Tõenäosusteooria ja statistika II

Kodutöö

Ülesanne 1 – hinnangu täpsuse kirjeldamine

Sissejuhatus:

Soovime kogutud andmete põhjal (elementaarosakeste kokkupõrgete arvud ajaühiku jooksul) hinnata tõenäosust, et saame uuritava tunnuse väärtuseks nulli. Kogutud andmed ise on järgmised:

2 3 2 3 3 2 1 0 3 0 2 0 1 0 1 0 1 1 3 1 1 0 2 0 0

Nulli saamise tõenäosust võime hinnata kahel viisil:

Meetod 1: $\hat{p}_0 = \frac{nullide\ arv}{vaatluste\ arv}$ ehk võime kasutada nulli nägemise suhtelist sagedust.

Meetod 2: Oletame, et uuritav tunnus on Poissoni jaotusega (kokkupõrgete arv võiks käituda kui Poissoni jaotusega juhuslik suurus). Hindame Poissoni jaotuse parameetri (suurima tõepära hinnang Poissoni jaotuse parameetrile on kõigi vaatluste keskmine, $\hat{\lambda} = \overline{X}$). Leiame, milline on Poissoni jaotusega juhusliku suuruse (parameetriga $\hat{\lambda}$ Poissoni jaotus) tõenäosus omandada väärtust 0 ja kasutame saadud tõenäosust kui oma hinnangut.

Küsimused:

a) Hinda nulli nägemise tõenäosust mõlema kirjeldatud meetodi abil.

b) Lisa mõlemale hinnangule ka hinnangu standardviga. Kui 1. meetodi korral saad standardvea hinnata ka klassikalisel meetodil teadaoleva valemi abil (suhtelise sageduse dispersiooni sa ju oskad leida/hinnata) , siis teise hinnangu jaoks võib standardvea leidmine olla veidi keerulisem ettevõtmine. Hinda teise hindamismeetodi täpsust nii Delta-meetodi abil kui ka bootstrap meetodil (otsusta ise, kas kasutad parameetrilist või mitteparameetrilist bootstrap-meetodit, põhjenda miks valisid just selle variandi). Delta meetodi jaoks lisa tuletuskäik/arvutusvalem kuidas täpselt sa otsitava standardvea arvutasid; bootstrap meetodi korral lisa lisaks saadud hinnangu standardveale ka R-i programm, mida kasutasid arvutuste tegemiseks.

c) Põhjenda kumba meetodit kasutaksid sina nulli esinemistõenäosuse hindamisel ja miks eelistad just seda meetodit.

Ülesanne 2 - Suurima tõepära hinnangu leidmine numbrilise maksimiseerimise abil

Näidised.

Vahel on tõepära või log-tõepärafunktsioon piisavalt keeruline, et tõepärafunktsiooni maksimumi leidmine osutub tülikaks ülesandeks. Sellisel juhul võime hinnangu (ligikaudseks) leidmiseks kasutada numbrilisi meetodeid. Vaatame alljärgnevalt näidet, kuidas saab R-is kasutada numbrilist maksimiseerimist.

Oletame antud juhul, et vaatlused peaksid olema normaaljaotusega (sellisel juhul oskame parameetrite hinnanguid suurima tõepära meetodil leida ka täpselt ja seega saame hiljem numbriliste meetodite abil leitud hinnanguid võrrelda täpsete tulemustega):

```
# Vaatlused
x=c(12, 14, 16)

# Suurima tõepära meetodil leitud hinnangud keskväärtusele ja
#     standardhälbele:
mean(x)
sqrt(mean((x-mean(x))**2))

# Hinnangud numbriliste meetodite abil

# Variant 1

# defineerime log-tõepära
l=function(arg, andmed){
 mu=arg[1]
 sigma=arg[2]
 l=sum(log(   dnorm(andmed, mean=mu, sd=sigma )   ))
 return(l)
}

# Maksimiseerime log-tõepära (funktsiooni l):
tul=optim(c(2,2), l, andmed=x, control=list(fnscale=-1))
tul

# Leitud parameetrite hinnangud
tul$par
# Log-tõepära maksimaalne väärtus mida õnnestus saavutada:
tul$value
```

Võrdle saadud hinnanguid tavapäraste hinnangutega:

```
> mean(x)
[1] 14
> sd(x)
[1] 2
```

Kas saadud hinnangud tulid samasugused kui `optim`-käsu abil leitud hinnangud? Ühe parameetri hinnang peaks tulema üsna sarnane tavapärasele hinnangule, teine on aga mõnevõrra erinev. Kas oskad öelda miks?

