Estimatzaileak

Demagun, egunero ordu ezberdin batean autobus geltokira joaten zarela, ordua begiratu gabe, eta bertan zain egoten zarela L1 lineako autobusa etorri arte. Suposatuz autobusa zehazki θ minuturo pasatzen dela, geltokian itxaroten pasatzen dudan denborak (minututan) $\mathcal{U}(0,\theta)$ banaketari jarraitzen diola esan dezakegu, honako dentsitate funtzioarekin:

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0 & \text{beste kasuetan} \end{cases}$$

Parametroen estimazioa momentuen eta egiantz handieneko metodoaren bitartez

Suposatu dezagun, θ, parametroa, hau da autobusaren frekuentzia (minututan), estimatu nahi dugula. Horretarako $n = 150$ egunetan zehar itxarondako denbora (minututan) gorde dugu, honako behaketak lortuz:

```
lagina <- c(1.12, 2.58, 2.36, 0.40, 3.29, 1.54, 0.25, 4.87, 1.00, 4.26,
            1.43, 0.58, 0.21, 1.84, 3.75, 0.56, 1.38, 4.46, 3.37, 3.41,
            2.53, 3.38, 3.16, 3.26, 2.17, 3.64, 2.31, 3.26, 1.82, 3.26,
            0.87, 1.07, 1.74, 4.32, 4.65, 0.87, 0.76, 4.51, 0.11, 1.38,
            4.34, 4.55, 4.13, 1.42, 2.12, 0.01, 3.36, 0.01, 0.20, 2.85,
            2.15, 4.30, 1.06, 3.06, 1.58, 1.32, 1.53, 0.94, 0.55, 0.18,
            1.04, 0.00, 3.84, 1.86, 2.84, 4.39, 3.85, 0.85, 4.27, 1.72,
            1.31, 0.69, 4.40, 3.22, 4.23, 4.49, 4.89, 0.44, 1.55, 3.70,
            3.51, 1.13, 0.18, 0.47, 4.65, 4.60, 1.01, 4.15, 0.54, 4.48,
            0.17, 1.08, 1.41, 4.48, 4.14, 1.36, 1.20, 1.19, 2.32, 3.21)
```

Momentuen metodoa

Lehenengo atal honetan banaketa honen parametroa (θ, autobusaren frekuentzia) momentuen metodoaren bitartez estimatzen saiatuko gara. Honetarako, banaketa uniformearen itxaropena $EX = \frac{\theta}{2}$ izanik, $EX = \bar{x}$ ekuazioa ebatzi behar dugu, edo baliokidea dena, $f(X) = EX - \bar{x}$ funtzioaren erroak bilatu. Kasu honetan, ekuazio honen soluzioa bilatzea tribiala da (eskuz egin daiteke):

```
#Estimatzailea edozein laginerako lortzeko funtzio orokorra
getThetaMM <- function(lagina){
  2*mean(lagina)
}

#Gure laginerako estimazioaren kalkulua
getThetaMM(lagina)
```

```
## [1] 4.605
```

