Resolucions de problemes

Estadística

Mario Vilar

4 de juny de 2024

Universitat de Barcelona

Índex

I

Llista I

Exercici I.I. *El temps d'emissió de la primera partícula d'una font radioactiva segueix una llei exponencial de paràmetre $\lambda > 0$. A la pràctica només es pot observar el temps d'emissió quan aquest està en $[0, T]$, on T és conegut. Calculeu la probabilitat de què la partícula sigui emesa en un temps menor o igual a t, sabent que ha estat emesa en $[0, T]$. Siguin $t_1, \ldots, t_n$, n observacions del fenomen descrit anteriorment.*

1. *Descriviu el model estadístic associat a aquestes observacions.*
2. *Doneu la funció de versemblança.*
3. *Mostreu que el model és regular.*
4. *Trobeu la informació de Fisher del model.*

Demostració. Si X és una variable aleatòria amb llei exponencial de paràmetre λ, aleshores $f(x) = \lambda e^{-\lambda x}\mathbb{1}_{(0,+\infty)}(x)$. Per tant, la funció de distribució n'és:

$$F(t) = P(X \leq t) = \begin{cases} 0, & \text{si } t < 0; \\ \lambda \int_0^x e^{-\lambda x} = 1 - e^{-\lambda x}, & \text{si } t \geq 0.^{1} \end{cases}$$

Si sabem que ha estat emesa en $[0, T]$, és a dir, condicionem, tenim:

$$P(X \leq t | 0 \leq X \leq T) = \frac{P(X \leq t, 0 \leq X \leq T)}{P(0 \leq X \leq T)} = \frac{P(X \leq t)}{P(X \leq T)} = \begin{cases} 0, & \text{si } t < 0; \\ \frac{\int_0^t \lambda e^{-\lambda x}\, dx}{\int_0^T \lambda e^{-\lambda x}\, dx} = \frac{1-e^{-\lambda t}}{1-e^{-\lambda T}}, & \text{si } 0 \leq t \leq T; \\ 1, & \text{si } t > T; \end{cases}$$

Si posem $G(t) = P(X \leq t | 0 \leq X \leq T)$, qualsevol variable amb aquesta llei (que és C^1 excepte, *potser*, en 0 i T) tindrà com a funció de densitat:

$$f(t) = \frac{\lambda e^{-\lambda t}}{1 - e^{-\lambda T}} \mathbb{1}_{[0,T]}(t).$$

Ara hem de fer el model estadístic: $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, on $(\mathcal{X}, \mathcal{A})$ és un espai mesurable, $\mathcal{X}$ és un espai mostral i $\mathcal{A}$ és la σ-àlgebra sobre $\mathcal{X}$ formada pels successos aleatoris, i $\mathcal{P}$ és una família de probabilitats en $(\mathcal{X}, \mathcal{A})$. Prendrem $\mathcal{X} = [0, T]^n$, $\mathcal{A} = \mathcal{B}([0, T]^n)$, on per $\mathcal{B}([0, T]^n)$ entenem els borelians de $[0, T]^n$ i $\mathcal{P} = \{P_\lambda \mid \lambda > 0\}$, on P_λ és la probabilitat absolutament contínua en $[0, T]^n$ amb densitat:

$$f_\lambda(t_1, \ldots, t_n) = \prod_{i=1}^n f(t_i, \lambda) = \prod_{i=1}^n \frac{\lambda e^{-\lambda t_i}}{1 - e^{-\lambda T}} = \frac{\lambda^n e^{-\lambda \sum_{i=1}^n t_i}}{(1 - e^{-\lambda T})^n}, \ \forall t_i \in [0, T], \ i = 1, \ldots, n.$$

Tenim, doncs, un model estadístic *paramètric* amb l'espai de paràmetres $\Theta = (0, +\infty)$ i el paràmetre λ. Ara, la funció de versemblança per a VAIID és:

$$L(t_1, \ldots, t_n; \lambda) = \prod_{i=1}^n f(t_i, \lambda) = \frac{\lambda^n e^{-\lambda \sum_{i=1}^n t_i}}{(1 - e^{-\lambda T})^n}$$

[1] La detallarem breument aquí, tot i que no sigui necessari. Fent el canvi de variable $y = -\lambda x$ ($dy = -\lambda\, dx$), es converteix en $\int_{-\lambda x}^0 e^x\, dx$, obtenint el resultat desitjat.

Per a trobar la informació de Fisher cal suposar que el model és regular, i calcular $I_F(\lambda) = E_\lambda(|\partial_\lambda \ln L(x,\lambda)|^2)$, que en el cas absolutament continu es pot fer per $\int_{\mathcal{X}} |\partial_\lambda \ln L(x,\lambda)|^2 L(x,\lambda)\,dx = -E_\lambda\left(\partial^2_\lambda \ln L(x,\lambda)\right)$. Es pot calcular anar calculant poc a poc:

$$\log L(t_1,\dots,t_n;\lambda) = n\log(\lambda) - \lambda\sum_{i=1}^n t_i - n\log(1-e^{-\lambda T}).$$

$$\partial_\lambda \log L(t_1,\dots,t_n;\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n t_i - \frac{nTe^{-\lambda T}}{1-e^{-\lambda T}}.$$

$$\partial^2_\lambda \log L(t_1,\dots,t_n;\lambda) = -\frac{n}{\lambda^2} + \frac{nT^2e^{-\lambda T}}{(1-e^{-\lambda T})^2}.$$

Ens interessa utilitzar la de la segona derivada ja que així ens desfem de les observacions $t_1,\dots,t_n$, i en termes calculístics ens facilita molt la vida. Com que la segona derivada és una constant que no depèn de les observacions $t_1,\dots,t_n$. Si k constant, $E_\lambda(k) = k$, pel que:

$$I_F(\lambda) = -E_\lambda(\partial^2_\lambda \log L(t_1,\dots,t_n;\lambda)) = \frac{n}{\lambda^2} - \frac{nT^2e^{-\lambda T}}{(1-e^{-\lambda T})^2},\ (t_1,\dots,t_n)\in[0,T]^n. \qquad \blacksquare$$

Exercici 1.2. *Considerem una mostra $\{x_1,\dots,x_n\}$ d'una llei de Pareto amb paràmetres (k,c). Es tracta d'una llei absolutament contínua amb funció de distribució de probabilitat:*

$$F_c(x) = \left(1 - \left(\frac{k}{x}\right)^c\right)\mathbb{1}_{[k,+\infty)}(x),\ k,c>0.$$

1. *Descriviu el model estadístic associat a aquestes observacions.*
2. *Doneu la funció de versemblança.*
3. *Mostreu que el model és regular.*
4. *Trobeu la informació de Fisher del model.*

Demostració. Posem $(\mathcal{X},\mathcal{A},\mathcal{P})$, on $(\mathcal{X},\mathcal{A})$ és un espai mesurable, $\mathcal{X}$ és un espai mostral i $\mathcal{A}$ és la σ-àlgebra sobre $\mathcal{X}$ formada pels successos aleatoris, i $\mathcal{P}$ és una família de probabilitats en $(\mathcal{X},\mathcal{A})$. Prendrem $\mathcal{X} = [k+\infty)^n$ (la pista ens la dona com està definida $F_c(x)$), $\mathcal{A} = \mathcal{B}([k+\infty)^n)$, on per $\mathcal{B}([k+\infty)^n)$ entenem els borelians de $[k+\infty)^n$ i $\mathcal{P} = \{P_c \mid c>0\}$, on P_c és la probabilitat absolutament contínua en $[k+\infty)^n$ amb densitat:

$$f_c(x_1,\dots,x_n) = \prod_{i=1}^n f_c(x_i) = \prod_{i=1}^n F_c'(x_i) = \prod_{i=1}^n c\frac{k^c}{{x_i}^{c+1}} = c^n\cdot k^{cn}\left(\prod_{i=1}^n \frac{1}{x_i}\right)^{c+1},\ x_i\in[k,+\infty),\ i=1,\dots,n.$$

Tenim, doncs, un model estadístic *paramètric* amb l'espai de paràmetres $\Theta = (0,+\infty)$ i el paràmetre c. Ara, la funció de versemblança per a VAIID és:

$$L(x_1,\dots,x_n;c) = \prod_{i=1}^n f_c(x_i) = c^n k^{nc}\left(\prod_{i=1}^n \frac{1}{x_i}\right)^{c+1},\ x_i\in[k,+\infty),\ i=1,\dots,n.$$

Finalment, hem de trobar la informació de Fisher, la qual d'alguna manera mesura la quantitat d'informació que conté la mostra sobre el paràmetre (com reacciona la mostra si variem lleugerament el paràmetre). Hem de suposar que el nostre model estadístic paramètric és regular i, aleshores, tenim:

$$I_F(c) = E_c(|\partial_c \log L(x;c)|^2) = -E_c(\partial_c^2 \log L(x;c)), \; x = (x_1, \dots, x_n) \in [k, +\infty), \; i = 1, \dots, n. \quad (1.1)$$

Anem, doncs, a calcular el logaritme i les respectives derivades del logaritme:

$$\begin{aligned} \log L(x;c) &= n \log c + nc \log k - (c+1) \sum_{i=1}^{n} \log x_i. \\ \partial_c \log L(x;c) &= \frac{n}{c} + n \log k - \sum_{i=1}^{n} \log x_i. \\ \partial_c^2 \log L(x;c) &= -\frac{n}{c^2} \end{aligned}$$

Com que la segona derivada és una constant que no depèn de les observacions $t_1, \dots, t_n$ aleshores per a obtenir la informació de Fisher usarem (1.1) i el fet que l'esperança d'una constant és la mateixa constant.

$$I_F(c) = -E_c(\partial_c^2 \log L(x_1, \dots, x_n; c)) = \frac{n}{c^2}.$$ ■

Observació 1.3. Si c hagués sigut la variable coneguda en lloc de la k, tindríem $((0, +\infty)^n, \mathscr{B}((0, +\infty)^n), \{P_k \mid k > 0\})$ l'espai estadístic, de manera que la funció de versemblança L seria:

$$L = \frac{c^n k^{nc}}{(\prod_i x_i)^{c+1}} \prod_{i=1}^{n} \mathbb{1}_{(k, +\infty)}(x_i).$$

Aquesta L no seria sempre positiva, de manera que no tindríem un model regular. Notem també que hauríem de fer servir els indicadors, mentre que si k és la coneguda no ho hem necessitat. En més detall, $(k, +\infty) \subset (0, +\infty)$, però k no és coneguda així que ens cal l'indicador.

Exercici 1.4. *Sigui $(X_1, \dots, X_n)$ una mostra d'una llei de Bernoulli de paràmetre p.*

1. *Per a $n \geq 2$, trobeu un estimador sense biaix de p^2. Demostreu que si $n = 1$, aleshores no existeix cap estimador sense biaix de p^2.*
2. *Trobeu una condició sobre m ($m \geq 1$) per tal que existeixi un estimador sense biaix de p^m.*

Demostració.

1. L'espai estadístic és $(\Omega, \mathcal{A}, \{P_p \mid p \in (0,1)\})$. Definim $E_p(\bar{X}_n) = g(p)$, g l'estimador de p per $\bar{X}_n$:

$$E_p(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^{n} E_p(X_i) = \frac{1}{n} np = p.$$

$$\begin{aligned} E_p(\bar{X}_n^2) &= \frac{1}{n^2} E_p\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) = \frac{1}{n^2} E_p\left(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} E(X_i)E(X_j)\right) = \frac{1}{n^2} \cdot (np + n(n-1)p^2) \\ &= \frac{p}{n} + \frac{n-1}{n} p^2. \end{aligned}$$

D'aquí, podem proposar com a estimador no esbiaixat de p^2 el següent[2]:

$$T = \frac{n\bar{X}_n^2 - \bar{X}_n}{n-1}.$$

Aquest estimador no existeix si $n = 1$. En efecte, suposem que existeix un $T(X_1)$ i arribem a contradicció:

$$p^2 = E_p(T(X_1)) = T(1)p + T(0)(1-p), \ \forall p \in (0,1).$$

però això és impossible perquè un polinomi de grau 2 en $\mathbb{R}[p]$ no pot ser igual a un polinomi de grau 1.

2. Fem el cas $n = 2$, $T(X_1, X_2)$ l'estimador. Aleshores, la seva esperança és:

$$\begin{aligned} E_p(T(X_1, X_2)) &= T(0,0)P(X_1 = 0, X_2 = 0) + T(0,1)P(X_1 = 0, X_2 = 1) \\ &\quad + T(1,0)P(X_1 = 1, X_2 = 0) + T(1,1)P(X_1 = 1, X_2 = 1). \end{aligned}$$

Suposem que $T(X_1, \dots, X_n)$ és un estimador no esbiaixat de p^m:

$$p^m = E_p(T(X_1, \dots, X_n)) = \sum_{(\alpha_1, \dots, \alpha_n) \in [0,1]^n} T(a_1, \dots, a_n) p^{\sum_{i=1}^n a_i}(1-p)^{n - \sum_{i=1}^n a_i} = p^m, \ m \leq n.$$

Ara tenim un polinomi de grau m en $\mathbb{R}[p]$ (amb coeficients de grau m no nul) igual a un polinomi de grau n; per tant, els coeficients han de coincidir i això només podria tenir solució si $n \geq m$. ■

Exercici 1.5. *Siguin $\{x_1, \dots, x_n\}$ observacions de n variables aleatòries independents amb densitat $f_\lambda(x) = \lambda e^{-\lambda x}$, $x > 0, \lambda > 0$.*

1. *Descriviu el model estadístic associat a aquestes observacions.*
2. *Doneu la funció de versemblança.*
3. *És un model exponencial?*
4. *Trobeu la informació de Fisher del model.*
5. *Per $n > 1$, trobeu un estimador sense biaix de λ basat en $S(x) = \sum_{i=1}^n x_i$. I per $n > 2$, calculeu la seva variància.*

Exercici 1.6. *La distribució de Weibull de paràmetres (λ, α), $\lambda > 0$, $\alpha > 0$ té per densitat:*

$$f(x, \lambda, \alpha) = \frac{\alpha}{\lambda} x^{\alpha - 1} \exp\left\{-\frac{x^\alpha}{\lambda}\right\} \mathbb{1}_{(0,+\infty)}(x).$$

Disposem una mostra aleatòria simple $(X_1, \dots, X_n)$ d'aquesta distribució amb α conegut. Calculeu la informació de Fisher de la mostra i trobeu un estimador eficient de λ.

[2] Desconec si té rellevància, però fixem-nos que $nE_p(\bar{X}_n^2) - E_p(\bar{X}_n) = (n-1)p^2$, i si dividim per $n-1$ obtenim, efectivament, p^2.

