P6 - Estadística Bayesiana I

Laborategi honetan, telebistako lehiaketa bat prestatzeko helburuarekin gauzatu den analisi batean lortutako datuak aztertuko ditugu. Lehiaketa hau irabazteko, lehiakideek 25 galdera erantzun behar dituzte, gehienez 2 akats eginez. Hirugarren akats bat egin ezkero, jokaldia amaitu egiten da eta eta lehiakideak galdu egiten du.

Saria kalibratu nahi dugu lehiaketa irabazteko probabilitatearen arabera, eta, horretarako, irabazteko probabilitatea zein den jakin behar dugu. Horretarako, 20 boluntariorekin proba bat egin da eta galderak erantzun dituzte 3 akats egin arte (edo 25 galderara iritsi arte). Bakoitzak 3. akatsa baino lehen **zuzen** erantzundako galdera kopurua honakoa da:

```
emaitza.zuzenak <- c(13, 2, 7, 9, 18, 16, 5, 4, 2, 11, 8, 11, 10, 6, 16, 10, 23, 1, 1, 7)
```

Irabazteko probabilitatea zein den kalkulatzeko, estatistika Bayesiarrean oinarritutako analisi bat burutuko dugu. Nahiz eta interesatzen zaiguna irabazteko probabilitatea kalkulatzea den, problema osoaren analisi zabalago bat egingo dugu.

Eredua

Modu naturalean, *3. akatsa baino lehen zuzen erantzundako galdera kopurua*, Binomial Negatibo banaketa batekin adierazi dezakegu. Arrakasta probabilitatea p izanik, banaketa honek, k arrakasta lortu arte x porrot gertatzeko probabilitatea neurtzen du. Adibidez, aurpegia ateratzeko 0.43ko probabilitatea duen txanpon bat izanda, 5 aurpegi lortu aurretik 10 gurutze lortzeko probabilitatea $p = 0.43$ eta $k = 5$ parametroko Binomial Negatibo bat erabiliz kalkulatu dezakegu: $BinNeg(k = 5, p = 0.43)$. Zehazki, $k = 5$ aurpegi lortu aurretik 10 gurutze lortzeko probabilitatea hauxe da:

```
dnbinom(x=10, size=5, prob=0.43)
```

```
## [1] 0.05327518
```

Binomial Negatibo baten probabilitate masa funtzioa honakoa da:

$$P(X = x) = \binom{k + x - 1}{x} p^k (1 - p)^x$$

Gure probleman, 3. akatsa egitean esperimentua moztu egiten da. Beraz, arrakasta galdera gaizki erantzutea da, eta $k = 3$. Bestalde, estimatu nahi dugun parametro ezezaguna, p, galdera bat gaizki erantzuteko probabilitatea izango da. Ordea, benetan interesatuko zaiguna $1 - p$ da, galdera bat ondo erantzuteko probabilitatea.

Ariketa: Suposatuz galderak **zuzen** erantzuteko probabilitatea 0.87 dela, zein da, eredu honen arabera, lehiaketa **ez irabazteko** probabilitatea?

```
pnbinom(q=22, size=3, prob=0.13)
```

```
## [1] 0.648294
```

Problema hau maiztasunetan oinarritutako estatistikaren bidez ebatzi genezake, zuzenean p datuetatik estimatuz, baina kasu honetan metodo Bayesiarrak erabiliko ditugu. Horretarako, bi oinarrizko elementu definitu behar ditugu: egiantza funtzioa eta p parametroaren a priori banaketa.

- **Egiantza funtzioak**, p parametro bat finkatzean, laginaren probabilitatea ematen digu: $P(D \mid k=3, p)$. Suposatuz lagineko balioak askeak direla, eta guztiek $BinNeg(k=3, p)$ banaketa Binomial Negatibo bat jarraitzen dutela, orduan egiantza funtzioa lagineko elementu guztien Binomial Negatiboaren probabilitate-masa funtzioaren balioak biderkatuz lortuko dugu:

$$P(D \mid k=3,p) = P(x_1,\dots,x_n|k=3,p) = \prod_{i=1}^{n} P(X=x_i|k=3,p) = \prod_{i=1}^{n} \binom{3+x_i-1}{x_i} p^3(1-p)^{x_i}$$

- p **parametroaren a priori banaketa:** $f(p)$. Egiantza funtzio honekin, a prioriarentzat aukera tipikoa Beta banaketa bat izango da. Era honetan, parametroaren a posteriori banaketa ere Beta bat izango da (banaketa konjugatuak). Hau da, p parametroaren a priori banaketa $Beta(\alpha, \beta)$ izanik, a posteriori banaketa ere Beta banaketa bat izango da, hurrengo parametroekin: $\alpha' = \alpha + 3n$ eta $\beta' = \beta + \sum_{i=1}^{n} x_i$.