Eelnev programm võib anda välja ka hoiatusi, sest numbrilised meetodid võivad proovida ka negatiivseid sigma väärtuseid – me pole antud programmile ette öelnud, et teise parameetri väärtused peavad olema mittenegatiivsed. Üheks lahenduseks oleks tagada, et sigma väärtus on alati mittenegatiivne – ükskõik mis siis teise parameetri väärtuseks ka poleks. Seda saab teha näiteks nii:

```
l2=function(arg, andmed){
 mu=arg[1]
 sigma=exp(arg[2])
 l=sum(log(   dnorm(andmed, mean=mu, sd=sigma )   ))
 return(l)
}

x=c(10,12,14)
tul=optim(c(2,2), l2, andmed=x, control=list(fnscale=-1))
# Hinnatud keskväärtus
tul$par[1]
# Hinnatud standardhälve
exp(tul$par[2])
```

Sissejuhatus:

Ühe küsitluse käigus mõõdeti tudengite pikkuseid. Mõõtmistulemused saad R-i lugeda järgmise käsu abil:

```
print(load(url("http://www-1.ms.ut.ee/mart/TS2/pikkused.RData")))
hist(pikkused)
# Esimese kümne tudengi pikkused:
pikkused[1:10]
```

Soovime hinnata nii nais- kui ka meestudengite keskmist pikkust; nii naiste kui ka meeste pikkuste varieeruvust (standardhälbeid) ja muidugi ka naiste osakaalu. Paraku pole meil kirjas mis soost tudeng ühe või teise pikkusega oli – meil on ununenud mõõtmata tudengi sugu.

Õnneks teame üsna kindalt, et nii meeste kui ka naiste pikkuste jaotuseks võiks olla normaaljaotus – pikkuse tinglikud jaotused tingimusel et sugu on antud on normaaljaotused. Seega on pikkuste marginaaljaotus kirja pandav kui jaotuste segu (meenuta täistõenäosuse valemit):

$$F_{PIKKUS}(x) = naiste\_osakaal\ F_{PIKKUS|Naine}(x) + (1-naiste\_osakaal)\ F_{PIKKUS|Mees}(x)$$

Ehk, alternatiivselt kirja pandult:

$$f_{PIKKUS}(x) = naiste\_osakaal\ f_{PIKKUS|Naine}(x) + (1-naiste\_osakaal)\ f_{PIKKUS|Mees}(x)$$

kus $f_{PIKKUS|Naine}(x)$ on parameetritega μ_{naine}; σ_{naine} normaaljatuse tihedusfunktsioon ja $f_{PIKKUS|Mees}(x)$ on parameetritega μ_{mees}; σ_{mees} normaaljatuse tihedusfunktsioon.

Küsimused:

1. **Milline näeb välja log-tõepära tudengi pikkuste andmete jaoks? Pane kirja R-i funktsioon mis leiab etteantud parameetrite vektori jaoks log-tõepära väärtuse.** Antud juhul on tudengite pikkuste jaotusel viis tundmatut parameetrit: μ_{naine}; σ_{naine}; μ_{mees}; σ_{mees} ja naiste_osakaal.

2. **Hinda nende viie parameetri väärtused numbriliste meetodite (näiteks optim-käsu) abil**. Muretse ka selle pärast, et kõigi parameetrite hinnangud jääksid lubatud piiridesse (näiteks ei soovi me näha

negatiivseid standardhälbeid). Millised hinnangud saad? Kommenteeri saadud hinnanguid, on need usutavad? Kui mitte, siis arutle selle üle, mis võis valesti minna.

3. **Lisa tööle kasutatud R-i programm** (milline näeb välja log-tõepära väärtust arvutav programm; kuidas kasutasid optim-käsku; kuidas optim-käsu tulemustest lugesid välja hinnangute väärtused…