Demostració. L'espai de distribució és $((0,+\infty)^n, \mathscr{B}((0,+\infty)^n), \{P_\lambda \mid \lambda > 0\})$. La funció de versemblança és:

$$L(x_1, \dots, x_n; \lambda) = \frac{\alpha^n}{\lambda^n} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} \exp\left(-\frac{1}{\lambda} \sum_i x_i^\alpha \right).$$

Volem aplicar $I_F(\lambda) = E((\partial_\lambda \log L(x, \lambda))^2) = -E(\partial_\lambda^2 \log L(x, \lambda))$, pel que seguim calculant:

$$\log L(x_1, \dots, x_n; \lambda) = n \log \alpha - n \log \lambda + (\alpha - 1) \sum_i \log x_i - \frac{1}{\lambda} \sum_i x_i^\alpha.$$

$$\partial_\lambda \log L(x_1, \dots, x_n; \lambda) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i^\alpha$$

$$\partial_\lambda^2 \log L(x_1, \dots, x_n; \lambda) = \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} x_i^\alpha.$$

Per tant:

$$I_F(\lambda) = -E(\partial_\lambda^2 \log L(x, \lambda)) = -\frac{n}{\lambda^2} + \frac{2}{\lambda^3} \sum_{i=1}^{n} E_\lambda(x_i^\alpha) = -\frac{n}{\lambda^2} + 2\frac{n\lambda}{\lambda^3} = \frac{n}{\lambda^2},$$

on hem usat que:

$$E_\lambda(x_i^\alpha) = \int_0^\infty x^\alpha \frac{\alpha}{\lambda} x^{\alpha-1} e^{-\frac{x^\alpha}{\lambda}} \, dx = \lambda.$$

Ara hem de calcular un estimador eficient de λ:

$$\lambda(\theta) \left(\frac{\partial}{\partial \theta} \log L(x, \theta) \right) = T(x) - g(\theta), \ \ g \text{ estimador.}$$

Anem a demostrar que $T(x) = \frac{1}{n} \sum_{i=1}^{n} x_i^\alpha$ és un estimador eficient de λ. Usant que $\partial_\lambda \log L(x_1, \dots, x_n; \lambda) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_i x_i^\alpha$, multipliquem $\partial_\lambda \log L(x_1, \dots, x_n; \lambda)$ per $\frac{\lambda^2}{n}$:

$$\frac{\lambda^2}{n} \left(\partial_\lambda \log L(x, \lambda) \right) = \frac{\sum_i x_i^\alpha}{n} - \lambda.$$

Com que existeix una funció $f(\lambda) = \frac{\lambda^2}{n}$ tal que $f(\lambda) \left(\partial_\lambda \log L(x, \lambda) \right) = T(x) - g(\lambda)$, $T(x) = \frac{1}{n} \sum_{i=1}^{n} X_i^\alpha$ és estimador eficient de λ amb informació de Fisher:

$$I_F(\lambda) = \frac{g'(\lambda)}{f(\lambda)} = \frac{1}{\frac{\lambda^2}{n}} = \frac{n}{\lambda^2}.$$

■

Exercici 1.7. *Sigui X una variable aleatòria discreta que pot prendre valors a, b, c amb probabilitats:*

$$P(X = a) = \frac{1 - \theta}{2}, \ \ P(X = b) = \frac{1}{2}, \ \ P(X = c) = \frac{\theta}{2}, \ \ 0 \leq \theta \leq 1.$$

Considerem una mostra de mida n.

1. *Escriviu la funció de versemblança. Sigui n_1 el nombre de vegades que apareix a en la mostra, n_2 el de b i n_3 el de c. Utilitzeu n_1, n_2, n_3 en l'expressió de la funció de versemblança.*
2. *Calculeu l'estimador del màxim de versemblança de θ.*

3. *Calculeu la cota de Cramer-Rao en l'estimació de θ.*

Demostració.

1. Posem $L(x_1, \dots, x_n; \theta) = P_\theta(X = x_1) \cdots P_\theta(X = x_n) = \left(\frac{1-\theta}{2}\right)^{n_1} \left(\frac{1}{2}\right)^{n_2} \left(\frac{\theta}{2}\right)^{n_3}$.
2. Com que en termes de θ la funció és contínua i derivable, podem maximitzar la funció derivant i igualant a zero. Com que el màxim de $f = g \circ L$, $g(x) = \log x$ és el de L,
$$\log L(x_1, \dots, x_n; \theta) = n_1 \log\left(\frac{1-\theta}{2}\right) + n_2 \log\frac{1}{2} + n_3 \log\frac{\theta}{2}$$
$$\partial_\theta \log L(x, \theta) = -n_1 \frac{1}{1-\theta} + n_3 \frac{1}{\theta} \implies -\frac{n_1}{1-\theta} + \frac{n_3}{\theta} = 0 \iff n_1\theta = n_3(1-\theta),$$
que queda: $\theta = \frac{n_3}{n_1+n_3}$. Com que:
$$\partial^2_\theta \log L(x, \theta) = -n_1 \frac{1}{(1-\theta)^2} - \frac{n_3}{\theta^2} < 0.$$
Aleshores $\theta = \frac{n_3}{n_1+n_3}$ és un màxim i l'EMV queda com $\hat{\theta} = \frac{n_3}{n_1+n_3}$. Si $n_1, n_3 = 0$, qualsevol estimador ens serveix com a EMV, ja que ens quedaria una funció de versemblança constant i tot punt seria extrem.
3. Val a dir que $n_1 \sim B(n, \frac{1-\theta}{2})$ i $n_3 = B(n, \frac{\theta}{2})$. Calculem la cota de Cramer-Rao a partir de la informació de Fisher:
$$\begin{aligned} I_F(\theta) &= E_\theta\left(\frac{n_1}{(1-\theta)^2} + \frac{n_3}{\theta^2}\right) \\ &= \frac{1}{(1-\theta)^2} E_\theta(n_1) + \frac{1}{\theta^2} E_\theta(n_3) = \frac{1}{(1-\theta)^2} n \cdot \left(\frac{1-\theta}{2}\right) + \frac{1}{\theta^2} \cdot n \cdot \frac{\theta}{2} \\ &= \frac{n}{2}\left(\frac{1}{1-\theta} + \frac{1}{\theta}\right) = \frac{n}{2} \frac{1}{\theta(1-\theta)}, \end{aligned}$$
Per tant, la cota de Cramer-Rao queda:
$$\frac{(g'(\theta))^2}{I_F(\theta)} = \frac{1}{\frac{n}{2} \cdot \frac{1}{\theta(1-\theta)}}.$$
■

Exercici 1.8. *Sigui una mostra aleatòria simple, una mostra d'una variable aleatòria amb densitat:*
$$f(x) = 2\theta x \exp(-\theta x^2) \mathbb{1}_{(0,+\infty)}(x),$$
$\theta > 0$ *és un paràmetre desconegut. Trobeu l'estimador del màxim de versemblança de θ i calculeu el seu biaix.*

Demostració. L'estimador de màxima versemblança es calcula imposant $\partial_\theta \log \mathscr{L}(x, \theta) = 0$, i veient quins θ ho compleixen. Efectivament, la funció de versemblança és:
$$\mathscr{L}(x, \theta) = 2^n \theta^n \left(\prod_{i=1}^n x_i\right) \exp\left\{-\theta \sum_{j=1}^n x_j^2\right\}$$
$$\log \mathscr{L}(x, \theta) = n \log 2 + n \log \theta + \sum_{i=1}^n \log x_i - \theta x_i^2$$
$$\partial_\theta \log \mathscr{L}(x, \theta) = \frac{n}{\theta} - \sum_{i=1}^n x_i^2 = 0 \iff \hat{\theta} = \frac{n}{\sum_{i=1}^n x_i^2}.$$

No ho comprovarem, però efectivament és un màxim. Per tant, ja tenim l'EMV. Si no tingués biaix, $E_\theta(\hat{\theta}) = \theta$. Anem a comprovar-ho:

$$E_\theta\left(\frac{n}{\sum_{i=1}^n x_i^2}\right) = nE_\theta\left(\frac{1}{\sum_{i=1}^n x_i^2}\right).$$

Si X és la variable aleatòria de la mostra $(x_1, \dots, x_n)$, es pot definir $Y = X^2$, de manera que $Y_i = X_i^2 \sim \operatorname{Exp}(\theta) \implies Z = \sum_i Y_i \sim \operatorname{Gamma}(\theta, n)$. Aplicant que la densitat de la $\operatorname{Gamma}(\theta, n)$ és $f(x) = \frac{\theta^n}{(n-1)!}x^{n-1}e^{-\theta x}\mathbb{1}_{(0,+\infty)}(x)$,

$$E_\theta(\hat{\theta}) = n\int_0^\infty \frac{1}{x}\cdot\frac{\theta^n}{(n-1)!}x^{n-1}e^{-\theta x}\,dx = \frac{n\theta}{n-1}\int_0^\infty \frac{\theta^{n-1}}{(n-2)!}x^{n-2}e^{-\theta x}\,dx = \frac{n\theta}{n-1}\cdot 1.$$

Per tant, $\hat{\theta}$ és esbiaixat amb biaix $\frac{\theta}{n-1}$. Un estimador sense biaix podria ser $\frac{n-1}{n}\cdot\hat{\theta}$. ∎

Exercici 1.9. *Sigui $(X_1, \dots, X_n)$ VAIID, amb variància σ^2 finita. Determineu la constant c per tal que*

$$c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2.$$

sigui un estimador no esbiaixat de σ^2.

Demostració. Anomenem $T = c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2$, volem veure que $E_{\sigma^2}(T) = \sigma^2$. Aleshores, simplement computem:

$$E_{\sigma^2}\left(c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2\right) = c\sum_{i=1}^{n-1}E(X_{i+1} - X_i)^2 = c(n-1)E((X_2 - X_1)^2).$$

Com que $E(X_1) = E(X_2)$:

$$\begin{aligned} E((X_2 - X_1)^2) &= E(X_2^2) + E(X_1^2) - 2E(X_1X_2) \overset{\text{VAIID}}{=} E(X_2^2) + E(X_1^2) - 2E(X_1)E(X_2) \\ &\overset{\text{VAIID}}{=} 2E(X_1^2) - 2(E(X_1))^2 = 2\operatorname{Var}(X_1) = 2\sigma^2. \end{aligned}$$

De manera que $E(T) = c(n-1)2\cdot\sigma^2$ i prenent $c = \frac{1}{2(n-1)}$, en efecte, T és un estimador no esbiaixat de σ^2. ∎

Exercici 1.10. *Considerem $(X_1, \dots, X_n)$ VAIID amb densitat:*

$$f_\theta(x) = e^{\theta - x}\mathbb{1}_{(0,+\infty)}(x),\ \theta > 0.$$

1. *Especifiqueu el model estadístic associat i escriviu la funció de versemblança.*
2. *Sigui $T(X_1, \dots, X_n) = \min(X_1, \dots, X_n)$. És T un estimador sense biaix de θ? En cas que tingui biaix, deduïu-ne un sense biaix, que designarem per U, i trobeu la variància.*
3. *Considerem ara l'estimador $V = \bar{X} - 1$. Té biaix? És millor que U?*

Demostració.

1. Posem $(\Omega, \mathcal{A}, \mathcal{P}) = ((0,+\infty)^n, \mathcal{B}((0,+\infty)^n), \{P_\theta \mid \theta > 0\})$ el model estadístic[3], i la funció de versemblança:

$$L(x_1,\dots,x_n;\theta) = \prod_{i=1}^n e^{\theta - x_i}\mathbb{1}_{(\theta,+\infty)}(x_i) = e^{n\theta - \sum_i x_i}\prod_{i=1}^n \mathbb{1}_{(\theta,+\infty)}(x_i) = e^{n\theta-\sum_i x_i}\mathbb{1}_{(\theta,+\infty)}(x_{(1)}).$$

On $x_{(1)} = \min(X_1,\dots,X_n)$, i $n\theta - \sum_i x_i = n(\theta - \bar{x})$, si es vol. Fixem-nos que el valor que maximitza $L(x_1,\dots,x_n;\theta)$ és $\theta = x_{(1)}$, ja que $e^{n\theta-\sum_i x_i}\mathbb{1}_{(\theta,+\infty)}(x_{(1)}) = e^{n(\theta-\bar{x})}\mathbb{1}_{(0,x_{(1)}]}(\theta)$.

2. Posem $T = \min(X_1,\dots,X_n)$, mirem $E_\theta(T) = E_\theta(\min(X_1,\dots,X_n))$. Necessitem abans $f_T(X)$, que calculem a través de la cdf:

$$\begin{aligned} P(T \le X) &= P(\min(X_1,\dots,X_n) \le x) = 1 - P(\min(X_1,\dots,X_n) > x) = 1 - \prod_{i=1} P(X_i > x) \\ &= 1 - (P(X_1 > x))^n = 1 - \left(\int_x^{+\infty} e^{\theta - y}\,dy\right)^n = 1 - e^{n(\theta-x)}, \end{aligned}$$

pel que $f_T(x)$ és la derivada respecte x, que és $ne^{n(\theta-x)}\mathbb{1}_{(\theta,+\infty)}(x)$. Aleshores,

$$\begin{aligned} E_\theta(T) = E_\theta(\min(X_1,\dots,X_n)) &= \int_\theta^{+\infty} x\cdot ne^{n(\theta-x)}\,dx = ne^{n\theta}\int_\theta^{+\infty} xe^{-nx}\,dx \\ &= ne^{n\theta}\left(\frac{1}{n}[-xe^{-nx}]_\theta^{+\infty} + \frac{1}{n}\int_\theta^{+\infty} e^{-nx}\,dx\right) = ne^{n\theta}\left(\frac{1}{n}\theta e^{-n\theta} + \frac{1}{n^2}e^{-n\theta}\right). \end{aligned}$$

El resultat d'aquesta integral és $\theta + \frac{1}{n}$. Per tant, T té biaix, que és:

$$\text{biaix}_\theta(T) = E_\theta(T) - \theta = E_\theta(T(X)) - \theta = E_\theta(X_{(1)}) - \theta = \frac{1}{n}.$$

Per tant, podem definir U com $U(x) = x_{(1)} - \frac{1}{n}$ i U serà no esbiaixat (ras i curt, li traiem el biaix). Amb tot, podem calcular la variància d'U, $\text{Var}_\theta(U) = \text{Var}_\theta(T) = \text{Var}_\theta(X_{(1)})$:

$$\begin{aligned} \text{Var}_\theta(X_{(1)}) = E_\theta((X_{(1)})^2) - (E_\theta(X_{(1)}))^2 &= \int_\theta^\infty x^2 ne^{n(\theta-x)}\,dx - \left(\theta + \frac{1}{n}\right)^2 \\ &= ne^{n\theta}\left(\frac{1}{n}\theta^2 e^{-n\theta} + \frac{2}{n}\left(\frac{1}{n}\theta e^{-n\theta} + \frac{1}{n^2}e^{-n\theta}\right)\right) - \left(\theta + \frac{1}{n}\right)^2 \\ &= \theta^2 + \frac{2\theta}{n} + \frac{2}{n^2} - \left(\theta^2 - \frac{2\theta}{n} + \frac{1}{n^2}\right) = \frac{1}{n^2}. \end{aligned} \tag{1.2}$$

3. Per veure si té biaix, tornem a calcular l'esperança $E_\theta(V)$ i mirem $E_\theta(V) - \theta$.

$$E_\theta(V) = E_\theta(\bar{X} - 1) = E_\theta(\bar{X}) - 1 = \frac{1}{n}\sum_{i=1}^n E_\theta(X_i) - 1 \overset{\text{VAIID}}{=} E_\theta(X_1) - 1 = \int_\theta^\infty xe^{\theta-x}\,dx - 1 \overset{4}{=} \theta.$$

Queda clar que V no és esbiaixat. Com $E_\theta(T) = \theta + \frac{1}{n}$, si posem $n = 1$ tenim directament que $\text{Var}_\theta(X_1) = 1$. Ja per últim:

$$\text{Var}_\theta(V) = \text{Var}_\theta(\bar{X}) = \frac{\text{Var}_\theta(X_1)}{n} \overset{(1.2)}{=} \frac{1}{n}.$$

[3] Ω no és $(\theta,+\infty)^n$, ja que θ és desconegut; de fet, per això la funció de versemblança té el producte d'indicadors.
[4] Estem usant que $E_\theta(T) = \theta + \frac{1}{n}$; com $n = 1$ el resultat de la integral queda $\theta + 1$.