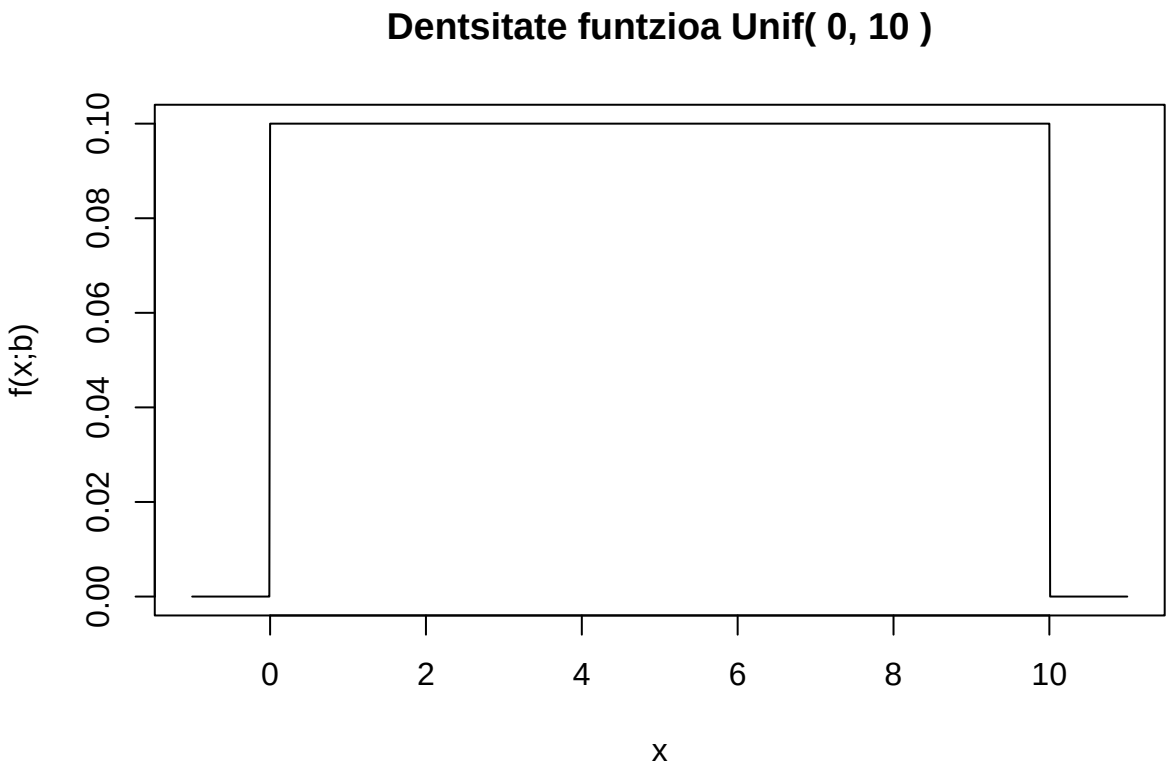
Ariketa: Bete aurreko kodean falta dena eta kalkulatu gure laginerako momentuen metodoaren bidezko estimazioa.

Hala ere, kasu guztietan ez da hain erraza izango ekuazio hau ebaztea eta batzuetan zenbakizko metodoetara (hurbilketa metodoetara) jo beharko dugu. Horretarako, itxaropenaren funtzioa definitu beharko dugu θ parametroaren menpe:

```
itxaropena_unif <- function(theta){
  return(theta/2)
}
```

Ariketa: Erabili 'itxaropena_unif' funtzioa banaketa uniforme ezberdinen itxaropenak kalkulatzeko. Irudikatu banaketa uniforme ezberdinen dentsitate funtzioa eta lerro berde batekin bere itxaropena. Saiatu ulertzen zergatik den banaketa uniformearen itxaropena $\theta/2$.

```
b <- 10
x <- seq(-1, b+1, 0.01)
plot(x, dunif(x, min = 0, max = b), t="l", main=paste("Dentsitate funtzioa Unif( 0,", b, ")"), xlab = "
```



Orain, `uniroot` funtzioa erabiliz $f(X) = EX - \bar{x}$ funtzioaren erroak bilatu ditzakegu:

```
getThetaMM2 <-uniroot(function(theta) itxaropena_unif(theta)-mean(lagina),
                      interval=c(-10,10))$root
getThetaMM2
```

```
## [1] 4.605
```