Ereduaren aurretiazko analisia

Datuen analisiarekin hasi baino lehen, ikus dezagun nola aldatzen den p parametroaren a posteriori banaketa behatutako datuen kopurua handitzen denean. Horretarako, har dezagun datuak $BinNeg(k=3, p=0.1)$ banaketa jarraitzen dutela (hau da, asumitu dezagun $p=0.1$ dela benetako balioa). Sortu ditzagun bi lagin hasteko: bat txikia (5 tamainakoa) eta bat handiagoa (50 tamainakoa).

```
k <- 3
p <- 0.1  #Asumitzen ari gara hau dela benetako balioa, laginak sortzeko
lagina.5  <- rnbinom(5, size=k, prob=p)
lagina.50 <- rnbinom(50, size=k, prob=p)

lagina.5 #Lagin txikia bistaratuko dugu
```

```
## [1] 46 17  9 44 13
```

Bestalde, p parametroaren a priori banaketa moduan $\alpha = \beta = 1$ parametroko Beta banaketa bat hartuko dugu. Azken hau p parametroarentzat banaketa uniforme baten baliokidea da (hau da, $p$ren balio guztiek *probabilitate* bera dute). Lehen ikusitako formula erabiliz, p parametroaren a posteriori banaketaren parametroak zehaztuko ditugu kasu bakoitzean:

```
#A priori parametroak
a.prior <- 1  #alpha (a priori)
b.prior <- 1  #beta  (a priori)

#Lagin txikirako (n=5)  a posteriori parametroak
a.lagina.5 <- a.prior + k*length(lagina.5) #alpha (a posteriori)
b.lagina.5 <- b.prior + sum(lagina.5)      #beta  (a posteriori)

#Lagin handirako (n=50) a posteriori parametroak
a.lagina.50 <- a.prior + k*length(lagina.50) #alpha (a posteriori)
b.lagina.50 <- b.prior + sum(lagina.50)      #beta  (a posteriori)
```

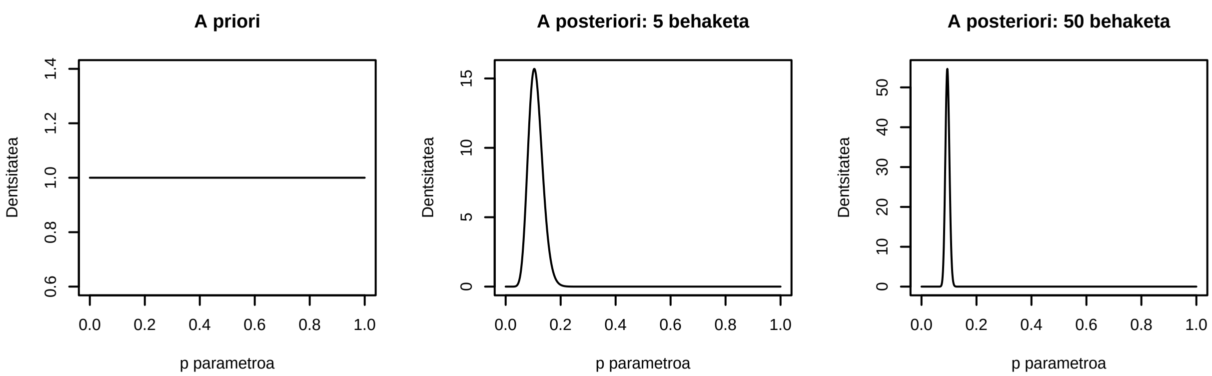
Azkenik, hiru banaketak irudikatuko ditugu, a priori banaketa (datuak behatu aurretik), eta a posteriori banaketak 5 eta 50 behaketako laginak ikusi eta gero:

```r
p.seq <- seq(0, 1, 0.001)
dens.prior   <- dbeta(x=p.seq, shape1=a.prior,     shape2=b.prior)     #a priori banaketa
dens.post.5  <- dbeta(x=p.seq, shape1=a.lagina.5,  shape2=b.lagina.5)  #a posteriori (n=5)
dens.post.50 <- dbeta(x=p.seq, shape1=a.lagina.50, shape2=b.lagina.50) #a posteriori (n=50)

#Dentsitate-funtzioen bistaraketa:
layout(matrix(1:3, nrow=1))
par(cex=0.5)
plot(p.seq, dens.prior, type="l",
     main="A priori", xlab="p parametroa", ylab="Dentsitatea")

plot(p.seq, dens.post.5, type="l",
     main="A posteriori: 5 behaketa", xlab="p parametroa", ylab="Dentsitatea")

plot(p.seq, dens.post.50, type="l",
     main="A posteriori: 50 behaketa", xlab="p parametroa", ylab="Dentsitatea")
```