Per tant, U, V no són esbiaixats però $\mathrm{Var}_\theta(U) < \mathrm{Var}_\theta(V)$ per a $n > 1$ i $\theta > 0$. Per tant, U és clarament millor estimador que V. ■

Exercici 1.11. *Tenim una mostra aleatòria simple $(X_1, \ldots, X_n)$ d'una distribució uniforme a l'interval $(\theta, \theta + 1)$. Sigui $X_{(1)} = \min(X_1, \ldots, X_n)$ i $X_{(n)} = \max(X_1, \ldots, X_n)$. Determineu la funció de versemblança i a partir d'aquests dos estimadors trobeu dos estimadors no esbiaixats de θ. Compareu-los.*

Demostració. Una variable aleatòria X segueix una distribució uniforme a l'interval $(\theta, \theta + 1)$, és a dir, $X \sim U(\theta, \theta + 1)$, si té com a funció de densitat:

$$g(x, \theta) = \mathbb{1}_{(\theta,\theta+1)}(x), \quad \mathbb{1}_{(\theta,\theta+1)}(x) = \begin{cases} 1, & \text{si } \theta < x < \theta + 1; \\ 0, & \text{otherwise.} \end{cases}$$

Així doncs, si prenem la funció de densitat conjunta obtenim que:

$$L(x_1, \ldots, x_n; \theta) = f(x, \theta) = \prod_{i=1}^{n} g(x_i, \theta) = \prod_{i=1}^{n} \mathbb{1}_{(\theta,\theta+1)}(x_i), \;\; x = (x_1, \ldots, x_n).$$

Ara bé, si volem que $f(x, \theta) \neq 0$, necessitem que $\theta < x_i < \theta + 1$ per a tot $i = 1, \ldots, n$ (per definició de la funció indicatriu) o, equivalentment, que $x_{(1)} > \theta$ i $x_{(n)} < \theta + 1$. Per tant, en particular podem escriure la funció de densitat conjunta en termes de $x_{(1)}$ i $x_{(n)}$ de la següent manera:

$$f(x, \theta) = \mathbb{1}_{(\theta.\infty)}(x_{(1)})\mathbb{1}_{(-\infty,\theta+1)}(x_{(n)}) \iff f(x, \theta) = \mathbb{1}_{(x_{(n)}-1, x_{(1)})}(\theta), \;\; x = (x_1, \ldots, x_n).$$

En particular, per a què la funció de densitat no sigui sempre nul·la necessitem que $x_{(n)} < x_{(1)} + 1$. A més, per definició, notem que sempre tindrem $x_{(1)} < x_{(n)}$. En resum, $0 < x_{(n)} - x_{(1)} < 1$. Volem trobar dos estimadors no esbiaixats de θ a partir dels dos estadístics $x_{(n)}$ i $x_{(1)}$:

1. Comencem amb l'estimador $x_{(n)}$. Primer ens caldria saber quina és la seva funció de distribució:

$$\begin{aligned} F_{x_{(n)}}(y) &= P(x_{(n)} \leq y) = P(x_1 \leq y, \ldots, x_n \leq y) = P(x_1 \leq y)^n \\ &= \begin{cases} 0, & y \leq \theta; \\ \left(\int_{-\infty}^{y} \mathbb{1}_{(\theta,\theta+1)}(x)\, dx\right)^n = (y - \theta)^n, & \theta < y < \theta + 1; \\ 1, & y \geq \theta + 1. \end{cases} \end{aligned}$$

   Pel que la funció de densitat queda $f_n(y, \theta) = n(y - \theta)^{n-1}\mathbb{1}_{(\theta,\theta+1)}(y)$. Observem que si calculem l'esperança d'aquest estimador:

$$E_\theta(x_{(n)}) = \int_{-\infty}^{+\infty} y f_n(y, \theta)\, dy = \int_{\theta}^{\theta+1} n y (y - \theta)^{n-1}\, dy = \frac{n}{n+1} + \theta.$$

   Prenent $M = x_{(n)} - \frac{n}{n+1}$, tenim un estimador sense biaix. En efecte:

$$E_\theta(M) = E_\theta(x_{(n)}) - \frac{n}{n+1} = \theta \implies \mathrm{biaix}_\theta(T) = 0.$$

Mario Vilar, 2024

Anem a comparar M i m. Considerem que un estimador és millor quan té menor variància:

$$\begin{aligned}\operatorname{Var}_\theta(M) &= E_\theta((M - E_\theta(M))^2) = E_\theta\left(\left(x_{(n)} - \frac{n}{n+1} - \theta\right)^2\right)\\ &= \int_{-\infty}^{+\infty}\left(y - \frac{n}{n+1} - \theta\right)^2 f_n(y,\theta)\,dy = \int_\theta^{\theta+1} n\left(y - \frac{n}{n+1} - \theta\right)^2 (y-\theta)^{n-1}\,dy\\ &= \frac{n}{n+2} - \frac{n^2}{(n+1)^2}.\end{aligned}$$

D'altra banda:

$$\begin{aligned}\operatorname{Var}_\theta(m) &= E_\theta((m - E_\theta(m))^2) = E_\theta\left(\left(x_{(1)} - \frac{1}{n+1} - \theta\right)^2\right) = \int_{-\infty}^{+\infty}\left(y - \frac{1}{n+1} - \theta\right)^2 f_1(y,\theta)\,dy\\ &= \int_\theta^{\theta+1} n\left(y - \frac{1}{n+1} - \theta\right)^2 (\theta+1-y)^{n-1}\,dy = \frac{n}{n+2} - \frac{n^2}{(n+1)^2}.\end{aligned}$$

Com que tenen la mateixa variància, podem dir que són *igual de bons*. ■

Exercici 1.12. *Siguin $X_1, \ldots, X_n$ variables aleatòries independents i idènticament distribuïdes, amb funció de densitat $f_\theta(x) = \frac{2x}{\theta^2}$, $0 < x \le \theta$, $\theta > 0$.*

1. *Determineu l'estimador de màxima versemblança del paràmetre θ.*
2. *Determineu, també, un estimador de θ, T, sense biaix.*
3. *Compareu-lo amb $S = \frac{3}{2}\bar{X}$.*

Demostració.

1. Ens situem en $\Omega = [0, +\infty)^n$. La funció de versemblança és:

$$L(x,\theta) = \prod_{i=1}^n \frac{2}{\theta^2} x_i \mathbb{1}_{[0,\theta]}(x_i) = \frac{2^n}{\theta^{2n}} \prod_{i=1}^n x_i \prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i) = \frac{2^n}{\theta^{2n}} \prod_{i=1}^n x_i \mathbb{1}_{[0,\theta]}(x_{(n)}).$$

Atès que la funció de versemblança no és estrictament positiva, no podem aplicar logaritmes; com que tampoc és contínua respecte de θ, no pode derivar. La versemblança és 0 excepte en l'interval $[x_{(n)}, \infty)$, on és decreixent en θ. Trobarem el seu màxim en el límit inferior de l'interval; per tant, l'estimador de màxima versemblança és $\hat{\theta} = X_{(n)}$.

$$\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i) = 1 \iff 0 \le x_i \le \theta.$$

2. Hem de calcular $E_\theta(\hat{\theta}) = E_\theta(\bar{X}_n)$, necessitarem $F_{X_{(n)}}(z) = P_\theta(X_{(n)} \le z)$.

$$P_\theta(X_{(n)} \le z) = (P_\theta(X_1 \le z)^n = \left(\int_0^z \frac{2}{\theta^2} x\,dx\right)^n = \left(\frac{z^2}{\theta^2}\right)^n, \quad 0 \le z < \theta.$$

Aleshores, derivant:

$$f_{X_{(n)}}(z) = 2n\frac{z^{2n-1}}{\theta^{2n}}\mathbb{1}_{(0,\theta)}(z).$$

És clar que quan $z > \theta$, $F_{X_{(n)}}(z) = 1$. Per tant:

$$E_\theta(\hat{\theta}) = \int_0^\theta z \cdot 2n \frac{z^{2n-1}}{\theta^{2n}}\, dz = \frac{2n}{2n+1}\theta \implies E_\theta(\bar{X}_n^2) = \frac{2n}{2n+2}\theta^2$$

A partir d'aquí, deduïm que $T = \frac{2n+1}{2n} X_{(n)}$ no té biaix.

3. Ara, per a $S = \frac{3}{2}\bar{X}_n$.

$$E_\theta(S) = \frac{3}{2} E_\theta(\bar{X}) = \frac{3}{2} E_\theta(X_1) = \frac{3}{2} \int_0^\theta x \frac{2}{\theta^2} x\, dx = \theta.$$

Comparem els riscos $R_T(\theta)$ i $R_S(\theta)$ i ens quedem amb el més petit. En aquest sentit:

$$\begin{aligned} R_T(\theta) = \operatorname{Var}_\theta(T) = \frac{(2n+1)^2}{(2n)^2} \operatorname{Var}_\theta(\bar{X}_n) &= \frac{(2n+1)^2}{(2n)^2} (E_\theta(X_{(n)}^2) - (E_\theta(X_{(n)}))^2) \\ &= \frac{(2n+1)^2}{(2n)^2} \left(\frac{2n}{2n+2}\theta^2 - \left(\frac{2n}{2n+1}\right)^2 \theta^2 \right) = \theta^2 \frac{1}{4n^2+4n}. \end{aligned}$$

Per altra banda:

$$R_S(\theta) = \operatorname{Var}_\theta\left(\frac{3}{2}\bar{X}_n\right) = \frac{9}{4}\operatorname{Var}_\theta(\bar{X}_n) = \frac{9}{4} \cdot \frac{1}{n} \operatorname{Var}_\theta(X_1) = \frac{9}{4} \cdot \frac{1}{n} (E_\theta(X_1^2) - E_\theta^2(X_1)) = \frac{\theta^2}{8n}.$$

Hem usat que $\operatorname{Var}_\theta(\bar{X}_n) = \frac{1}{n} \operatorname{Var}_\theta(X_1)$. Resulta:

$$\operatorname{Var}_\theta(T) = \frac{\theta^2}{4n(n+1)} < \frac{\theta^2}{8n} = \operatorname{Var}(\hat{\theta}).$$

Pel que és preferible T perquè n^2 domina sobre n, quan $n \to \infty$. ■

Exercici 1.13. *Sigui $X_1, \dots, X_n$ una mostra d'una variable aleatòria X amb llei de Poisson de paràmetre $\lambda > 0$. Volem estimar $g(\lambda) = \lambda^2$. Definim:*

$$T = \frac{1}{n^2}\left(\left(\sum_{i=1}^n X_i\right)^2 - \sum_{i=1}^n X_i \right).$$

És T un estimador sense biaix de $g(\lambda)$? Calculeu la fita de Cramer-Rao.

Demostració. Primer provarem que T és un estimador sense biaix i després trobarem la fita de Cramer-Rao. Per veure-ho, provarem que $E_\lambda(T) = \lambda^2 = g(\lambda)$. L'esperança d'una variable aleatòria X amb llei de Poisson de paràmetre λ és λ:

$$E(X) = \sum_{k=0}^\infty k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda \sum_{k=1}^\infty e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

I el moment de segon ordre és:

$$\begin{aligned} E(X^2) &= \sum_{k=0}^\infty k^2 e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^\infty k e^{-\lambda} \frac{\lambda^k}{(k-1)!} = \sum_{k=1}^\infty (k-1) e^{-\lambda} \frac{\lambda^k}{(k-1)!} + \sum_{k=1}^\infty e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\ &= \lambda \sum_{k=2}^\infty (k-1) e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} + \lambda \sum_{k=1}^\infty e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda^2 \sum_{k=2}^\infty e^{-\lambda} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

La suma de n variables independents i idènticament distribuïdes amb llei de Poisson de paràmetre λ és una variable amb llei de Poisson de paràmetre $n\lambda$ (cf. *Probabilitats*). Per tant, $Z = \sum_{i=1}^{n} X_i$ és una variable de Poisson de paràmetre $n\lambda$ i tenim $E(Z) = n\lambda$ i $E(Z^2) = n^2\lambda^2 + n\lambda$. Per tant:

$$E(T) = \frac{1}{n^2}(E(Z^2) - E(Z)) = \frac{1}{n^2}(n^2\lambda^2 + n\lambda - n\lambda) = \lambda^2.$$

Per tant, T és un estimador no esbiaixat de $g(\lambda) = \lambda^2$. Calculem ara la fita de Cramer-Rao. Deduïm primer la informació de Fisher d'un model de Poisson de paràmetre λ. Observem que la funció de probabilitat es pot escriure com:

$$\begin{aligned} p(k,\lambda) &= \exp(-\lambda + k\log\lambda - \log k!) \\ \log p(k,\lambda) &= -\lambda + k\log\lambda - \log k! \\ \partial_\lambda \log p(k,\lambda) &= -1 + \frac{k}{\lambda} \\ \partial_\lambda^2 \log p(k,\lambda) &= -\frac{k}{\lambda^2}. \end{aligned}$$

En definitiva, $p(k,\lambda)$ és un model exponencial regular amb informació de Fisher $I_n(\lambda) = \frac{n}{\lambda}$[5]. La fita de Cramer-Rao queda com:

$$V_\lambda(T) \geq \frac{(g'(\lambda))^2}{I_n(\lambda)} = \frac{(2\lambda)^2\lambda}{n} = \frac{4\lambda^3}{n}.$$

Podríem calcular $V_\lambda(T)$ si volguéssim, ja que $V_\lambda(T) = E_\lambda(T^2) - (E(T))^2 = E_\lambda(T^2) - \lambda^4$. ■

Exercici 1.14. *Tenim n observacions independents amb densitat* $f_\theta(x) = \frac{\theta}{(1+x)^{1+\theta}}$, $x > 0, \theta > 1$.