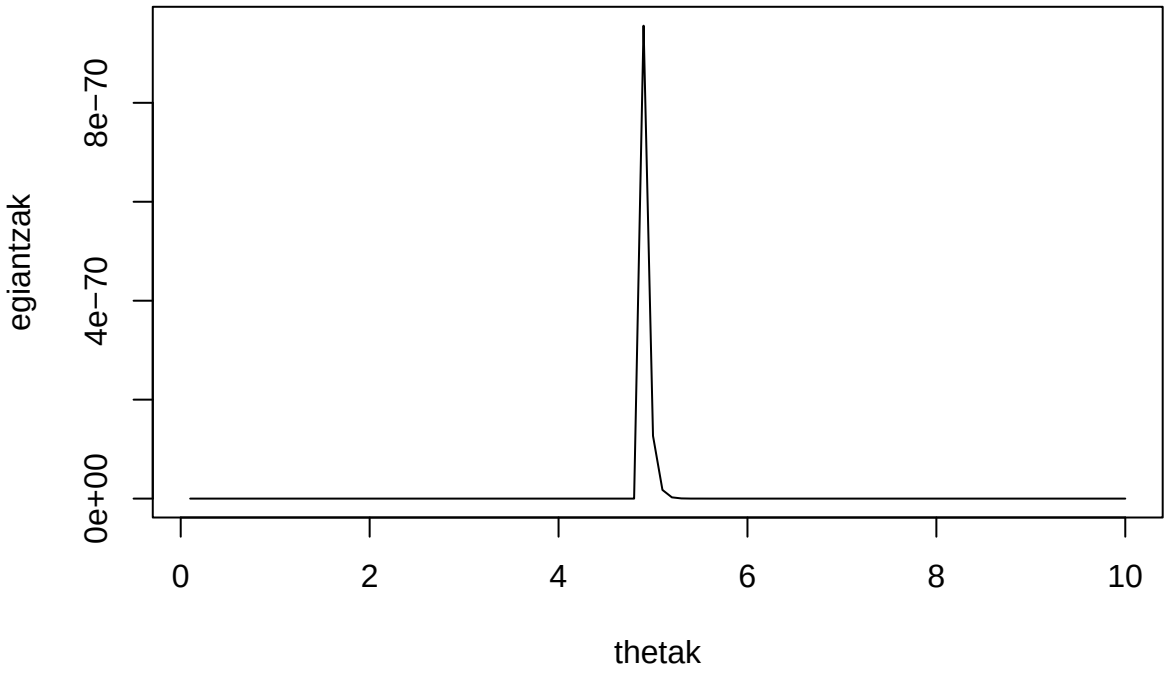
Kasu honetan ere, $\hat{\theta}_{MM} = 4.605$ lortu dugu ekuazioa ebazteko oso sinplea izanik, zenbakizko metodoak balio zehatza topatu duelako.

Egiantz handieneko metodoa

Orain, egiantza handieneko metodoaren bidezko estimazioa lortzen saiatuko gara. Horretarako, definitu dezagun, hasteko, egiantz funtzioa $\mathcal{U}(0, \theta)$ banaketarako eta irudikatu dezagun funtzioa gure laginerako:

```
getLikelihood <- function(theta, sample){
  l <- 0
  if(all(sample<=theta) & all(sample>=0)){
    return(prod(dunif(sample, min=0, max=theta)))
  }
  return(l)
}


thetak <- seq(0.1, 10, by=0.1)
egiantzak <- sapply(thetak, getLikelihood, sample=lagina)
plot(thetak, egiantzak, type="l")
```



Ariketa: Zein balioren inguruan dago egiantz funtzioaren maximoa? Lortutako balioa momentuen metodoarekin lortutakoaren antzekoa al da?

Erantzuna: 4.89 balioraen inguruan dago, momentuen metodoarekin lortutakoaren antzekoa da, baina ez berdina.

Orain egiantz handieneko estimazioa lortzeko, funtzio honen maximoa bilatu behar dugu, baina errazagoa izango da bere logaritmoa aztertzea (leunagoa izango da orohar eta propietate egokiagoak izango ditu zenbakizko metodoak erabiltzeko).

