Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Barcelona

Seminari 5: Tests de Hipótesi 2

Aquesta pràctica té com a objectiu implementar en R els tests de quocient de versemblances (LRT) que hem estat estudiant, i els altres tests asimptòtics (Wald i Score).

1. Tests asimptòtics (LRT, Wald Test)

LRT és més o menys universal, i es pot aplicar en la majoria dels casos, almenys quan el nombre de paràmetres és finit: Per provar la hipòtesi nul·la $H_0 : \theta \in \Theta_0$ contra l'alternativa $H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$, el LRT rebutja H_0 per a valors petits de

$$\frac{\sup_{\theta \in \Theta_0} L(x;\theta)}{\sup_{\theta \in \Theta} L(x;\theta)}.$$

Hi ha altres tests que també es basen en la normalitat asimptòtica del EMV: ja sabem que sota suficients condicions de regularitat, $\hat{\theta}$ convergeix en distribució a una V. A. amb distribució normal, de mitjana θ i variància $I(\theta)^{-1}$. A més, com $\hat{\theta}$ és consistent com a estimador de θ, llavors $I(\hat{\theta})$ és un estimador consistent de $I(\theta)$.

Observem que si

$$\Lambda(\mathbf{x}) = -2\log\lambda(\mathbf{x}) = 2\left[l(\widehat{\theta};\mathbf{x}) - l(\theta_0;\mathbf{x})\right]$$

la regió crítica del LRT es pot escriure com

$$C_1 = \{\mathbf{x} : \Lambda(\mathbf{x}) \geq c\}$$

Sota les mateixes condicions de regularitat que necessitem per garantitzar que el EMV és asimptòticament normal, tenim

$$\Lambda(\mathbf{X}) \stackrel{D}{\to} \chi^2_p \qquad \text{si} \quad p = dim(\Theta) - dim(\Theta_0) \geq 1$$

Wald va proposar utilitzar com a estadístic per provar $H_0 : \theta \in \Theta_0$:

$$W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) \sim \chi^2_1 \quad \text{asimptóticament, sota}\, H_0,$$

o en la seva versió vectorial

$$W = (\hat{\theta} - \theta_0)^{tr} I(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi^2_p \quad \text{asintóticament, sota}\, H_0, \text{ on}\, p = dim(\Theta).$$

Observem que W és el quadrat de la distància entre θ_0 i el EMV de θ, ponderada per una estimació consistent de la informació continguda en la mostra.

- el LRT requereix el càlcul del EMV i el EMV restringit a Θ_0.
- el test de Wald només requereix el càlcul del EMV.

El LRT per tant usa més informació i hi ha estudis empírics que suggereixen que per a mostres de grandària moderada resulta més confiable, encara que asintóticamente són equivalents.

1. Sigui $X_1, X_2, \ldots, X_n$ una mostra aleatòria d'una població amb distribució de Poisson truncada, amb funció de massa de probabilitats

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

   a) Escriure la versemblança, el Score i la informació observada.

   S'han anotat les grandàries dels grups observats en llocs públics en una tarda de primavera.

Grandaria del grup:	1	2	3	4	5	6
Freqüencia:	1486	694	195	37	10	1

   Si suposem que un model raonable per a aquestes dades és la distribució de Poisson truncada,

   b) Mostrar que el EMV correspon al màxim en θ de

$$3663 \log\theta - 2423\theta - 2423\log(1-e^{-\theta}).$$

   c) Trobar (numèricament) l'estimador de màxima versemblança de θ.

   d) Obtenir un interval de confiança aproximat de 95 % per θ.

2. *Datos de Pielou sobre la enfermedad en las raíces de los abetos (Douglas fir trees) causada por la Armillaria*

   L'ecologista I.C. Pielou, va estudiar el patró d'arbres sans i malalts (infectats amb *Armillaria*) en una plantació d'avets. Va registrar les longituds de 109 successions d'arbres malalts.

Longitud de las sucesiones de árboles enfermos

Longitud	1	2	3	4	5	6
Número de sucesiones	71	28	5	2	2	1

   Basant-se en consideracions d'índole biològic, Pielou va proposar un model geomètric per a la distribució de les longituts.

   Nota: Si $X \sim$Geo(p), $\mathbf{P}(X = h) = (1-p)^{h-1}p$, para $h = 1, 2, \ldots$, es decir, X cuenta el número de ensayos hasta el primer éxito en una sucesión de ensayos independientes de Bernoulli con probabilidad de éxito p.

   a) Un grup de defensors dels boscos afirmen que $4/5$ dels arbres están malalts. Fer un test de raó de versemblançes per confirmar o rebutjar aquesta afirmació.

   b) Com es construeix un interval de confiança (aproximat) per a p?

   c) Escriu la regió crítica del test de Wald per provar la hipòtesi nul·la anterior. Quina és la conclusió del test?

3. Usar los datos del número de goles de las ligas de fútbol europeas correspondientes a las temporadas 1993-1994 hasta 2003-2004

```
goles=read.csv2("http://mat.uab.cat/~acabana/data/goles.csv")
```

   a) Probar la hipótesis nula de que la media del número de goles en la liga española es 3.

   b) Hacer un test de cociente de veromismilitudes para la hipótesis nula de que la media del número de goles/partido en cada una de las ligas es el mismo. Suponer que los datos de cada liga tienen distribución de Poisson con parámetros respectivos $\lambda_i, i = 1, \ldots, 5$.

2. Del examen de pràctiques 2016

4. Les següents dades corresponen a la mitjana mensual de la velocitat del vent (km/h) a Castelldefells entre els anys 2006 i 2012 (`http://www.castelldefels.org/es/doc_generica.asp?dogid=2015`).

```
wind=c(5.86,6.64,8.8,7.81,7.78,7.63,7.51,6.95,5.13,5.2,4.79,5.3)
```

Els meteoròlegs pensen que un bon model per a les velocitats mitjanes del vent és la distribució de Weibull, amb funció de distribució

$$F(t) = 1 - \mathrm{e}^{-(t/\beta)^\alpha} \qquad x \geq 0$$

amb paràmetre d'escala $\beta > 0$ i paràmetre de forma $\alpha > 0$ desconeguts.

a) Calcular (numèricament) els estimadors de màxima versemblança amb `nlm` escrivint una funció que tingui $-\log L$ on L és la versemblança de la mostra.

HINT: Si es vol fer més fàcilment, convé usar que, si Y ∼Weibull(α, β), llavors $X = -\log(Y)$ ∼Gumbel$(-\log(\beta), \alpha^{-1})$. Una variable aleatoria $X \sim$ Gumbel(ξ, θ) té funció de distribució

$$F_X(x) = \mathrm{e}^{-\mathrm{e}^{-(x-\xi)/\theta}} \quad \text{i densitat} \quad f_X(x) = \frac{1}{\theta}\mathrm{e}^{-(z+\mathrm{e}^{-z})} \quad \text{amb} \quad z = \frac{x-\xi}{\theta}$$

$$\mathbf{E}X = \xi + \theta\gamma \quad \mathbf{Var}X = \frac{\pi^2}{6}\theta^2 \quad \text{ón} \quad \gamma \approx 0{,}5772 \quad \text{és la constant d' Euler.}$$

`https://en.wikipedia.org/wiki/Gumbel_distribution`

b) Fes un *qq*-plot per verificar gràficament si el model Weibull és raonable.

c) Quins són els estimadors de α i β?

d) Verifica els resultats obtinguts fent servir

```
install.packages("fitdistrplus");library(fitdistrplus)
fitdist(wind, "weibull")
```