Grafikoan ikus dezakegu nola, banaketa uniforme batetik abiatuta ere, soilik 5 behaketekin p parametroaren ia probabilitate masa osoa 0 eta 0.2 balioen artean metatzen den. Lagin tamaina 50ra handitzen dugunean, probabilitate masa *"benetako"* balioaren inguruan (hau da, 0.1 inguruan) are gehiago metatzen da.

Ariketa: Errepikatu atal honetan ikusitakoa a priori banaketa ezberdinak hartuz (Beta banaketaren parametroak aldatuz). Probatu lagin tamaina ezberdinekin eta aztertu zein lagin tamaina den beharrezkoa kasu bakoitzean estimazio on bat lortzeko.

A posteriori banaketa izanda, estimazio puntual bat lor dezakegu p parametroarentzat banaketa horren itxaropena kalkulatuz. Gure kasuan, a posteriori banaketa $Beta(\alpha', \beta')$ banaketa bat izanik, badakigu bere itxaropena $\alpha'/(\alpha' + \beta')$ izango dela.

```r
#Lagin txikia erabilita (n=5)
estimazioa.p.5 <- a.lagina.5 / (a.lagina.5 + b.lagina.5)
estimazioa.p.5
```

```
## [1] 0.109589
```

```r
#Lagin handia erabilita (n=50)
estimazioa.p.50 <- a.lagina.50 / (a.lagina.50 + b.lagina.50)
estimazioa.p.50
```

```
## [1] 0.09443402
```

Puntu-estimazioaz gain, tarte bidezko estimazio bat ere lor dezakegu p parametroarentzat, a posteriori banaketaren kuantil egokiak aukeratuz. Adibidez, %95ko tarte bat lortzeko:

```
#Lagin txikia erabilita (n=5)
p.inf.5 <- qbeta(0.025, a.lagina.5, b.lagina.5)
p.sup.5 <- qbeta(0.975, a.lagina.5, b.lagina.5)
message("%95ko tarte Bayesiarra: [", round(p.inf.5, 3), ",", round(p.sup.5,3), "]")
```

```
## %95ko tarte Bayesiarra: [0.064,0.165]
```

```
#Lagin handia erabilita (n=50)
p.inf.50 <- qbeta(0.025, a.lagina.50, b.lagina.50)
p.sup.50 <- qbeta(0.975, a.lagina.50, b.lagina.50)
message("%95ko tarte Bayesiarra: [",  round(p.inf.50, 3), ",", round(p.sup.50,3), "]")
```

```
## %95ko tarte Bayesiarra: [0.081,0.109]
```

Ariketa: Beste estimatzaile puntual posible bat MAP (maximum a posteriori) da, hau da, a posteriori banaketaren moda. Kalkulatu estimatzaile puntual hau aurreko kasuetan eta konparatu emaitzak itxaropenarekin lortutako balioekin.

$$X \sim B(\alpha, \beta) \implies moda_X = \frac{\alpha - 1}{\alpha + \beta - 2}$$

```
minusf.lagina.5 <- function(x) {-dbeta(x, a.lagina.5, b.lagina.5)}
minusf.lagina.50 <- function(x) {-dbeta(x, a.lagina.50, b.lagina.50)}

# Zenbakizko metodoekin
# Lagin Txikia:
nlm(minusf.lagina.5, 0.1)$estimate
```

```
## [1] 0.1041667
```

```
# Lagin Handia
nlm(minusf.lagina.50, 0.1)$estimate
```

```
## [1] 0.09392614
```

```
# Matematikoki
# Lagin Txikia
(a.lagina.5 - 1)/(a.lagina.5 + b.lagina.5 - 2)
```

```
## [1] 0.1041667
```

```r
# Lagin Handia
(a.lagina.50 - 1)/(a.lagina.50 + b.lagina.50 - 2)
```

```
## [1] 0.09392611
```

Datuen analisia

Aurreko tutoriala lantzeko, p parametroaren balio bat finkatu dugu (guk asmatutakoa). Azken atal honetan, lehiaketa irabazteko probabilitatearen galderari erantzungo diogu, hori baitzen gure helburu nagusia hasieratik. Baina galdera hau tribiala ez denez, aurretik. . .