1. *Determina l'estimador de màxima versemblança de θ i el de $\frac{1}{\theta}$. Calcula el biaix en ambdós casos.*
2. *Calcula la fita de Cramer-Rao per a estimador sense biaix de θ i de $\frac{1}{\theta}$. Són eficients els estimadors obtinguts?*
3. *Troba un estimador de θ usant el mètode dels moments.*

Demostració.

1. Definim l'espai estadístic com $((0,+\infty)^n, \mathscr{B}((0,+\infty)^n), \{P_\theta \mid \theta > 1\})$. La funció de versemblança es pot escriure com:

$$\mathscr{L}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \theta^n \prod_{i=1}^{n} \frac{1}{(1+x_i)^{1+\theta}}.$$

No tenim cap indicatriu en la densitat que ens donen, així que no cal que ens preocupem per ella. En

[5] Hem usat que $I_1(\lambda) = \frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$.

qualsevol cas, sabem que l'EMV compleix $\partial_\theta \log L(x_1, \dots, x_n; \theta) = 0$:

$$\begin{aligned}
\log \mathscr{L}(x_1, \dots, x_n; \theta) &= n \log \theta - (1+\theta) \sum_{i=1}^{n} \log(1 + x_i), \\
\partial_\theta \log \mathscr{L}(x, \theta) &= \frac{n}{\theta} - \sum_{i=1}^{n} \log(1 + x_i). \\
\partial_\theta^2 \log \mathscr{L}(x, \theta) &= -\frac{n}{\theta^2}.
\end{aligned}$$

Igualant aquesta expressió a zero queda clar que l'estimador màxim versemblant de θ és $\frac{n}{S}$, on $S = \sum_{i=1}^{n} \log(1 + X_i)$. Observem que la segona derivada $\partial_\theta^2 \log L(x, \theta) < 0$, de manera que estem davant d'un màxim. Per la *propietat d'invariància de l'estimador màxim versemblant*, l'estimador màxim versemblant de $\alpha := \frac{1}{\theta}$ és $\widehat{\left(\frac{1}{\theta}\right)} = \frac{1}{\hat{\theta}} = \frac{S}{n}$. Això també es pot provar directament escrivint:

$$\mathscr{L}(x, \theta) = \mathscr{L}(x, \alpha) = \frac{1}{\alpha^n} \prod_{i=1}^{n} \frac{1}{(1 + x_i)^{1+\frac{1}{\alpha}}}.$$

I calculant directament l'EMV d'α. Per a calcular el biaix de $\frac{S}{n}$ i $\frac{n}{S}$, el millor és veure quina és la llei de S. Definim $Y_i = \log(1 + X_i)$. Tenim, per a tot $y \geq 0$[6]:

$$P(Y_i \leq y) = P(1 + X \leq e^y) = P(X_i \leq e^y - 1) \implies F_{Y_i}(y) = F_{X_i}(e^y - 1).$$

Si derivem, obtenim la densitat $f_{Y_i}(y)$:

$$f_{Y_i}(y) = f_{X_i}(e^y - 1)e^y = e^y \cdot \theta \cdot e^{-y(1+\theta)} = \theta e^{-\theta y}.$$

Per tant, $Y_i \sim \exp(\theta)$, per a tot $i = 1, \dots, n$. $S = \sum_{i=1}^{n} Y_i$ té llei Gamma de paràmetres n i θ. $E_\theta(S) = nE_\theta(Y) = \frac{n}{\theta}$; per tant, $\frac{S}{n}$ és un estimador no esbiaixat de $\frac{1}{\theta}$:

$$E(S) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E_\theta(\log(1 + X_i))}_{\frac{1}{\theta}} = \frac{1}{n} \cdot n \cdot \frac{1}{\theta} = \frac{1}{\theta}, \ S \text{ no té biaix.}$$

La seva densitat és:

$$f_S(x) = \frac{\theta^n}{(n-1)!} x^{n-1} e^{-\theta x}.$$

Calculem l'esperança de $\frac{n}{S}$, tenim:

$$\begin{aligned}
E_\theta\left(\frac{n}{S}\right) = \int_0^\infty \frac{n\theta^n}{(n-1)!} x^{n-2} e^{-\theta x}\, dx = \int_0^\infty \frac{n\theta}{(n-1)!} y^{n-2} e^{-y}\, dy &= \frac{n\theta}{(n-1)!} \int_0^\infty y^{n-2} e^{-y}\, dy \\
&= \frac{n\theta}{(n-1)}.
\end{aligned}$$

[6] Es pot fer amb un canvi de variable $h : (0, +\infty) \longrightarrow (0, +\infty)$ tal que $x \longmapsto y := \log(1 + x)$, però dependríem de la indicatriu (per forçar la bijectivitat en h) i quedaria $\theta e^{-\theta y} \mathbb{1}_{(0,+\infty)}(y)$.

Per tant, $\frac{n}{S}$ és un estimador amb biaix de θ; està clar que:

$$\text{biaix}_\theta\left(\frac{n}{S}\right) = E_\theta\left(\frac{n}{S}\right) - \theta = \left(\frac{n}{n-1} - 1\right)\theta = \frac{\theta}{n-1}.$$

Ens val amb prendre $T = \frac{n-1}{S}$ per assegurar que T és un estimador no esbiaixat de θ.

2. Calculem la fita de Cramer-Rao per a θ. Podem calcular la informació de Fisher $I_F(\theta) = -E_\theta(\partial_\theta^2 \log L(x,\theta))$:

$$\begin{aligned}\log L(x_1,\dots,x_n;\theta) &= n\log\theta - (1+\theta)\sum_{i=1}^n \log(1+x_i),\\ \partial_\theta \log L(x,\theta) &= \frac{n}{\theta} - \sum_{i=1}^n \log(1+x_i).\\ \partial_\theta^2 \log L(x,\theta) &= -\frac{n}{\theta^2}.\end{aligned}$$

Per tant, $I_F(\theta)$ no depèn d'x i $E(\partial_\theta^2 \log L(x,\theta)) = \partial_\theta^2 \log L(x,\theta) = -\frac{n}{\theta^2}$, pel que $I_F(\theta) = \frac{n}{\theta^2}$. Com la fita de Cramer-Rao per a un estimador no esbiaixat T de θ és $V_\theta(T) \geq \frac{\theta^2}{n}$, i si tenim un estimador U no esbiaixat de $g(\theta) = \frac{1}{\theta}$, la seva fita de Cramer-Rao és:

$$V_{\frac{1}{\theta}}(U) \geq \frac{\theta^2}{n}\left(-\frac{1}{\theta^2}\right)^2 = \frac{1}{n\theta^2}.$$

Per a veure si $\frac{S}{n}$ i $\frac{n-1}{S}$ són estimadors eficients de $\frac{1}{\theta}$ i θ respectivament, hem de calcular les seves variàncies i veure si coincideixen amb la corresponent fita de Cramer-Rao. D'una banda:

$$V_{\frac{1}{\theta}}\left(\frac{S}{n}\right) = \frac{1}{n}V(Y) = \frac{1}{n\theta^2}.$$

Per tant, $\frac{S}{n}$ és un estimador eficient de $\frac{1}{\theta}$. D'altra banda:

$$\begin{aligned}E_\theta\left(\left(\frac{n-1}{S}\right)^2\right) &= (n-1)^2\int_0^\theta \frac{\theta^n}{(n-1)!}x^{n-3}e^{-\theta x}\,dx = (n-1)^2\int_0^\infty \frac{\theta^2}{(n-1)!}y^{n-3}e^{-y}\,dy\\ &= (n-1)^2\frac{\theta^2}{(n-1)!}\int_0^\infty y^{n-3}e^{-y}\,dy = (n-1)^2\frac{\theta^2}{(n-1)(n-2)} = \frac{n-1}{n-2}\theta^2.\end{aligned}$$

Aleshores:

$$V_\theta\left(\frac{n-1}{S}\right) = \frac{n-1}{n-2}\theta^2 - \theta^2 = \left(\frac{n-1}{n-2} - 1\right)\theta^2 = \frac{\theta^2}{n-2}.$$

Observem que és un valor més gran que la fita de Cramer-Rao, que és $\frac{\theta^2}{n}$. Per tant, $\frac{n-1}{S}$ no és pas un estimador eficient.

3. Ara busquem l'estimador de θ pel mètode dels moments. Tenim:

$$\begin{aligned}E_\theta(X) &= \int_0^\infty x\cdot\frac{\theta}{(1+x)^{(1+\theta)}}\,dx = \int_1^{+\infty}\frac{\theta(u-1)}{u^{(1+\theta)}}\,du = \theta\int_1^{+\infty}\frac{1}{u^\theta}\,du - \theta\int_1^{+\infty}\frac{1}{u^{(1+\theta)}}\,du\\ &= -\frac{\theta}{1-\theta} - 1 = \frac{1}{\theta-1}.\end{aligned}$$

Per tant, el mètode dels moments ens diu que $\bar{X}_n$ és un bon estimador de $\frac{1}{\theta-1}$, en particular, sense biaix. Per tant, $\hat{\theta} = 1 + \frac{1}{\bar{X}_n}$ és l'estimador adequat de θ segons el mètode dels moments. ■

Exercici 1.15. *Considerem una mostra de mida n d'una variable aleatòria X amb llei lognormal de paràmetres (μ, σ^2) és a dir,* $\log X \sim N(\mu, \sigma^2)$.

1. *Demostreu que si σ^2 és conegut, l'estimador de màxima versemblança de μ és:*

$$\hat{\mu} = \frac{1}{n}\left(\sum_{i=1}^{n} \log x_i\right).$$

2. *Trobeu l'estimador de màxima versemblança de $E(X)$.*
3. *Si μ és coneguda, trobeu l'estimador de màxima versemblança de σ^2.*
4. *Trobeu l'estimador de màxima versemblança de (μ, σ^2).*

Demostració. Sabem que X segueix una distribució lognormal, és a dir, $Y = \log X$ segueix una distribució $N(\mu, \sigma^2)$ amb funció de densitat:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \ y \in \mathbb{R}.$$

Aleshores, quina es la funció de densitat de X? Considerem:

$$\begin{array}{cccc} g: & \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ & y & \longmapsto & e^y \\ & \log x & \longmapsto & x \end{array}$$

X té, doncs, per funció de densitat la següent:

$$\begin{aligned} f_X(y) &= f_Y(g^{-1}(x)) \cdot |(g^{-1}(x))'| \cdot \mathbb{1}_{g(\mathbb{R})}(x) = f_X(\log x)\left|\frac{1}{x}\right| \cdot \mathbb{1}_{(0,+\infty)}(x) \\ &= \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{(0,+\infty)}(x). \end{aligned}$$

Per tant, el logaritme de la funció de versemblança serà:

$$\log \mathscr{L}(x_1, \dots, x_n; \mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\log x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2,$$

tal que $(x_1, \dots, x_n) \in (0, +\infty)^n$.

1. Calcularem l'estimador de màxima versemblança de μ suposant σ^2 conegut.

   Observació 1.16. Sigui $\theta = (\theta_1, \dots, \theta_k)$, recordem el procediment per obtenir un EMV de θ_j donada una mostra $(x_1, \dots, x_n)$:

   1.1. Escriure la funció de versemblança $\mathscr{L}(\theta) = \mathscr{L}(x_1, \dots, x_n; \theta)$.
   1.2. Escriure el logaritme de la funció de versemblança, $\log \mathscr{L}(x_1, \dots, x_n; \theta)$.
   1.3. Obtenir el θ_j tal que $\frac{\partial}{\partial\theta_j}\log\mathscr{L}(x_1, \dots, x_n; \theta) = 0$ i denotar-lo per $\hat{\theta}_j$.
   1.4. Comprovar que és, efectivament, un màxim, fent la segona derivada per exemple.

La primera derivada és zero quan $\mu = \frac{1}{n}\sum_{i=1}^{n}\log x_i$:

$$\partial_\mu \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sigma^2} \sim \sum_{i=1}^{n}(\log x_i - \mu) = 0 \iff \mu = \frac{1}{n}\sum_{i=1}^{n}\log x_i.$$

La segona derivada és negativa, pel que tenim un màxim:

$$\partial_\mu^2 \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{\sigma^2} < 0.$$

Per tant, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\log x_i$ és l'estimador de màxima versemblança de μ.

2. Volem trobar l'estimador de màxima versemblança de $E(X)$. El que hem de fet, doncs, és calcular $E(X)$ i aplicar després el teorema d'invariància funcional. Observem, doncs:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} e^y \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2 - 2y\mu + \mu^2 - 2y\sigma^2}{2\sigma^2}\right) dy \\ &= e^{\mu+\frac{\sigma^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-(\mu+\sigma^2))^2}{2\sigma^2}\right) dy = e^{\mu+\frac{\sigma^2}{2}} \int_{-\infty}^{+\infty} h(y)\, dy, \end{aligned}$$

però l'integrand $h(y)$ és la funció de densitat d'una distribució $N(\mu+\sigma^2, \sigma^2)$. Per tant, $\int_{-\infty}^{+\infty} h(y)\, dy = 1$ i obtenim que $E(X) = e^{\mu+\frac{\sigma^2}{2}}$. Com que $E(X)$ és una funció bijectiva de μ i $\hat{\mu}$ és l'EMV de μ, podem aplicar el teorema d'invariància funcional, pel que:

$$\hat{m} = e^{\hat{\mu}+\frac{\sigma^2}{2}} = e^{\frac{\sigma^2}{2}} \prod_{i=1}^{n} \sqrt[n]{x_i}$$

és l'estimador de màxima versemblança de $E(X)$.

3. L'estimador de màxima versemblança de σ^2 suposant μ conegut es calcula de manera similar a com hem fet al primer apartat.

$$\partial_{\sigma^2} \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(\log x_i - \mu)^2 = \frac{n}{\sigma^2}\left(-1 + \frac{1}{n\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2\right)$$

$$\partial_{\sigma^2} \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) = 0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \mu)^2$$

$$\partial_{\sigma^2}^2 \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(\log x_i - \mu)^2 = \frac{n}{2\sigma^4}\left(1 - \frac{2}{n\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2\right).$$

I $\partial_{\sigma^2}^2 \log \mathscr{L}(x_1, \ldots, x_n; \mu, \sigma^2) < 0$ quan $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \mu)^2$, és fàcil comprovar-ho. Per tant, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \mu)^2$.

Observació 1.17. També podríem haver trobat l'estimador de màxima versemblança de σ i aleshores aplicar el principi d'invariància; és a dir, si $\hat{\sigma}$ és l'estimador de màxima versemblança de σ, llavors $\hat{\sigma}^2$ ho és de σ^2.