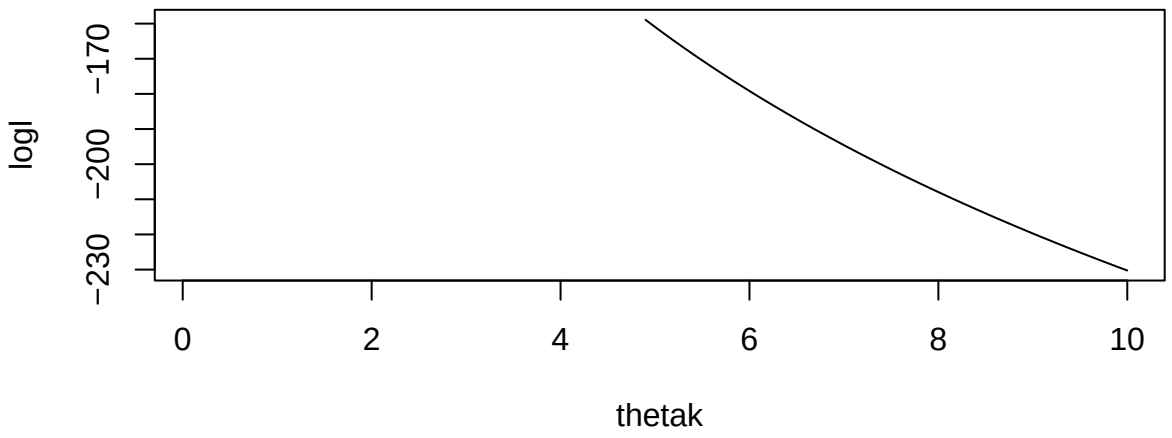
```r
getLogLikelihood  <- function(theta, sample){
  l <- getLikelihood(theta, sample)
  if(l!=0){
    return(log(l))}
  else{
    # -Inf balioa ekiditeko
    return(-Inf)
  }
}
```

Aurreko funtzioak ez du arazorik emango lagin-tamaina txikiekin. Lagin-tamaina handiekin ordea, arazoak sor daitezke egiantzaren balioa 0-ra hurbilduko delako. Arazo hori ekiditeko probabilitateen logaritmoen batura erabiliz kalkulatuko dugu, probabilitateen biderkaduraren logaritmoaren ordez.

```r
getLogLikelihood <- function(theta, sample){
  probs <- dunif(sample, min=0, max=theta)
  ll <- sum(log(probs))
  return (ll)
}
```

Grafikoki irudikatuko dugu.

```r
logl <- sapply(thetak, FUN=getLogLikelihood, sample=lagina)
plot(thetak, logl, type="l")
```



Erreparatu zenbait baliotarako $-\infty$ balioa duela egiantzaren logaritmoak eta ondorioz ezin dela irudikatu.

Funtzioaren maximoa begiz non kokatzen den ikusten dugu baina balio zehatza lortzeko zenbakizko metodoak erabili behar ditugu. Horretarako `nlm` funtzioa erabiliko dugu. Deribatuan oinarritutako zenbakizko metodoak erabiliz (hurbilketa metodoak) funtzioen *minimoak* topatzeko balio du funtzio honek. Guk maximoa bilatu nahi dugunez, maximizatu nahi dugun funtzioaren negatiboa definitu beharko dugu hasteko. Gainera, optimizazio prozesuan infinitoruntz doazen balioak arazoak sortu ditzaketenez, balio negatibo oso altuekin ordezkatuko ditugu.

```r
minusL <- function(theta, sample){
  ll <- getLogLikelihood(theta, sample)
  if (is.infinite(ll)) {
    ll <- -1000
  }
```

```
  return(-ll)
}
```

Ondoren `nlm` funtzioa aplikatuko dugu. Funtzioak hasierako balio bat behar du eta balio hori egiantza 0 ez den eremuan kokatuta egotea beharrezkoa da, optimizazio algoritmoak deribatuarekin arazorik ez izateko. Hasierako parametro hori bakoitzean ez emateko funtzioari defektuzko balio handi bat erabiliko dugu. Komeni denean, defektuzko balio hori aldatu ahalko dugu. Behin funtzioa definituta, laginak ematen digun estimazioa kalkulatuko dugu.

```
#Estimatzailea edozein laginerako lortzeko funtzioa
getThetaML <- function(sample, initial.theta=50){
  return(nlm(f=minusL, p=initial.theta, sample=sample))
}
getThetaML(lagina)
```

```
## $minimum
## [1] 158.7192
##
## $estimate
## [1] 4.89
##
## $gradient
## [1] -860195
##
## $code
## [1] 3
##
## $iterations
## [1] 37
```

```
theta.ehe <- getThetaML(lagina)$estimate
```

Beraz, lortutako estimazioa $\hat{\theta}_{EHE} = 4.8900004$ da.

