Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Saucedo

Seminari 5: Estimadors de màxima versemblança

La versemblança d'una mostra.

La densitat conjunta avaluada a la mostra s'anomena **funció de versemblança/likelyhood de la mostra**. És una variable aleatòria amb valors a $\mathbb{R}^+$.

Denotarem per

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

on $f(x;\theta)$ és la densitat de les X_i (o la funció de probabilitats en el cas discret), a la funció de versemblança, i per

$$\lambda(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

al logaritme de L.

El mètode de màxima versemblança consisteix en utilitzar el valor de $\hat{\theta}$ que maximitza la versemblança de la mostra com a estimador del paràmetre θ.

L'**estimador de màxima versemblança**[1] $\hat{\theta}$ és aquell que verifica

$$L(X, \hat{\theta}) \geq L(X, \theta)$$

per a tot θ a l'espai de paràmetres.

La funció de densitat $f(x;\theta)$ (o en el cas discret, la probabilitat $p(x;\theta) = \mathbf{P}\{X = x\}$) ens dóna una mesura intuïtiva de quant podem confiar que succeeixi x una vegada coneixem θ, però també ens permet comparar aquelles mesures per diferents valors del paràmetre θ. Particularment, una vegada hem observat x, $f(x;\theta)$ ens indica fins a quin punt era esperable el resultat observat quan el paràmetre val θ.

Trobarem aleshores que el model probabilístic donat per $f(x;\theta')$ resulta més creïble (o més versemblant) que el model definit per $f(x;\theta'')$ quan es compleix que $f(x;\theta') > f(x;\theta'')$.

En aquest context podem dir que el mètode de la màxima versemblança proposa utilitzar com estimador del paràmetre el valor que fa més creïble o més versemblant el resultat que de fet ha estat observat experimentalment.

Calculeu el MLE per distribucions normals i exponencials.

1. Càlcul dels estimador de màxima versemblança amb R

La llibreria `MASS` té una funció que es diu `fitdistr` que calcula el MLE para algunes distribucions:

```
library(MASS)
?fitdistr

###############
```

[1] que pot no ser únic, pot no existir

```
# Exponencials
###############
x=rexp(100,rate=2) #rate es lambda
fitdistr(x,"exponential")
1/mean(x) # comprovem que l'estimador es el que coneixem

##########
# Normals
##########
n=100
y=rnorm(n,3,6) #mean=3, desv=6
fitdistr(y,"normal")
# el MLE és
mean(y);sqrt(var(y))
# R fa una correció a la variància anomenada correció de Bessel
mean(y);sqrt(var(y)*(n-1)/n)
```

- Hi ha una llibreria (`fitdistrplus`) que inclou més models (consulteu l'ajuda).
- El càlcul de la màxima versemblança també es pot fer amb les funcions `nlm` i `optim` (rutines d'optimització de `R`). Aquestes rutines minimitzen funcions, per tant s'ha d'escriure una funció que calculi $-\log L(X_1, \ldots, X_n; \theta)$, on $\theta \in \mathbb{R}^p$.

1.1. Exemple: el MLE pel paràmetre λ d'una distribució exponencial

Comencem generant un mostra de mida $n = 100$ i paràmetre $\lambda = 2$.

```
x=rexp(100,rate=2)
mean(x);1/mean(x)
```

Ara definim una funció que rebi un vector x, la mostra, i un paràmetre λ que és el que voldrem trobar. La funció retorna la $-\log L(X; \lambda)$

```
mlogL=function(lam,x){return(-(length(x)*log(lam)-lam*sum(x)))}
```

Podem dibuixar gràficament com es comporta la funció $\log L(X; \lambda)$ en funció del valor de λ:

```
npoint <- 100
lambda <- seq(0.1, 10, length = npoint)
plot(lambda, mlogL(x,lambda), type = "l")
```

Finalment, utilitzem les funcions mencionades per trobar el valor de $\hat{\lambda}$ que minimitza la funció definida:

```
nlm(mlogL, 1, x)
optim(1,mlogL,x=x)
1/mean(x)
```

Exercicis

1. Escriviu una funció que calcula la -log-versemblança per a una mostra de variables aleatòries amb distribució normal de mitjana i variància desconegudes. Utilitzeu-la per estimar els paràmetres d'una mostra de mida 100 generada amb $\mu = 3$, $\sigma = 2$ amb `nlm`.

2. Calculeu els estimadors de màxima versemblança de μ i σ^2 per a les dades de Michelson de la llibreria `MASS`. Repetiu el càlcul eliminant els possibles outliers. Si teniu temps, compareu els resultats obtinguts amb el que trobaríeu utilitzant tècniques gràfiques d'ajust.