4. Per definició d'estimador de màxima versemblança de μ (el qual no depèn de σ^2) tenim que:

$$\mathscr{L}(x_1, \dots, x_n; \mu, \sigma^2) \leq \mathscr{L}(x_1, \dots, x_n; \hat{\mu}, \sigma^2).$$

   Ara bé, com $\hat{\mu}$ és conegut, obtenim ara prenent $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \hat{\mu})^2$ que:

$$\mathscr{L}(x_1, \dots, x_n; \hat{\mu}, \sigma^2) \leq \mathscr{L}(x_1, \dots, x_n; \hat{\mu}, \hat{\sigma}^2).$$

   En particular, si ho ajuntem tot i prenem el suprem entre tots els parells de paràmetres (μ, σ^2) obtenim que:

$$\sup_{(\mu,\sigma^2)} \mathscr{L}(x_1, \dots, x_n; \mu, \sigma^2) = \mathscr{L}(x_1, \dots, x_n; \hat{\mu}, \hat{\sigma}^2),$$

   i l'estimador de màxima versemblança de (μ, σ^2) és $(\hat{\mu}, \hat{\sigma}^2)$, on:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log x_i \text{ i } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \hat{\mu})^2.$$ ■

2

Llista 2

Exercici 2.1. *Sigui $X_1, \dots, X_n$ VAIID amb llei $P(X = m) = \frac{a^m}{(1+a)^{m+1}}$, $m = 0, 1, 2, \dots$, a paràmetre desconegut.*

1. *Trobeu l'estimador del màxim de versemblança del paràmetre a.*
2. *Suposem $a \in (0, 1]$. Trobeu l'estimador del màxim de versemblança d'a.*
3. *És eficient l'estimador trobat al primer apartat? Trobeu la cota de Cramer-Rao i la funció d'informació de Fisher.*

Exercici 2.2. *Considerem $X_1, \dots, X_n$ VAIID amb llei normal de mitjana μ i variància $\sigma^2 = 16$. Trobeu el menor valor de n per tal que $[\bar{X} - 1, \bar{X} + 1]$ sigui un interval de confiança per a μ de nivell 0.95.*

Demostració. Tenim $X_i \sim \mathcal{N}(\mu, 4^2)$, i a més:

$$\operatorname{Var}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i \operatorname{Var}(X_i) = \frac{16}{n} \implies \frac{1}{n}\sum_i X_i \sim N\left(\mu, \frac{16}{n}\right).$$

Ara, podem definir $T = \frac{\bar{X}_n - \mu}{\sqrt{\frac{16}{n}}} \sim N(0, 1)$, de manera que ja tenim la nostra funció pivotant i:

$$P_\mu(-B_\alpha \leq T \leq B_\alpha) = \alpha \iff P_\mu(-B_{0.95} \leq T \leq B_{0.95}) = 0.95.$$

El nostre objectiu ara és aïllar la μ per trobar-ne l'interval de confiança més òptim. Notem que podem reescriure $T = \frac{\bar{X}\mu}{4}\sqrt{n}$, i:

$$-\frac{1.96 \cdot 4}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{1.96 \cdot 4}{\sqrt{n}} \iff -\frac{1.96 \cdot 4}{\sqrt{n}} \leq \mu - \bar{X} \leq \frac{1.96 \cdot 4}{\sqrt{n}} \iff \bar{X} - \frac{1.96 \cdot 4}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96 \cdot 4}{\sqrt{n}}.$$

Així doncs, l'interval de confiança al 0.95 és $[\bar{X} - \frac{1.96\cdot 4}{\sqrt{n}}, \bar{X} + \frac{1.96\cdot 4}{\sqrt{n}}]$. Com que volem que $\frac{1.96\cdot 4}{\sqrt{n}} = 1$, necessitem una $n \geq 61.47$. ■

Exercici 2.3. *Suposant que el contingut de nicotina dels cigarrets de determinada marca segueix una distribució $N(30, \sigma^2)$. Agafem a l'atzar deu cigarrets d'aquesta marca i obtenim:*

$$\frac{1}{10}\sum_{i=1}^{n}(X_i - 30)^2 = 12.4.$$

Trobeu un interval de confiança per a σ amb un nivell de confiança de 0.9.

Demostració. Tenim una mostra d'una variable aleatòria X amb distribució normal de mitjana $\mu = 30$ i σ^2 desconeguda. Tenim que $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$. Està clar que $\frac{X-30}{\sigma}$ té una distribució normal estàndard. Per tant:

$$T := \sum_{i=1}^{n}\left(\frac{X_i - 30}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - 30)^2$$

és un estadístic amb la distribució $\chi^2_{(n)}$ de Pearson[7], amb n graus de llibertat. Observem que T és una funció pivotant perquè la seva llei no depèn del paràmetre σ^2. En aquest cas, assumim $n = 10$ i, aleshores, podem calcular a, b tals que $P(a \leq T \leq b) = P(a \leq \chi^2_{(10)} \leq b) = 0.9$.

$$P(\chi^2_{(10)} \leq a) = 0.05 \implies a = 3.94$$
$$P(\chi^2_{(10)} \leq b) = 0.95 \implies b = 18.30$$

Volem arribar a demostrar que l'interval observat és $[\frac{124}{18.30}, \frac{124}{3.94}] = [6.78, 31.47]$. Efectivament, tenim que:

$$a \leq \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \leq b \iff \frac{1}{b} \leq \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \mu)^2} \leq \frac{1}{a} \iff \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{b} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{a}.$$

En el nostre cas tenim que $\sum_{i=1}^{n}(X_i - 30)^2 = 12.4 \cdot 10 = 124$, pel que l'interval de confiança resulta:

$$I_{0.9}(\sigma^2) = \left[\frac{\sum_{i=1}^{n}(X_i - 30)^2}{b}, \frac{\sum_{i=1}^{n}(X_i - 30)^2}{a}\right] = \left[\frac{124}{18.30}, \frac{124}{3.94}\right] = [6.78, 31.47].$$

Si volem un interval per a σ tenim $I_{0.9}(\sigma) = [\sqrt{6.78}, \sqrt{31.47}] = [2.60, 5.61]$. ■

Exercici 2.4. *Considerem una mostra de mida n d'una variable aleatòria amb funció de densitat:*

$$f(x) = \beta x^{\beta-1}\mathbb{1}_{(0,1)}(x),\ \beta > 0.$$

1. *Determineu la llei de $Y = -\ln X$.*
2. *Trobeu una funció pivotant per construir un interval de confiança per β ($n \geq 1$).*

[7] A la literatura matemàtica sempre és anomenada χ-quadrat, però a Teoria se l'ha anomenada (al criteri de l'escriptor d'aquestes notes, erròniament) χ^2 de Pearson. Val a dir, doncs, que Pearson engloba certes distribucions, entre elles la χ^2.

3. *Trobeu un interval de confiança per β al 95% si $n = 1$ (i $x_1 = 0.5$).*

Demostració.

1. Hem de fer un canvi de variable, que anomenarem g:

$$\left.\begin{array}{cccc} g: & (0,1) & \longrightarrow & (0,+\infty) \\ & x & \longmapsto & -\ln x \\ & e^{-y} & \longleftarrow & y \end{array}\right\} \Longrightarrow f_Y(y) = e^{-y}\beta(e^{-y})^{\beta-1} \cdot \mathbb{1}_{(0,+\infty)}(y) = \beta e^{-\beta y} \cdot \mathbb{1}_{(0,+\infty)}(y).$$

Per tant, $Y \sim \text{Exp}(\beta)$, pel que és conegut que $P(Y \leq y) = 1 - e^{-\beta y}$, $y > 0$.

2. Prenent $y = \frac{t}{\beta}$, és clar que $P(Y \leq \frac{t}{\beta}) = 1 - e^{-t}$. Definint, doncs, $T := \beta Y = -\beta \ln X$, $P(T \leq t) = 1 - e^{-t}$, és a dir, T és una variable amb llei exponencial de paràmetre 1; per tant, és independent del paràmetre β.

$$Z := \sum_{i=1}^{n} T_i = -\beta \sum_{i=1}^{n} \ln X_i \sim \text{Gamma}(1, n) \implies Z \text{ és funció pivotant.}$$

Aleshores, si prenem $\pi(x, \beta)$ la funció pivotant:

$$P_\beta(a \leq \pi(x,\beta) \leq b) = P_\beta(a \leq \text{Gamma}(1,n) \leq b) = \gamma = 0.95.$$

Com que $P_\beta(\text{Exp}(1) \leq b) = \frac{1+\gamma}{2} = 0.975$ i $P_\beta(\text{Exp}(1) \leq a) = \frac{1-\gamma}{2} = 0.025$, $1 - e^{-b} = 0.975$ i $1 - e^{-a} = 0.025$. Aïllant, $b = -\ln 0.025 \approx 3.689$ i $a = -\ln 0.975 \approx 0.0253$. L'interval de confiança quedaria com:

$$P_\beta\left(a \leq -\beta \sum_{i=1}^{n} \ln X_i \leq b\right) \implies I_\gamma(\beta) = \left[\frac{a}{-\sum_{i=1}^n \log X_i}, \frac{b}{-\sum_{i=1}^n \log X_i}\right].$$

3. Hem d'aplicar l'interval anterior al cas $n = 1$ i $\gamma = 0.95$. Observem que en aquest cas, $\text{Gamma}(1,1) \sim \text{Exp}(1)$, és a dir, la funció de distribució de la Gamma seria $F(x) = 1 - e^{-x}$, per a $x > 0$, i a, b quedarien igual; la diferència, doncs, radicaria en el denominador. Aleshores, si $x_1 = 0.5$:

$$I_{0.95}(\beta) = \left[\frac{0.025}{-\ln(0.5)}, \frac{3.689}{-\ln(0.5)}\right] = [0.036, 5.322].$$

∎

Exercici 2.5. *Siguin $X_1, \ldots, X_n$ VAIID amb llei uniforme $U(0,\theta)$, $\theta > 0$. Designem per $X_{(n)}$ el màxim de $X_1, \ldots, X_n$. Calculeu la distribució de $\frac{X_{(n)}}{\theta}$ i utiliteu aquest resultat per a construir un inverval de confiança per a θ de nivell 0.9.*

Demostració. Per calcular la funció de distribució de $X_{(n)}$, $F_{X_{(n)}}(t)$, considerem $P(X_{(n)} \leq t)$. Aleshores:

$$\begin{aligned} P(X_{(n)} \leq t) = P(\max(X_1,\ldots,X_n) \leq t) &= P(X_1 \leq t) \cdots P(X_n \leq t) \\ &= P(X_1 \leq t)^n \overset{8}{=} \left(\int_0^t \frac{1}{\theta}\, dx\right)^n = \left(\frac{t}{\theta}\right)^n. \end{aligned}$$

Recordem que volem calcular la distribució de $\frac{X_{(n)}}{\theta}$, ara solament hem calculat la de $X_{(n)}$. Com $P(\frac{X_{(n)}}{\theta} \leq t)$ és $P(X_{(n)} \leq t\theta)$, definim $t = y\theta$ i calculem $F_{X_{(n)}}(y\theta)$:

$$F_{X_{(n)}}(y\theta) = \begin{cases} 0, & \text{si } \theta y < 0 \xrightarrow{\theta>0} y < 0; \\ y^n, & \text{si } 0 \leq \theta y < \theta \xrightarrow{\theta>0} 0 < y < 1; \\ 1, & \text{si } \theta y \geq \theta \xrightarrow{\theta>0} y \geq 1. \end{cases} \implies F_{X_{(n)}}(y\theta) = P(X_{(n)} \leq y\theta) = P\left(\frac{X_{(n)}}{\theta} \leq y\right),$$

com volíem veure. Ara hem de construir un interval de confiança per a θ de nivell 0.9:

$$P_\theta\left(a \leq \frac{X_{(n)}}{\theta} \leq b\right) = 0.9 \iff P_\theta\left(\frac{1}{b} \leq \frac{\theta}{X_{(n)}} \leq \frac{1}{a}\right) = 0.9 \iff P_\theta\left(\frac{X_{(n)}}{b} \leq \theta \leq \frac{X_{(n)}}{a}\right) = 0.9.$$

A partir d'aquí no hi ha una manera única d'arribar a la solució, depèn de com triem a i b. Optarem per fer-ho usant simetria (és a dir, si volem una confiança de 0.9, fixarem $P\left(\frac{X_{(n)}}{\theta} \leq a\right) = \frac{1-0.9}{2} = 0.05$ i $P_\theta\left(\frac{X_{(n)}}{\theta} \leq b\right) = \frac{1+0.9}{2} = 0.95$, de manera que:

$$P_\theta\left(\frac{X_{(n)}}{\theta} \leq a\right) = F_{X_{(n)}}(a\theta) = a^n \implies a^n = 0.05 \iff a = \sqrt[n]{0.05}$$
$$P_\theta\left(\frac{X_{(n)}}{\theta} \leq b\right) = F_{X_{(n)}}(b\theta) = b^n \implies b^n = 0.95 \iff b = \sqrt[n]{0.95}$$

I l'interval de confiança seria: $\left[\frac{X_{(n)}}{\sqrt[n]{0.95}}, \frac{X_{(n)}}{\sqrt[n]{0.05}}\right]$. ■

Exercici 2.6. *Considerem $X_1, \ldots, X_n$, n còpies independents d'una variable aleatòria X amb funció de densitat $f_\theta(x) = (1-\theta)\mathbb{1}_{[-\frac{1}{2},0]}(x) + (1+\theta)\mathbb{1}_{(0,\frac{1}{2}]}(x)$, on $\theta \in [-1, 1]$.*

1. *Determineu un estimador del màxim de versemblança del paràmetre θ.*
2. *Determineu un estimador de θ pel mètode dels moments.*
3. *Determineu la cota de Cramer-Rao per estimadors sense biaix de θ.*
4. *Quins dels estimador calculats és preferible? Algun d'ells és eficient?*

Demostració.