Ariketa: Dokumentuaren hasierako datuak hartuz lagin moduan (erantzun zuzen kopuruak), eta a priori uniforme batetik abiatuz ($Beta(1, 1)$ banaketa batetik), irudikatu p parametroaren a posteriori banaketa, estimatu bere baliorik probableena (MAP) eta sortu %95ko tarte Bayesiarra.

```r
k <- 3
a.prior <- 1
b.prior <- 1

a.lagina <- a.prior + k*length(emaitza.zuzenak) #alpha (a posteriori)
b.lagina <- b.prior + sum(emaitza.zuzenak)      #beta  (a posteriori)

# MAP atera
minusf <- function(x) {-dbeta(x, a.lagina, b.lagina)}

# phat <- nlm(minusf, 0.4)$estimate
phat <- (a.lagina - 1) / (b.lagina + a.lagina - 2)
phat
```

```
## [1] 0.25
```

Lehiaketa irabazteko probabilitatea ezagutzeko, 3 akats egin aurretik erantzundako galdera zuzenen kopuruaren probabilitate-banaketa ezagutu behar dugu. Ikusi dugu probabilitate banaketa hau $k = 3$ parametroko Binomial Negatibo batekin adierazi dezakegula. Arazoa da p ez dugula ezagutzen.

Gainera, kontuan izan behar dugu, Binomial Negatiboak **galdera zuzen** kopurua adierazten duela k akats egin aurretik, baina galdera kopurua, berez, ez dagoela mugatuta banaketa honetan. Lehiaketa errealean soilik 3 modu daude programa irabazteko: 23, 24 edo 25 galdera ondo erantzunez. Baina, Binomial Negatiboa erabiltzen badugu, asumitu behar dugu 25 galdera baino gehiago (hipotetikoak) zuzen erantzuten dituen edonork ere lehiaketa irabaziko duela. Hau honela, onartuko dugu lehiaketa irabazteko probabilitatea $P(X \geq 23)$ dela, $k = 3$ eta p (ezezaguna) parametroko Binomial Negatibo bat izanik.

Ariketa: Kalkulatu lehiaketa irabazteko probabilitatea suposatuz p parametroak aurreko ariketan lortutako estimazio puntualaren (MAP) balioa hartzen duela.

```r
prob.irabazi <- 1 - pnbinom(q=22, size=k, p=phat)
prob.irabazi
```

```
## [1] 0.03210852
```

Nahiz eta ariketan egindako puntu-estimazioa nahiko zentzuzkoa izan, estatistika Bayesiarrean asumitzen dugu banaketaren parametroek **edozein balio** har dezaketela (probabilitate jakin batekin), hau da, zorizko aldagai moduan tratatzen ditugu. Beraz, lehiaketa irabazteko probabilitatearen kalkuluak p parametroaren balio posible guztiak hartu beharko genituzke kontuan. Hau *a posteriori banaketa prediktiboa* deritzon banaketaren bidez egiten da, lagin bat emanik, X aldagaiaren hurrengo behatutako balio baten ($\tilde{x}$) probabilitatea nolakoa den esango diguna. Hau da, $P(\tilde{x}|D)$.

