea hinnang 7 on leitud siin bootstrap-meetodit kasutades.

b) Leiame hinnangu täpsuse ka Delta-meetodil. Valimi keskmise ehk suuruse

$$\hat{\theta} = \overline{\ln(100 + tootlusprotsent)}$$

dispersiooni on kerge leida — hindame lihtsalt suuruste $\ln(100+tootlusprotsent)$ dispersiooni ja jagame saadud tulemuse vaatluste arvuga:

$$\begin{aligned}\hat{\mathrm{D}}(\hat{\theta}) &= \hat{\mathrm{D}}\left(\overline{\ln(100 + tootlusprotsent)}\right) \\ &= \hat{\mathrm{D}}(\ln(100 + tootlusprotsent))/n \\ &= 0{,}04853828/12 \\ &= 0{,}004044857.\end{aligned}$$

Meil on aga tarvis leida geomeetrilise keskmise hinnangu dispersiooni ehk juhusliku suuruse

$$\exp(\hat{\theta}) - 100$$

dispersiooni. Konstandi liitmine või lahutamine juhusliku suuruse dispersiooni ei muuda — seega võime suuruse -100 dispersiooni leidmiseks vajalikke arvutusi tehes ka unustada.

Delta meetod ütleb, et

$$\mathrm{D}\left(f(\hat{\theta})\right) \approx \mathrm{D}(\hat{\theta}) \cdot \left(f'(\hat{\theta})\right)^2 .$$

Antud juhul $f(x) = \exp(x)$, $f'(x) = \exp(x)$ ja $(f'(\hat{\theta}))^2 = \exp(\hat{\theta})^2$, seega

$$\hat{\mathrm{D}}(f(\hat{\theta})) \approx 0{,}004044857 \cdot \exp(4.737454)^2 = 52{,}699\ldots$$

Antud arvutust tehes tasub meeles pidada, et praegu

$$\hat{\theta} = \overline{\ln(100 + tootlusprotsent)} = 4{,}737\ldots.$$

Seega annab Delta meetod standardvea ligikaudseks hinnanguks $\sqrt{52{,}699} = 7{,}259\ldots$.

2. Ülesanne

Otsitava parameetri p hindamiseks võime kasutada suurima tõepära hinnangut. Suurima tõepära hinnangu leidmiseks paneme esmalt kirja tõepärafunktsiooni:

$$L(p) = \left(\frac{2p}{3}\right)^{n_0} \left(\frac{p}{3}\right)^{n_1} \left(\frac{2(1-p)}{3}\right)^{n_2} \left(\frac{1-p}{3}\right)^{n_3},$$

Kus n_0 on valimis olevate nullide arv, n_1 on nähtud ühtede arv jne.

Maksimiseerida on lihtsam log-tõepära, seega leiame ka selle:

$$l(p) = n_0 \log(2p/3) + n_1 \log(p/3) + n_2 \log(2(1-p)/3) + n_3 \log((1-p)/3),$$

Maksimumi leidmiseks võtame tuletise otsitava parameetri p järgi:

$$\frac{\partial l}{\partial p} = n_0/p + n_1/p + n_2/(2(1-p)/3) \cdot (-2/3) + n_3/((1-p)/3) \cdot (-1/3),$$

ja võrdsustame saadud tulemuse nulliga:

$$\begin{aligned}
&\frac{\partial l}{\partial p} = 0 \\
&0 = n_0/p + n_1/p + n_2/(2(1-p)/3) \cdot (-2/3) + n_3/((1-p)/3) \cdot (-1/3) \mid \cdot p(1-p), \\
&0 = n_0(1-p) + n_1(1-p) - n_2 p - n_3 p \\
&0 = (n_0 + n_1) - p(n_0 + n_1 + n_2 + n_3) \\
&p = (n_0 + n_1)/(n_0 + n_1 + n_2 + n_3).
\end{aligned}$$

Kas saime tulemuseks ikka maksimumi? Kontrollime seda. Log-tõepära teine tuletis

$$\frac{\partial^2 l}{\partial p^2} = -n_0/p^2 - n_1/p^2 - n_2/(1-p)^2 - n_3/(1-p)^2$$

on negatiivne mistahes p väärtuse korral mis jääb vahemikku $0 \ldots 1$, seega oleme tõepoolest leidnud tõepärafunktsiooni maksimumi.