1. Hem de calcular la funció de versemblança i veure que $\partial_\theta \log \mathscr{L}(x_1, \ldots, x_n; \hat{\theta}) = 0$ (recordem que els

[8] Sabem que $X_1 \sim U(0,\theta)$, pel que $f_{X_1}(x) = \frac{1}{\theta} \cdot \mathbb{1}_{(0,\theta)}(x)$ i $F_{X_1}(x)$ n'és la integral entre zero i $x \leq \theta$: $\int_0^x \frac{1}{\theta}\, dt$.

logaritmes conserven els extrems relatius) i $\partial_\theta^2 \log \mathscr{L}(x_1,\dots,x_n;\hat\theta) < 0$.

$$\begin{aligned}
\mathscr{L}(x_1,\dots,x_n;\theta) &= \prod_{i=1}^n \left((1-\theta)\mathbb{1}_{[-\frac12,0]}(x) + (1+\theta)\mathbb{1}_{(0,\frac12]}(x)\right)\\
&= (1-\theta)^{\sum_i \mathbb{1}_{[-\frac12,0]}(x_i)} + (1+\theta)^{\sum_i \mathbb{1}_{[0,\frac12]}(x_i)}.\\
\log \mathscr{L}(x_1,\dots,x_n;\theta) &= \left(\sum_{i=1}^n \mathbb{1}_{[-\frac12,0]}(x_i)\right)\cdot \log(1-\theta) + \left(\sum_{i=1}^n \mathbb{1}_{[0,\frac12]}(x_i)\right)\cdot\log(1+\theta)\\
&= N_1\log(1-\theta) + N_2\log(1+\theta).\\
\partial_\theta \log \mathscr{L}(x_1,\dots,x_n;\theta) &= -\frac{N_1}{1-\theta} + \frac{N_2}{1+\theta} = \frac{N_2 - N_1 - \theta(N_1+N_2)}{1-\theta^2}\\
\partial_\theta \log \mathscr{L}(x_1,\dots,x_n;\hat\theta) = 0 &\iff \frac{N_2 - N_1 - \theta(N_1+N_2)}{1-\theta^2} = 0 \iff \hat\theta = \frac{N_2-N_1}{N_1+N_2} = \frac{N_2-N_1}{n}.\\
\partial_\theta^2 \log \mathscr{L}(x_1,\dots,x_n;\hat\theta) &= -\frac{N_1}{(1-\hat\theta)^2} - \frac{N_2}{(1+\hat\theta)^2} < 0.
\end{aligned}$$

Acabem de demostrar que $\hat\theta = \frac{N_2-N_1}{n}$.

2. Per trobar l'estimador pel mètode dels moments, recordem que $\hat m_j \xrightarrow[n\to\infty]{\text{qs}} m_j$; per tant, calcularem $m_1 = E_\theta(x)$:

$$E_\theta(x) = \int_{-\infty}^{+\infty} x f_\theta(x)\,dx = \int_{-\frac12}^{0} x(1-\theta)\,dx + \int_0^{\frac12} x(1+\theta)\,dx = (1-\theta)\cdot\left(-\frac18\right) + (1+\theta)\cdot\frac18 = \frac{\theta}{4}.$$

Com $E_\theta(X) = \bar X_n$, $\theta = 4\bar X_n$ i queda $\tilde\theta = 4\bar X_n$.

3. Ara, com el model és regular, tenim que $I_F(\theta) = -E_\theta(\partial_\theta^2 \ln \mathscr{L}(x,\theta))$:

$$-E_\theta\left(-\frac{N_1}{(1-\theta)^2} - \frac{N_2}{(1+\theta)^2}\right) = E_\theta\left(\frac{N_1}{(1-\theta)^2} + \frac{N_2}{(1+\theta)^2}\right) = \frac{1}{(1-\theta)^2}E_\theta(N_1) + \frac{1}{(1+\theta)^2}E_\theta(N_2).$$

Tenim dues maneres de calcular aquestes esperances. La primera és purament calculística:

$$E_\theta(N_1) = \sum_{i=1}^n E_\theta(\mathbb{1}_{[-\frac12,0]}(x_i)) = \sum_{i=1}^n P\left(-\frac12 \le x_i \le 0\right) = \sum_{i=1}^n \int_{-\frac12}^0 (1-\theta)\,dx = \frac{n(1-\theta)}{2}.$$

Anàlogament podem trobar que $E_\theta(N_2) = \frac{n(1+\theta)}{2}$. L'altra forma que tenim és adonar-nos que:

$$N_1 = \sum_{i=1}^n \mathbb{1}_{[-\frac12,0]}(x_i) \sim B\left(n, \frac{1-\theta}{2}\right),\ N_2 = \sum_{i=1}^n \mathbb{1}_{(0,\frac12]}(x_i) \sim B\left(n, \frac{1+\theta}{2}\right).$$

Pel que $E_\theta(N_1) = n\cdot\frac{1-\theta}{2}$ i $E_\theta(N_2) = n\cdot\frac{1+\theta}{2}$. Aleshores,

$$I_F(\theta) = \frac{n}{(1-\theta)^2}\cdot\frac{1-\theta}{2} + \frac{n}{(1+\theta)^2}\cdot\frac{1+\theta}{2} = \frac{n(1+\theta)+n(1-\theta)}{2(1+\theta)(1-\theta)} = \frac{n}{(1+\theta)(1-\theta)}. \tag{2.1}$$

4. Finalment, veure que ni $\tilde{\theta}$ ni tampoc $\hat{\theta}$ tenen biaix (i.e. $E_\theta(\tilde{\theta}) = E_\theta(\hat{\theta}) = \theta$):

$$E_\theta(\tilde{\theta}) = 4E_\theta(\bar{X}_n) = 4 \cdot \frac{\theta}{4} = \theta, \quad E_\theta(\hat{\theta}) = \frac{1}{n}\left(E_\theta(N_2) - E_\theta(N_1)\right) = \frac{1}{n}\left(n \cdot \frac{1+\theta}{2} - n \cdot \frac{1-\theta}{2}\right).$$

Pel que ara per ara cap dels dos és preferible. Haurem de calcular el segon moment (la variància) per a cadascun dels estimadors:

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathrm{Var}_\theta\left(\frac{2}{n}N_2 - 1\right) = \frac{4}{n^2}\mathrm{Var}_\theta(N_2) = \frac{4}{n^2}n\left(\frac{1+\theta}{2}\right)\left(1 - \frac{1+\theta}{2}\right) = \frac{1}{n}(1+\theta)(1-\theta).$$

Com que $\frac{1}{n}(1+\theta)(1-\theta)$ és la cota de Cramer-Rao (recordem (2.1)), $\hat{\theta}$ és un estimador eficient. Per calcular $\mathrm{Var}_\theta(\tilde{\theta})$ necessitem $E_\theta(X_1^2)$:

$$E_\theta(X_1^2) = \int_{-\frac{1}{2}}^{0} x^2(1-\theta)\,dx + \int_0^{\frac{1}{2}} x^2(1+\theta)\,dx = (1-\theta)\cdot\frac{1}{3}\cdot\frac{1}{8} + (1+\theta)\cdot\frac{1}{3}\cdot\frac{1}{8} = \frac{1}{12}.$$

Pel que:

$$\begin{aligned}\mathrm{Var}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta(4\bar{X}_n) = \frac{16}{n}\mathrm{Var}_\theta(X_1) &= \frac{16}{n}\left(E_\theta(X_1^2) - (E_\theta(X_1))^2\right) \\ &= \frac{16}{n}\left(\frac{1}{12} - \left(\frac{\theta}{4}\right)^2\right) = \frac{4 - 3\theta^2}{3n} = \frac{\frac{2}{3} - \theta^2}{n}.\end{aligned}$$

L'únic estimador eficient dels dos és $\hat{\theta}$. ■

Exercici 2.7. *Considerem una observació d'una variable amb densitat $f(x) = e^{-(x-\theta)}\mathbb{1}_{(\theta,+\infty)}(x)$, on $\theta > 0$. Trobeu una funció pivotant i un interval de confiança per a θ de nivell γ.*

Demostració. Definim $T = X - \theta$ i estudiem la funció de distribució:

$$F_T(x) = P_\theta(X - \theta \leq x) = P_\theta(X \leq x + \theta) \overset{9}{=} \int_\theta^{x+\theta} e^{-(y-\theta)}\,dy = e^\theta(e^{-\theta} - e^{-x-\theta}) = 1 - e^{-x}.$$

Ara, de probabilitats recordem que si Y és una variable aleatòria amb funció de distribució $F_Y(y)$ contínua, aleshores $F_Y(Y) \sim U(0,1)$. D'aquí, com $F_X(x) = 1 - e^{\theta - x}$, aleshores $1 - e^{\theta - X} \sim U(0,1)$.

- Se segueix que $-\log(1 - e^{\theta - X}) \sim \mathrm{Exp}(1)$, però és poc útil.
- D'altra banda, si $1 - e^{\theta - X} \sim U(0,1)$ aleshores $e^{\theta - X} \sim U(0,1)$. Ara sí, $-\log(e^{\theta - X}) = X - \theta$ i $-\log(e^{\theta - X}) \sim \mathrm{Exp}(1)$, pel que $T = X - \theta \sim \mathrm{Exp}(1)$. Per tant, com T no depèn del paràmetre $\pi(x,\theta) = x - \theta$ descriu una funció pivotant i:

$$P_\theta(a \leq X - \theta \leq b) = \gamma \iff \int_{-\infty}^{+\infty} f_{X-\theta}(y)\,dy = \int_a^b e^{-y}\,dy = e^{-a} - e^{-b} = \gamma.$$

[9] Tenim una funció de densitat clarament contínua, per tant, integrable i d'aquí en podem extreure la funció de distribució. Recordem que $f(x) = e^{-(x-\theta)}\mathbb{1}_{(\theta,+\infty)}(x)$, pel que hem d'integrar en $(\theta, x+\theta)$ (ja que busquem la probabilitat $P_\theta(X \leq x + \theta)$).

Es tracta de buscar a i b, però, com ja hem comentat, la solució no és única. Tenim diverses maneres: ens podem fixar en la cua (la cua val $1-\gamma$) o bé amb simetria. Utilitzem la segona i busquem $P_\theta(X-\theta \leq a)$ i $P_\theta(X-\theta \leq b)$:

$$P_\theta(X-\theta \leq a) = \frac{1-\gamma}{2} \iff 1-e^{-a} = \frac{1-\gamma}{2} \iff a = -\ln\left(\frac{1+\gamma}{2}\right),$$
$$P_\theta(X-\theta \leq b) = \frac{1+\gamma}{2} \iff 1-e^{-b} = \frac{1+\gamma}{2} \iff b = -\ln\left(\frac{1-\gamma}{2}\right).$$

Per últim, $P_\theta(a \leq X+\theta \leq b) = P_\theta(X-b \leq \theta \leq X-a)$, pel que l'interval de confiança haurà de ser $[X-b, X-a] = \left[X+\ln\left(\frac{1-\gamma}{2}\right), X+\ln\left(\frac{1+\gamma}{2}\right)\right]$. ■

Exercici 2.8. *El nombre de virus per centímetre cúbic en una certa dissolució segueix una distribució de Poisson de paràmetre λ. A partir d'una mostra de mida $n = 283$, s'obté:*

virus	0	1	2	3	4	5
freqüència	47	121	82	24	7	2

Trobeu un interval de confiança aproximat per a λ al 90%.

Demostració. Primer aplicarem la desigualtat de Txebitxev i després farem el mateix però per mètodes asimptòtics. Sigui $T = \bar{X}_n$ un estimador sense biaix de λ (ho és perquè $T \sim \text{Poisson}(\lambda)$ i per tant $E_\lambda(T) = \lambda$).

$$P_\lambda\left(\frac{|T-\lambda|}{\sqrt{\text{Var}_\lambda(T)}} \geq \delta\right) \leq \frac{1}{\delta^2} \iff P_\lambda\left(\frac{|T-\lambda|}{\sqrt{\text{Var}_\lambda(T)}} < \delta\right)$$
$$\overset{10}{=} P_\lambda\left(T-\delta\sqrt{\text{Var}_\lambda(T)} < T+\lambda < \delta\sqrt{\text{Var}_\lambda(T)}\right) \geq 1-\frac{1}{\delta^2}.$$

Definint $\alpha = 1-\frac{1}{\delta^2}$, obtenim que $\delta = \sqrt{\frac{1}{1-\alpha}}$ i per a un $\alpha = 0.9$, necessitem $\delta = \sqrt{10}$. Així doncs,

$$\text{Var}_\lambda(\bar{X}_n) = \frac{1}{n}\text{Var}_\lambda(X_1) = \frac{\lambda}{n}.$$

Tot plegat, l'interval de confiança seria el següent:

$$\left[\bar{X}_n - \delta\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + \delta\sqrt{\frac{\bar{X}_n}{n}}\right].$$

No hem acabat, ja que també podem procedir per mètodes asimptòtics. Pel teorema del límit central:

$$\frac{\sum_{i=1}^n x_i - n\lambda}{\sqrt{n\lambda}} \xrightarrow[n\to\infty]{\mathcal{L}} N(0,1) \iff \frac{\bar{X}_n-\lambda}{\sqrt{\frac{\lambda}{n}}} \xrightarrow[n\to\infty]{\mathcal{L}} N(0,1)$$

[10] Ens hem estalviat uns quants passos intermitjos, purament calculístics. Si ho volguéssim escriure fil per randa, seria: $P_\lambda\left(|T-\lambda| < \delta\sqrt{\text{Var}_\lambda(T)}\right) = P_\lambda\left(-\delta\sqrt{\text{Var}_\lambda(T)} - T < -\lambda < \delta\sqrt{\text{Var}_\lambda(T)} - T\right)$, i d'aquí obtenim $P_\lambda\left(T-\delta\sqrt{\text{Var}_\lambda(T)} < T+\lambda < \delta\sqrt{\text{Var}_\lambda(T)}\right)$, com volíem.