Ariketa: Zergatik ariketa guzti honetan zehar estimazioa hitza erabili dugu estimatzailea beharrean? Zein da ezberdintasuna?

Ariketa: Begiratu klaseko ariketetan edo interneten (wikipedian, adibidez), banaketa uniformearen egiantz handieneko estimatzailea zein den. Kalkulatu, formula hori erabiliz, egiantz handieneko estimazioa kasu honetarako. Aurreko adibideetan zenbakizko metodoekin lortutako balioa asko urruntzen al da benetako egiantz handieneko estimaziotik?

Ariketa: Jakinda autobusen frekuentziak zenbaki osoak izaten direla normalean, eta lortutako estimazioetan oinarrituta, zein uste dezu izan daitekeela autobusaren frekuentziaren benetako balioa? Nolako erroreak egin ditugu estimazioekin?

Estimatzaileen propietateak

Orain, jakinda benetako populazioak $\mathcal{U}(0,5)$ jarraitzen duela, azter dezagun zer gertatzen den momentuen ($\hat{\theta}_{MM} = 2\bar{X}$) eta egiantz handieneko ($\hat{\theta}_{EHE} = max\{X_1, X_2, ..., X_n\}$) estimatzaileekin lagina handitzen denean. Horretarako, sor ditzagun n tamaina ezberdinetako $N = 1000$ lagin eta ikus ditzagun lortutako estimazioen batezbestekoa eta bariantza nola aldatzen diren:

```
N <- 1000
lagin.tamainak <- c(10, 20, 30, 40, 50, 75, 100, 150, 200, 250, 300, 350)
set.seed(323)
#Hemen gordeko ditugu estimazioen batezbesteko eta bariantzak
est <- c()
for(n in lagin.tamainak){
  #Sortu laginak
  laginak <- matrix(runif(n*N, 0, 5), nrow=N, ncol=n)
  #Kalkulatu estimazioak lagin guztientzat
  aux<- t(apply(laginak, 1,
                function(x){c(getThetaMM(x),
                              getThetaML(x)$estimate)}))
  #Kalkulatu estimazio guztien batezbestekoak eta bariantzak
  bb <- colMeans(aux)
  vars <- apply(aux, 2, var)
  est <- rbind(est, c(bb,vars))
}
```

Erreparatu `est` matrizean batezbestekoak (lehen bi zutabetan) eta bariantzak (3. eta 4. zubatetan) jaso ditugula. Lagin tamainaren araberako eboluzioa ikusteko grafikoki irudikatuko ditugu.

```
#colnames(est) <- c("bb_MM", "bb_EHE", "var_MM", "var_EHE")

#Irudikatu batezbestekoak eta bariantzak
par(mfrow=c(1,2))
plot(lagin.tamainak, est[,1],type="l", ylim=c(4,5.5), main="Batezbestekoak", col="blue", xlab="Lagin-ta
lines(lagin.tamainak, est[,2], type="l", col="red", xlab="")
legend(1, 5.5, legend=c("MM", "EHE"),
       col=c("blue", "red"), lty=1, cex=0.8)

plot(lagin.tamainak, est[,3],type="l", main="Bariantzak", col="blue", xlab="Lagin-tamainak")
lines(lagin.tamainak, est[,4], type="l", col="red", xlab="")
legend(2, 0.8, legend=c("MM", "EHE"),
       col=c("blue", "red"), lty=1, cex=0.8)
```

Ikusten dugun moduan bi estimatzaileak asintotikoki alboragabeak direla dirudi grafikoki (n handitzen denean 5ra hurbiltzen dira batezbestekoak) eta bariantza ere asintotikoki 0ra doala ikusten dugu bi kasuetan. Beraz, tinkoak direla esan dezakegu.