Tuletatud üldvalemit kasutades saame leida hinnangu ka oma valimi jaoks:

$$\hat{p} = 5/10 = 0{,}5.$$

Hinnangu täpsuse ehk standardvea leidmiseks võime kasutada näiteks parameetrilist bootstrappi:

```
hinnang=function(x){sum(x<=1)/length(x)}
BShinnang=rep(NA, 100000)

for (i in 1:100000){
valim=sample(0:3, 10, prob=c(1/3, 1/6, 1/3, 1/6), replace=TRUE)
BShinnang[i]=hinnang(valim)
}
sd(BShinnang)
```

Ülaltoodud programmi jooksutades saame standardvea ligikaudseks väärtuseks $\approx 0{,}16$. Seega leitud hinnang parameetrile p on $0{,}5 \pm 0{,}16$.

Antud juhul on võimalik leida ka täpsem hinnang standardveale. Nimelt võime tähele panna, et hinnang $\hat{p}$, on vaadeldav kui juhuslike suuruste $X_i, i = 1, \ldots, n$ keskmine, kusjuures $X_i = 1$ kui esialgne juhuslik suurus omandab väärtuse 0 või 1. Muidu on juhusliku suuruse X_i väärtus null, $X_i = 0$. Bernoulli jaotusega juhuslike suuruste X_i keskmise dispersiooni on aga lihtne leida:

$$\mathrm{D}(\hat{p}) = P(0 \text{ või } 1)(1 - P(0 \text{ või } 1))/n = p(1-p)/n,$$

sest $P(0 \text{ või } 1) = 2p/3 + p/3 = p$. Seega alternatiivne hinanng hinnangu dispersioonile oleks $\hat{\mathrm{D}}(\hat{p}) = 0{,}5 \cdot 0{,}5/10 = 0{,}025$ ja standardveaks tuleks siis $\sqrt{0{,}025} = 0{,}158$.

3. Ülesanne

```
# Andmete tekitamine:
set.seed(1)
naised=rnorm(120, mean=168, sd=6)
mehed=rnorm(80, mean=182, sd=7)
pikkused=c(mehed, naised)
rm(naised, mehed)

# Paneme kirja log-tõepära
ll = function(arg, andmed){
 mu_mees=arg[1]
 sigma_mees=exp(arg[2])
 mu_naine=arg[3]
 sigma_naine=exp(arg[4])
 meeste_osakaal=80/200
 print(paste(mu_mees, sigma_mees, mu_naine, sigma_naine, meeste_
     osakaal))
 l=sum(log( dnorm(andmed, mean=mu_mees, sd=sigma_mees )*meeste_osakaal
      +
```

```
    dnorm(andmed, mean=mu_naine, sd=sigma_naine )*(1-meeste_osakaal)
       ))
 return(l)
}

# Maksimiseerime tõepärafunktsiooni numbriliste meetodite abil:
# Peame andma ligikaudsed (oletuslikud) väärtused otsitavatele
    parameetritele:
tul=optim(c(180, log(10), 160, log(10)), ll,
  andmed=pikkused, control=list(fnscale=-1))


# Tulemused:

# Hinnatud meeste pikkuste keskväärtus
tul$par[1]
# Hinnatud meeste pikkuste standardhälve
exp(tul$par[2])

# Hinnatud naiste pikkuste keskväärtus
tul$par[3]
# Hinnatud naiste pikkuste standardhälve
exp(tul$par[4])
```

Seega saame praegu meeste keskmise pikkuse hinnanguks 180,77 cm (standardhälve 7,78) ja naiste keskmiseks pikkuseks 169,08 (standardhälve 5,47).

Mõtteaineks:

1. Võrdle saadud tulemusi tegelike väärtustega!

2. Kas oskaksid seda ülesannet lahendada ka siis, kui meeste osakaalu poleks teada?