Quan tenim distribucions simètriques, solem buscar a, b amb el mètode dels simètrics. Prenem $Z = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \sim N(0,1)$ i busquem a, b tals que $P(a \leq Z \leq b) = 0.9$: essent $P(Z \leq b) = \frac{1+0.9}{2}$ i $P(Z \leq a) = \frac{1-0.9}{2}$, respectivament, ens queda que $a = -1.64$ i $b = 1.64$.

$$-1.64 \leq \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \leq 1.64 \iff \bar{X}_n - 1.64\sqrt{\frac{\lambda}{n}} \leq \lambda \leq \bar{X}_n + 1.64\sqrt{\frac{\lambda}{n}}. \qquad \blacksquare$$

Exercici 2.9. *Dos proveïdors A, B empaqueten un material en capses. Els pesos de les capses d'A presenten una distribució $N(\mu_1, 0.07^2)$ i les de B una $N(\mu_2, 0.04^2)$. Una mostra de 100 caixes d'A d'una mitjana de 0.99 kg i una mostra de 300 caixes de B d'una mitjana de 1.01 kg. Determineu un interval de confiança per a $\mu_1 - \mu_2$ al 95%.*

Demostració. Aquest exercici és immediat ja que ens demanen un interval amb condicions de normalitat per a una mostra, en particular la diferència de mitjanes amb variàncies conegudes. Per un corol·lari sabem que $\bar{X}_A \sim N(\mu_A, \frac{\sigma_A^2}{n_A}) = N(\mu_1, \frac{0.07^2}{n_A})$ i $\bar{X}_B \sim N(\mu_B, \frac{\sigma_B^2}{n_B}) = N(\mu_2, \frac{0.04^2}{n_B})$. Aleshores, $\bar{X}_A - \bar{X}_B \sim N(\mu_1 - \mu_2, \frac{0.07^2}{n_A} + \frac{0.04^2}{n_B})$ i podem definir:

$$Z \sim (X_A, X_B, \mu_1 - \mu_2) = \frac{\bar{X}_A - \bar{X}_B - (\mu_1 - \mu_2)}{\sqrt{\frac{0.07^2}{n_A} + \frac{0.04^2}{n_B}}} \sim N(0,1).$$

A partir d'aquí ja tenim la funció pivotant i el problema és trivial, usant que $\alpha = 0.95$ podem trobar un real positiu ζ_α tal que $Z([-\zeta_\alpha, \zeta_\alpha]) = \alpha$ (usant el mètode dels simètrics, surt $\zeta_\alpha = 1.96$) i, aleshores, l'interval de confiança per a $\mu_1 - \mu_2$ és el següent:

$$S(x) = \left[\bar{X}_A - \bar{X}_B - \zeta_\alpha\sqrt{\frac{\sigma_A^2}{n_1} + \frac{\sigma_B^2}{n_2}}, \bar{X}_A - \bar{X}_B + \zeta_\alpha\sqrt{\frac{\sigma_A^2}{n_1} + \frac{\sigma_B^2}{n_2}}\right].$$

Com ho coneixem tot, simplement substituïm i ja hem acabat. ■

3

Llista 3

Exercici 3.1. *Amb una observació d'una certa variable aleatòria X volem contrastar les hipòtesis H_0, H_1, X té una densitat $f_0(x) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}x^2\}$ i X té una densitat $f_0(x) = \frac{1}{\sqrt{2\pi}}\exp\{-\sqrt{2}|x|\}$, respectivament. Estudieu pels diferents valors $\alpha \in (0,1)$ quina és la regió d'acceptació d'un test UMP.*

Demostració. Considerem $\{\Omega, \mathscr{F}, \{P_\theta \mid \theta \in \Theta\}\}$, $H_0 : \theta = \theta_0$ i $H_1 : \theta = \theta_1$, $\theta_0 \neq \theta_1$. Considerem $\alpha \in (0,1)$. A_0 és una regió $A_0 = \{x \in \Omega \mid \mathscr{L}(x, \theta_1) \leq c_\alpha \cdot \mathscr{L}(x, \theta_0)\}$, $P_{\theta_0}(A_0^c) = \alpha$ implica que A_0 és test

$\text{ump}(\alpha)$. Assumint $n = 1$ la funció de versemblança coincideix amb la densitat, i suposant que $\mathscr{L}(x, \theta_0) \neq 0$ per a tot x, podem posar:

$$A_0 = \left\{ x \in \Omega \mid \frac{L(x,\theta_1)}{L(x,\theta_0)} \le c_\alpha \right\} = \left\{ x \in \Omega \mid \frac{\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}x^2\}}{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}} \le c_\alpha \right\} = \left\{ \sqrt{\pi}\exp\left\{\frac{1}{2}x^2 - \sqrt{2}x\right\} \le c_\alpha \right\}$$
$$= \left\{ x \in \Omega \mid \frac{1}{2}x^2 - \sqrt{2}x \le \log\frac{c_\alpha}{\sqrt{\pi}} \right\}.$$

Pel que si definim $k_\alpha := \log\frac{c_\alpha}{\sqrt{\pi}}$:

$$P_{\theta_0}(\frac{1}{2}x^2 - \sqrt{2}|x| > k_\alpha) = \alpha \implies g(x) = \frac{1}{2}x^2 - \sqrt{2}|x| \xrightarrow{x>0} x\left(\frac{1}{2}x - \sqrt{2}\right).$$

Pel que hem de distingir casos en funció d'A_0:

1. Si $A_0 = \emptyset$, $k_\alpha \le -1$. En aquest cas, $P_{\theta_0}(A_0^c) = 1$.
2. Si A_0 és unió de dos intervals, $-1 < k_\alpha < 0$.
3. Si A_0 és un interval, $k_\alpha \ge 0$:

$$P_{\theta_0}\left(\frac{1}{2}x^2 - \sqrt{2}|x| \ge 0\right) = P_{\theta_0}(x \le -2\sqrt{2}) + P_{\theta_0}(X \ge 2\sqrt{2}) = 2P_{\theta_0}(x \ge 2\sqrt{2})$$
$$= 2\int_{2\sqrt{2}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx = \alpha = 0.0048.$$

Hem usat que $P(X \ge 2\sqrt{2}) = \frac{0.0048}{2}$. Llavors, quan $\alpha \in (0.0048, 1)$ estem en el cas que A_0 són dos intervals i quan $\alpha \in (0, 0.0048)$, en el d'un interval. ■

Exercici 3.2. *Sabem que una certa variable aleatòria X té una distribució exponencial de paràmetre $\lambda = 1$ o bé $\lambda = 2$. Volem contrastar amb una mostra de mida 1 les hipòtesis:*

$$H_0 : \lambda = 1 \textit{ contra } H_1 : \lambda = 2.$$

1. *Calculeu les probabilitats dels errors de primera i segona espècie quan la regió d'acceptació és $A_0 = \{x \le 1\}$.*
2. *El mateix per a $A_0 = \{x \ge 0.07\}$.*
3. *Quin test triaríeu dels dos anteriors?*
4. *Trobeu el valor de k perquè la suma dels errors sigui mínima quan la regió d'acceptació és $A_0 = \{x \ge k\}$.*

Demostració. Tenim una variable aleatòria X amb llei exponencial de paràmetre $\lambda > 0$. Recordem que la seva funció de distribució és:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{si } x > 0; \\ 0, & \text{si } x \le 0. \end{cases} \implies f(x) = \lambda e^{-\lambda x}\mathbb{1}_{[0,+\infty)}(x).$$

Volem contrastar la hipòtesi $H_0 : \lambda = 1$ contra la hipòtesi $H_1 : \lambda = 2$. Denotarem per F_0 i f_0 les funcions de distribució i densitat de l'exponencial de paràmetre 1 i per F_1 i f_1 les mateixes funcions per a l'exponencial de paràmetre 2.

 Mario Vilar, 2024

1. Considerem la regió d'acceptació $A_0 = \{X \leq 1\}$. Això vol dir que acceptem H_0 si x, l'observació d'X, pertany a A_0 i acceptem H_1 si x no pertany a A_0, és a dir, $x \in A_0^c$. La regió A_0^c s'anomena *regió crítica*. L'error de primera espècie és la probabilitat de rebutjar H_0 quan és certa; és a dir, la probabilitat que X prengui un valor a A_0^c quan H_0 és certa. Per tant,

$$\alpha_1(H_0) = P_{H_0}(A_0^c) = P_{H_0}(X > 1) = 1 - P_{H_0}(X \leq 1) = 1 - F_0(1) = \frac{1}{e} \approx 0.368.$$

La probabilitat de segona espècie és la probabilitat d'acceptar H_0 quan és falsa. Per tant, la probabilitat que X prengui fora de la regió d'acceptació quan la llei d'X ve donada per H_1. Aleshores, tenim:

$$\alpha_2(H_1) = P_{H_1}(A_0) = P_{H_1}(X \leq 1) = F(1) = 1 - e^{-2} \approx 0.865.$$

2. Ara tenim la regió d'acceptació $A_0 = \{X \geq 0.07\}$, pel que fem exactament el mateix que a l'apartat anterior:

$$\alpha_1(H_0) = P_{H_0}(A_0^c) = P_{H_0}(X < 0.07) = F_0(0.07) = 1 - e^{0.07} \approx 0.068.$$
$$\alpha_2(H_1) = P_{H_1}(A_0) = P_{H_1}(X \geq 0.07) = 1 - F_1(0.07) = e^{-2 \cdot 0.07} \approx 0.869.$$

3. Els dos testos tenen l'error de segona espècie semblant, però el segon test té un error de primera espècie molt més petit. Per tant, en aquest sentit, és millor.
4. La potència es defineix com el complementari de l'error de segona espècie, és a dir, $P_{H_1}(A_0^c)$. Fent els mateixos càlculs:

$$P_{H_1}(A_0^c) = P_{H_1}(X > 1) = 1 - F_1(1) = \frac{1}{e^2} \approx 0.135.$$
$$P_{H_1}(A_0^c) = P_{H_1}(X < 0.07) = F_1(0.07) = 1 - e^{-2 \cdot 0.07} = 0.869.$$

Els dos testos tenen una potència semblant. Notem que la potència i l'error de segona espècie sumen 1.
5. Sigui $f(k)$ la suma dels errors de primera i segona espècie, és a dir, $f(k) = \alpha_1(\theta_0) + \alpha_2(\theta_1)$. Si la regió d'acceptació és $A_0 = \{X \geq k\}$:

$$\alpha_1(\theta_0) = P_{H_0}(A_0^c) = P_{H_0}(X < k) = F_0(k) = 1 - e^{-k}.$$
$$\alpha_2(\theta_1) = P_{H_1}(A_0) = P_{H_1}(X \geq k) = 1 - F_1(k) = e^{-2k}.$$

La suma dels errors és, doncs, $f(k) = 1 - e^{-k} + e^{-2k}$ per a $k \in (0, +\infty)$. Busquem la k que minimitza la funció f, derivant i igualant la derivada a zero. Fent els càlculs:

$$f'(k) = e^{-k} - 2e^{-2k} = 0 \iff k = -\ln\frac{1}{2} = \ln 2,\ f''(\ln 2) > 0 \implies k = \ln 2 \text{ és mínim.} \qquad \blacksquare$$

Exercici 3.3. *Sigui X una variable aleatòria amb densitat*

$$f(x) = 2px \exp\left\{-px^2\right\} \mathbb{1}_{(0,\infty)}(x),$$

on $p > 0$. Fixem dos nombres reals $0 < p_0 < p_1$ i considerem una mostra de mida n. Trobeu un test UMP per a contrastar

$$H_0 : p = p_0 \quad \textit{contra} \quad H_1 : p = p_1.$$

Apliqueu-ho a $n = 6$, $p_0 = 1$, $p_1 = 1.5$, $\alpha = 0.05$.

Demostració. Sigui $L(x, p_i)$ la funció de versemblança per a p_i, $i \in \{0, 1\}$:

$$L(x, p_i) = 2^n p_i^n \prod_{j=1}^{n} x_j \exp\left\{-p_i \sum_{j=1}^{n} x_j^2\right\}$$

Per tant, la regió d'acceptació serà A_0 que quedarà definida de la manera següent:

$$A_0 = \left\{\frac{L(x, p_1)}{L(x, p_0)} \le c_\alpha\right\} = \left\{\frac{p_1^n}{p_0^n} \exp\left\{-(p_1 - p_0)\sum_{i=1}^{n} x_i^2\right\} \le c_\alpha\right\} = \left\{\sum_{i=1}^{n} x_i^2 \ge \frac{\log\left(c_\alpha \cdot \frac{p_0^n}{p_1^n}\right)}{p_0 - p_1}\right\}$$

I definirem $k'_\alpha = \frac{\log\left(c_\alpha \cdot \frac{p_0^n}{p_1^n}\right)}{p_0 - p_1}$. Volem $P_{p_0}(A_0^c) = \alpha$, és a dir, $P_{p_0}(A_0^c) = P_{p_0}(\sum_{i=1}^n x_i^2 < k'_\alpha)$. Ara, bàsicament veurem que $Y = X^2 \sim \text{Exp}(p_0)$ i per tant $\sum_i Y_i \sim \text{Gamma}(n, p_0)$. Definint Y d'aquesta manera:

$$\left.\begin{array}{cccc} g: & (0, +\infty) & \longrightarrow & (0, +\infty) \\ & x & \longmapsto & x^2 \\ & \sqrt{y} & \longleftarrow & y \end{array}\right\} f_Y(y) = \frac{1}{2\sqrt{y}} 2p\sqrt{y}e^{-py} = pe^{-py}, \ y > 0.$$

Com volíem, $Y = X^2 \sim \text{Exp}(p_0)$ i $\sum_i Y_i \sim \text{Gamma}(n, p_0)$. Aleshores:

$$P_{p_0}(A_0^c) = P_{p_0}\left(\sum_{i=1}^{n} X_i^2 < k'_\alpha\right) = P(\text{Gamma}(n, p_0) < k'_\alpha) = \alpha.$$

Arribats a aquest punt simplement caldria substituir els valors coneguts n, p_0. Per a la potència, seria el mateix però canviant p_0 per p_1:

$$P_1(A_0^c) = P_1\left(\sum_{i=1}^{n} x_i^2 < k'_\alpha\right) = (\text{Gamma}(6, 1.5) < k'_\alpha).$$ ■

Exercici 3.4. *Sobre la mitjana d'una llei normal $\mathcal{G}(m, 1)$ es fan les hipòtesis $H_0 : m = -1$ contra $H_1 : m = m_0$, on $m_0 > -1$. Es van contrastar aquestes hipòtesis mitjançant un test de Neyman-Pearson aplicat a una mostra de mida $n = 100$. El test tenia risc de primer tipus igual a 0.0287 i potència 0.9032. Calculeu m_0.*

Demostració. Per Neyman-Pearson, sabem que $P_0(A_0^c) = 0.0287$ i $P_1(A_0^c) = 0.9032$. Anomenarem A_0 a la regió d'acceptació, que ve donada per $A_0 = \{L(x_1, m_1) \le c_\alpha \cdot L(x_1, -1)\}$.

$$A_0 = \left\{\frac{L(x_1, m_1)}{L(x_1, -1)} \le c_\alpha\right\} = \left\{\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^{100} \exp\left\{-\frac{1}{2}\sum_{i=1}^{100}(x_i - m_1)^2\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^{100} \exp\left\{-\frac{1}{2}\sum_{i=1}^{100}(x_i + 1)^2\right\}} \le c_\alpha\right\}.$$

Com que:

$$-\frac{1}{2}\sum_{i=1}^{100}\left[(x_i^2-2x_im_1+m_1^2)-(x_i^2+2x_i+1)\right]=-\frac{1}{2}\sum_{i=1}^{100}\left[-2x_i(m_1+1)+m_1^2-1\right]$$
$$=\left(\frac{1}{2}\sum_{i=1}^{100}2x_i(m_1+1)\right)-\frac{1}{2}\left(100m_1^2-100\right).$$

Ens queda:

$$\left\{\exp\left\{\left(\frac{1}{2}\sum_{i=1}^{100}2x_i(m_1+1)\right)-\frac{1}{2}\left(100m_1^2-100\right)\right\}\le c_\alpha\right\}=\left\{\sum_{i=1}^{100}(x_i(m_1+1))-50\left(m_1^2-1\right)\le\log(c_\alpha)\right\}$$
$$\left\{\sum_{i=1}^{100}x_i\le\frac{\log(c_\alpha)}{m_1+1}+50\,(m_1-1)\right\}=\left\{\sum_{i=1}^{100}x_i\le k_\alpha\right\}$$

Pel que $P_0(A_0^c)=P_0(\sum_{i=1}^{100}X_i>k_\alpha)$. No tenim absolutament cap dada de la distribució de les X_i *per se*. Se'ns diu que es fan les hipòtesis sobre la mitjana d'una llei normal $\mathcal{G}(m,1)$; per tant, pel teorema del límit central $Y_0=\frac{(\sum_{i=1}^{n}X_i)+100}{\sqrt{100}}\sim N(0,1)$. Així:

$$P_0\left(Y_0>\frac{k_\alpha+100}{\sqrt{100}}\right)=0.0287\iff P_0\left(Y_0\le\frac{k_\alpha+100}{\sqrt{100}}\right)=0.9713\xrightarrow{\text{taula}}\frac{k_\alpha+100}{\sqrt{100}}=1.9\implies k_\alpha=-81.$$

Ara que hem trobat k_α, podem calcular $P_1(\sum_i X_i>-81)$ definint $Z=\frac{(\sum_i X_i)-100m_1}{\sqrt{100}}\sim N(0,1)$:

$$P_1(\sum_i X_i>-81)=P_1\left(\frac{(\sum_i X_i)-100m_1}{\sqrt{100}}>\frac{-81-100m_1}{\sqrt{100}}\right)=P_1\left(Z>\frac{-81-100m_1}{\sqrt{100}}\right)=0.9032.$$

Per tant, usant una taula per a distribucions normals o R, $\frac{-81-100m_1}{\sqrt{100}}=-1.3$ i $m_1=-0.64$. ■

Exercici 3.5. *Fem una observació d'una variable aleatòria amb funció de densitat* $f(x,\theta)=\frac{2}{\theta^2}(\theta-x)\mathbb{1}_{(0,\theta)}(x)$, *on* $\theta>0$. *Volem contrastar les hipòtesis* $H_0:\theta=\theta_0$ *contra* $H_1:\theta=\theta_1$, *on* $\theta_0<\theta_1$.