Ariketa: Irudikatu sortutako laginetako batzuk. Nolakoak dira? Eta $\hat{\theta}_{MM}$ eta $\hat{\theta}_{EHE}$ estimazioak? Antzeko balioak lortzen dira lagin ezberdinetan? Zentzua du horrek?

Ariketa: Egiantz handieneko estimatzailearen batezbestekoa beti da 5 baino txikiagoa. Arrazonagarria al da hau? Nolakoa izango da estimatzaile honen alborapenaren zeinua beti?

Ariketak

Eraiki berri dugun sistemak eskaintzen dituen zerbitzuen itxaron denborak aztertu nahi ditugu. Denbora horiek egoki adierazten dira $X \sim Gamma(k, \theta)$ banaketarekin, hau da, gure populazioa adierazteko eredu egokia eskaintzen du $X \sim Gamma(k, \theta)$ banaketak. Gure helburua parametro horien balio egokiak topatzea da, behatutako lagineko informazioan oinarrituz. Honako lagina lortu dugu:

```
lagina <- c(77.551, 45.195, 50.626, 39.878, 29.137, 57.321, 39.140, 66.776,
            48.028, 42.325, 31.200, 38.632, 42.914, 60.969, 22.076, 52.446,
            45.257, 42.626, 62.504, 22.684, 69.196, 42.383, 61.339, 45.803,
            74.707, 33.048, 72.423, 43.670, 65.279, 42.714, 59.785, 101.742,
            59.641, 44.749, 44.161, 58.488, 46.448, 25.280, 67.619, 66.846,
            80.208, 98.492, 41.149, 40.395, 22.220, 34.628, 77.768, 48.161,
            48.909, 66.267)
```

Hasi baino lehen, gogoratu Gamma banaketaren dentsitate-funtzioa honela idatz daitekeela (R bidez `dgamma`):

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} I_{\{x \geq 0\}}(x)$$

Gainera, banaketa honen momentuak ere ezagunak dira:

$$E(X) = k\theta; \quad VAR(X) = k\theta^2$$

1. Momentuen metodoa:
   - Lortu momentuen metodoen bidezko estimazioa dagokion ekuazio sistema ebatziz. Horretarako, erabili `nleqslv` funtzioa, zenbakizko metodoen bitartez soluzioa hurbilduko duena.
   - Kalkulatu eskuz momentuen metodoaren bidezko estimatzaileak eta konparatu aurreko atalean lortutako estimazioa eskuz lortutako balioekin. Hurbilketa ona lortu al dugu?
2. Egiantz handieneko estimazioa:
   - Irakurri lagina eta eraiki dagokion egiantz-funtzioaren logaritmoa (L) kalkulatzen duen funtzioa.
   - Marraztu L k balioekiko $[3, 10]$ eremuan. θren zein balio finkatu duzu?
   - Marraztu L θ balioekiko $[3, 10]$ eremuan. $k$ren zein balio finkatu duzu?
   - Marraztu L-ren maila-kurbak $[1, 10] \times [1, 10]$ eremuan.
   - Erabili *nlm* funtzioa egiantz handieneko estimazioak topatzeko. Erabili hasiera puntu bezala $k = 1, \theta = 3$. (Gogoratu *nlm* minimoa bilatzen saiatuko dela).
3. Propietateak: Sortu Gamma(5, 8) banaketatik $n \in \{10, 20, 30, 40, 50, 75, 100, 150, 200, 250, 300, 350\}$ tamainuko laginak $N = 500$ aldiz.
   - Lortu EHEk ematen dizkizun estimazioak lagin bakoitzerako eta aztertu batezbesteko eta bariantzak n-ren arabera. Estimatzailea tinkoa al da? (Hasi $\hat{k} = \hat{\theta} = 3$ balioetatik eta zehaztu *stepmax*=2 `nlm` funtzioan)
   - Egin ariketa berdina momentuen metodoak ematen dizkizun estimazioak erabiliz. Estimatzailea tinkoa al da?