Hau lortzeko, parametroen balio posible guztiak hartu behar ditugu kontutan. Hau da, gure kasuan, $D = \{x_1, \ldots, x_n\}$ laginak baldintzatutako $\tilde{x}$ balioaren eta p parametroen baterako probabilitate banaketa lortuko dugu. Ondoren, banaketa hau $p$ren balio guztiekiko integratuko dugu (hau da, marginalizatuko dugu):

$$P(\tilde{x}|D) = \int_0^1 P(\tilde{x}, p|D)dp$$

Katearen legea aplikatuz, honakoa dugu:

$$P(\tilde{x}|D) = \int_0^1 P(\tilde{x}|p, D)P(p|D)dp$$

Gainera, p parametroaren balioa jakinda, balio berri ($\tilde{x}$) baten probabilitatea aurreko behaketetatik (D) askea da, $P(\tilde{x}|p, D) = P(\tilde{x}|p)$. Beraz:

$$P(\tilde{x}|D) = \int_0^1 P(\tilde{x}|p)P(p|D)dp$$

Integralean agertzen den lehenengo terminoa p balio jakin baterako k erantzun oker baino lehen emandako erantzun zuzenenen banaketa da, hau da, $k = 3$ eta $p = p$ parametroko Binomial Negatibo bat. Bigarren terminoa parametroaren a posteriori banaketa da, hau da, $\alpha' = \alpha + 3n$ eta $\beta' = \beta + \sum_i x_i$ parametroko Beta banaketa bat. Guzti hau ordezkatuz:

$$P(\tilde{x}|D) = \int_0^1 \binom{3 + \tilde{x} - 1}{\tilde{x}} p^3(1-p)^{\tilde{x}} \frac{1}{Beta(\alpha + 3n, \beta + \sum_{i=1}^n x_i)} p^{\alpha+3n}(1-p)^{\beta+\sum_{i=1}^n x_i} dp$$

Eta sinplikatuz:

$$P(\tilde{x}|D) = \int_0^1 \frac{\binom{3+\tilde{x}-1}{\tilde{x}}}{Beta(\alpha + 3n, \beta + \sum_{i=1}^n x_i)} p^{\alpha+3(n+1)}(1-p)^{\tilde{x}+\beta+\sum_{i=1}^n x_i} dp$$

Integral honen emaitza Beta-Binomial Negatibo izeneko banaketa bat da, hiru parametro dituena k, α eta β. Kasu jakin honetan, $k = 3$, $\alpha' = \alpha + 3n$ eta $\beta' = \beta + \sum_i x_i$. Banaketa hau ez dator Rko oinarrizko instalazioan baina `extraDistr` paketean topa dezakezue.

$P(\tilde{x}|D)$ banaketak behaketa berri baten probabilitatea adierazten du, aurretik ikusitako lagina kontuan izanda. Hau da, 21. jokalariaren erantzun zuzen kopuruaren banaketa (3 akats baino lehenago), aurreko 20 jokalarien datuak kontuan izanda. Honi esker, jokalari berri batek lehiaketa irabazteko probabilitatea kalkulatu dezakegu, hurrengo probabilitatea kalkulatuz: $P(\tilde{x} \geq 23|D) = 1 - P(\tilde{x} \leq 22|D)$. Kalkula dezagun balio hori R erabiliz:

```
a.prior <- 1
b.prior <- 1

k <- 3
```

```r
a <- a.prior + k*length(emaitza.zuzenak)
b <- b.prior + sum(emaitza.zuzenak)

library(extraDistr)
prob.irabazi <- 1 - pbnbinom(q=22, size=k, alpha=a, beta=b)
prob.irabazi
```

```
## [1] 0.03680945
```

Hau da, lehiaketa irabazteko probabilitatea hauxe da: 0.037.

Ariketa: Konparatu lehiaketa irabazteko probabilitatea hurrengo bi kasuetan: (1) $p$ren balio posible guztiak kontuan hartuz, eta (2) aurretik kalkulatutako estimazio puntualak erabiliz.

Ariketak (pentsatzeko)

- Denboraldiko aurrekontua (sarientzat) 250.000€ da eta lehiaketa astean behin burutuko da. Programazio bereziko asteak kenduz (gabonak, udara, etab.), denboraldiak 40 aste izango ditu. Programa bakoitzean 5 lehiakide egongo dira, batezbeste, eta 25 galderak 3 akats baino gutxiagorekin erantzuten dituztenak soilik irabaziko dute.
  - Nola zehaztuko zenuke irabazleen saria?
  - Aurreko atalean kalkulatutako kantitate hori kontuan izanda, zein da denboraldi amaieran aurrekontua baino diru gehiago gastatzeko probabilitatea?
  - Zein probabilitate dago denboraldi amaieran 50.000 € baino gehiagoko soberakina izateko?
- Suposatu 15. galderara iristen diren jokalariei erdibideko sari bat eman nahi diegula (bukaerako sariaz gain). Errepikatu kalkulu guztiak kasu honetarako.