3. La següent taula mostra la distribució del nombre de parells de sabates per a un grup de 60 corredors aficionats.

Nombre de parells:	1	2	3	4	5
Freqüència:	18	18	12	7	5

   La distribució de Poisson no pot ser un bon model per a aquestes dades perquè se suposa que els corredors tenen almenys un parell de sabates. La distribució de Poisson truncada al zero amb funció de massa de probabilitats

$$p(x;\theta) = \frac{\theta^x e^{-\theta}}{x!(1-e^{-\theta})}, \quad x = 1, 2, \ldots, \quad \theta > 0$$

   pot ser un model més adient.

   a) Escriviu la funció de versemblança.

   b) Trobeu numèricament l'estimador de màxima versemblança de θ.

4. Com ja vau veure, a la llibreria `evir` hi ha un conjunt de dades anomenat `danish` que es correspon amb les quantitats reclamades per incendis fetes a les asseguradores a Dinamarca entre el 3 de gener de 1980 i el 31 de desembre de 1990. Suposem que aquestes dades provenen d'una distribució de Pareto amb densitat

$$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \quad \text{si } x \geq x_m$$

   i 0 si $x < x_m$.

   a) Calculeu el MLE per als paràmetres α i x_m de la distribució de Pareto. Observeu que x_m està a la frontera de l'espai de paràmetres.

5. Les densitats Gamma pertanyen a una família de distribucions contínues amb dos paràmetres. L'exponencial i la χ^2 són casos particulars de densitats Gamma. Per una variable aleatòria $X \sim \Gamma(\alpha, \beta)$ la seva densitat

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}$$

   on $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}dx$.

   En aquest cas, els estimadors de màxima versemblança no admeten una forma tancada, així que cal calcular-los numèricament.

   a) Genereu una mostra de mida 100 d'una $\Gamma(10, 3)$, utilitzeu la funció `rgamma()`

   b) Calculeu els estimadors de màxima versemblança.

Examen 2022

Zero-inflated Poisson. Una variable aleatòria X que segueix una distribució habitual de Poisson amb paràmetre λ s'utilitza, en moltes ocasions, per modelar esdeveniments que succeeixen naturalment i on X representa el nombre d'esdeveniments per unitat de temps o espai. Tanmateix, la distribució de Poisson pot no ser útil quan X pren el valor 0 amb una alta probabilitat. En aquest

cas, és més convenient utilitzar una modificació de $Pois(\lambda)$ coneguda com *zero inflated Poisson* (ZIP). La distribució ZIP té paràmetres π i λ, on $X \sim ZIP(\pi, \lambda)$, té funció de distribució

$$P(X = k) = \begin{cases} \pi + (1-\pi)\exp(-\lambda) & \text{if } k = 0 \\ (1-\pi)\exp(-\lambda)\lambda^k/k! & \text{if } k \in \{1, 2, \dots\} \end{cases}$$

on $0 \leq \pi \leq 1$ i $\lambda \geq 0$.

1. Comenceu demostrant que

$$E(X) = (1-\pi)\lambda, \quad Var(X) = \lambda(1-\pi)(1+\lambda\pi).$$

   A partir d'aquí, trobeu l'expressió pels estimadors de π i λ que dóna el mètode dels moments.

2. Donat un conjunt de n observacions i.i.d. $X = (X_1, \dots, X_n)$ que segueixen una distribució $ZIP(\pi, \lambda)$, escriviu la seva log-versemblança $l(X_1, \dots, X_n; \pi, \lambda)$.

La taula següent conté el nombre d'huracans grans a l'Atlàntic (categories 4 o 5) per any que han arribat a Estats Units des de 1987 fins 2012.

Dècada	Nombre d'esdeveniments per any									
1980-1989	-	-	-	-	-	-	-	0	0	1
1990-1999	0	0	1	0	0	1	0	0	2	2
2000-2009	0	0	1	1	3	4	0	0	2	0
2010-2019	0	0	0	-	-	-	-	-	-	-

1. A partir de les dades de la taula anterior, calcula els estimadors $\hat{\lambda}_M$ i $\hat{\theta}_M$ que dóna el mètode dels moments.

2. Troba numèricament els estimadors de màxima versemblança per $\hat{\lambda}_L$ i $\hat{\pi}_L$. Compara la distribució de probabilitat empírica amb les calculades numèricament $ZIP(\hat{\pi}_M, \hat{\lambda}_M)$ i $ZIP(\hat{\pi}_L, \hat{\lambda}_L)$