1. *Busqueu un test UMP de nivell* $\alpha=0.01$. *Calculeu la potència d'aquest test.*
2. *Discutiu si el test anterior pot servir per a contrastar* $H_0:\theta=\theta_0$ *contra* $H_1:\theta>\theta_0$. *Quines propietats es conserven?*

Demostració.

1. La primera part torna a ser molt rutinària. Definim $A_0=\{x>0\mid L(x,\theta_1)\le c_\alpha L(x,\theta_0)\}$ com a regió d'acceptació. En aquest cas, el quocient $\frac{L(x,\theta_1)}{L(x,\theta_0)}$ manca de sentit en certs casos:

$$\frac{L(x,\theta_1)}{L(x,\theta_0)}=\begin{cases}\frac{\theta_0^2(\theta_1-x)}{\theta_1^2(\theta_0-x)}, & \text{si } x<\theta_0;\\ \infty, & \text{si } \theta_0\le x<\theta_1;\\ \frac{0}{0}, & \text{si } x\ge\theta_1.\end{cases}\tag{3.1}$$

En el cas senzill de contrastar hipòtesis simples, el lema de Neyman-Pearson ens permet de donar una regió d'acceptació A_0 d'un test UMP al nivell α, sempre que existeixi una constant c_α que satisfaci $P_{\theta_0}(A_0^c) = \alpha$. Pel que hem vist a (3.1), aquest quocient només està definit per a $x < \theta_0 < \theta_1$, pel que $c_\alpha \in (0, +\infty)$; per tant, $P_{\theta_0}(X > \theta_0) = P_{\theta_0}(X > \theta_1) = P_{\theta_1}(X > \theta_1) = 0$. Així doncs, la regió d'acceptació queda de la següent manera:

$$A_0 = \left\{ \frac{L(x, \theta_1)}{L(x, \theta_0)} \leq c_\alpha \right\} = \left\{ \frac{\theta_1 - x}{\theta_0 - x} \leq \frac{\theta_1^2}{\theta_0^2} c_\alpha \right\} = \left\{ x \in (0, \theta_0) \;\middle|\; \left(\frac{\theta_1^2}{\theta_0^2} c_\alpha - 1 \right) x \leq \frac{\theta_1^2}{\theta_0} c_\alpha - \theta_1 \right\}.$$

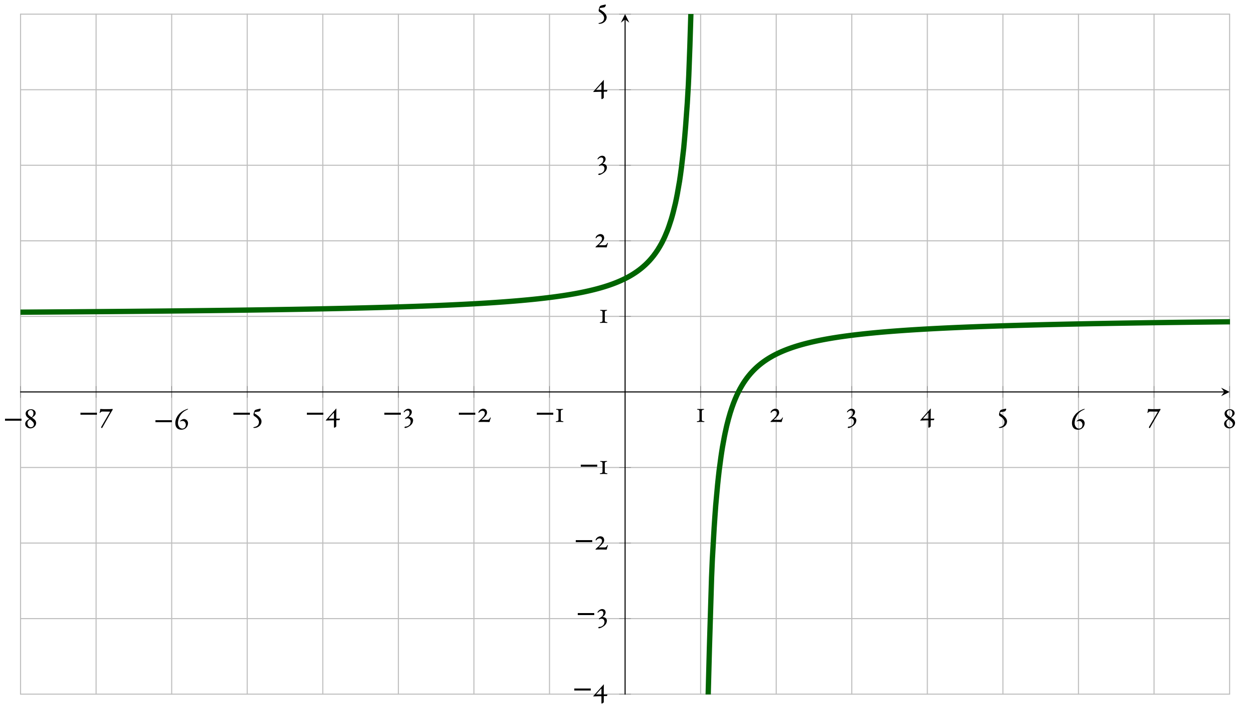

Figura 1: Gràfica de la situació en el cas especial $\theta_1 = 1.5$ i $\theta_0 = 1$.

Així doncs, $\{X \leq k'_\alpha\}$ per a cert k'_α, pel que $P_{\theta_0}(A_0^c)$ és $P_{\theta_0}(X > k'_\alpha)$:

$$P_{\theta_0}(X > k'_\alpha) = \begin{cases} 0, & \text{si } k'_\alpha \geq \theta_0; \\ \int_{k'_\alpha}^{\theta_0} \frac{2}{\theta_0^2}(\theta_0 - x)\, dx = \left(1 - \frac{k'_\alpha}{\theta_0}\right)^2 = \alpha, & \text{si } k'_\alpha < \theta_0. \end{cases}$$

Com $\left(1 - \frac{k'_\alpha}{\theta_0}\right)^2 = \alpha$, $1 - \frac{k'_\alpha}{\theta_0} = 0.1$ i per tant $k'_\alpha = \theta_0 \cdot 0.9$. La potència representa la probabilitat de rebutjar la hipòtesi nul·la quan és falsa, ja que el vertader valor del paràmetre és $\theta \in \Theta_1$. Per tant, volem $P_{\theta_1}(A_0^c)$:

$$P_{\theta_1}(A_0^c) = P_{\theta_1}(X > 0.9\theta_0) = \int_{0.9\theta_0}^{\theta_1} \frac{2}{\theta_1^2}(\theta_1 - x)\, dx = \frac{(\theta_1 - 0.9\theta_0)^2}{\theta_1^2}.$$

2. Hem vist $A_0 = \{x \leq k'_\alpha\}$. Com la regió no depèn del valor concret θ_1, A_0 és un test UMP per a $H_0 : \theta = \theta_0$ i $H_1 : \theta > \theta_0$. ■

Exercici 3.6. *Considerem una mostra de mida 1 d'una llei amb densitat* $f(x) = \frac{1}{2}\exp\{-|x-\theta|\}$, $x \in \mathbb{R}$, *on* $\theta \in \mathbb{R}$. *Volem contrastar les hipòtesis* $H_0 : \theta = \theta_0$ *contra* $H_1 : \theta = \theta_1$, *on* $\theta_0 < \theta_1$.

1. *Discutiu l'existència del test de Neyman-Pearson segons el valor del nivell de significació* α.
2. *Doneu la regió d'acceptació del test de Neyman-Pearson per a* $\theta_0 = 0$, $\theta_1 = 10$ *i* $\alpha = 0.01$.

Demostració. La regió d'acceptació ve donada per $A_0 = \{L(x,\theta_1) \le c_\alpha \cdot L(x,\theta_0)\}$ i $\frac{L(x,\theta_1)}{L(x,\theta_0)} = \exp\{|x-\theta_0| - |x-\theta_1|\}$ i com aquest quocient està ben definit $\forall x$:

$$A_0 = \{|x-\theta_0| - |x-\theta_1| \le \log(c_\alpha)\}, \quad |x-\theta_0| - |x-\theta_1| = \begin{cases} \theta_0 - \theta_1, & \text{si } x < \theta_0; \\ 2x - \theta_0 - \theta_1, & \text{si } \theta_0 \le x < \theta_1; \\ \theta_1 - \theta_0, & \text{si } x \ge \theta_1; \end{cases}$$

Ens ho podem imaginar una gràfica a trossos contínua i contínua en els punts de contacte, i k_α una funció constant que talla (o no) $|x-\theta_0| - |x-\theta_1|$. Hem de plantejar casos:

1. Talla $|x-\theta_0| - |x-\theta_1|$ entre $\theta_1 - \theta_0$ i $\theta_0 - \theta_1$. És a dir, $k_\alpha \in [\theta_0 - \theta_1, \theta_1 - \theta_0)$. Aleshores:
$$P_{\theta_0}(A_0^c) = P_{\theta_0}(|x-\theta_0| - |x-\theta_1| > k_\alpha) = P_{\theta_0}(x > k'_\alpha) = \int_{k'_\alpha}^{+\infty} \frac{1}{2} e^{-|x-\theta_0|}\, dx = \frac{1}{2}\int_{k'_\alpha}^{\infty} e^{-(x-\theta_0)}\, dx\,.$$
2. Si *no* talla $|x-\theta_0| - |x-\theta_1|$, és a dir, $k_\alpha \ge \theta_1 - \theta_0$ o bé $k_\alpha < \theta_0 - \theta_1$. En tals casos:
$$P_{\theta_0}(A_0^c) = P_{\theta_0}(|x-\theta_0| - |x-\theta_1| > k_\alpha) = P_{\theta_0}(\emptyset) = 0.$$

■

Exercici 3.7. *Sigui X una variable aleatòria amb densitat* $f(x) = 2\theta x \exp\{-\theta x^2\}\mathbb{1}_{(0,+\infty)}(x)$, *on* $\theta > 0$. *Fixem dos nombres reals* $0 < \theta_0 < \theta_1$, *i considerem una mostra de mida n. Trobeu un test* UMP *per a contrastar* $H_0 : p = p_0$ *contra* $H_1 : p = p_1$. *Apliqueu-ho a* $n = 6$, $p_0 = 1$, $p_1 = 1.5$ *i* $\alpha = 0.05$.

Demostració. El model estadístic paramètric és $(\Omega, \mathcal{A}, \{P_\theta \mid \theta > 0\})$, i $\Omega = (0,+\infty)^n$ no depèn del paràmetre. La funció de versemblança és:

$$\mathcal{L}(x,\theta) = 2^n\theta^n \left(\prod_{i=1}^n x_i\right) \cdot \exp\left\{-\theta \sum_i x_i^2\right\} > 0, \ \forall \theta > 0.$$

Per tant, $\mathcal{L}(x,\theta_0) > 0$ i el quocient $\frac{\mathcal{L}(x,\theta_1)}{\mathcal{L}(x,\theta_0)}$ està ben definit.

$$\frac{\mathcal{L}(x,\theta_1)}{\mathcal{L}(x,\theta_0)} = \frac{2^n\theta_1^n \left(\prod_{i=1}^n x_i\right) \cdot \exp\left\{-\theta_1 \sum_i x_i^2\right\}}{2^n\theta_0^n \left(\prod_{i=1}^n x_i\right) \cdot \exp\left\{-\theta_0 \sum_i x_i^2\right\}} = \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left\{(\theta_0 - \theta_1) \cdot \sum_i x_i^2\right\}.$$

La regió d'acceptació és $A_0 = \left\{\frac{\mathcal{L}(x,\theta_1)}{\mathcal{L}(x,\theta_0)} \le c_\alpha\right\}$:

$$\begin{aligned} A_0 &= \left\{\exp\{(\theta_0 - \theta_1) \cdot \sum_i x_i^2\} \le c_\alpha \cdot \frac{\theta_0^n}{\theta_1^n}\right\} \\ &= \left\{(\theta_0 - \theta_1) \cdot \sum_i x_i^2 \le \ln\left(c_\alpha \cdot \frac{\theta_0^n}{\theta_1^n}\right)\right\} \overset{\theta_0 < \theta_1}{=} \left\{\sum_i x_i^2 \le \frac{\ln\left(c_\alpha \cdot \frac{\theta_0^n}{\theta_1^n}\right)}{\theta_0 - \theta_1}\right\}. \end{aligned}$$

Així, si definim $k_\alpha = \frac{\ln\left(c_\alpha \cdot \frac{\theta_0^n}{\theta_1^n}\right)}{\theta_0 - \theta_1} \neq 0$, $A_0^c = \{\sum_i x_i^2 < k_\alpha\}$. Definim $Y_i = X_i^2$:

$$\begin{array}{rccc} g: & (0,+\infty) & \longrightarrow & (0,+\infty) \\ & x & \longmapsto & x^2 \\ & \sqrt{y} & \longleftarrow & y \end{array} \qquad \begin{array}{rccc} g': & (0,+\infty) & \longrightarrow & (0,+\infty) \\ & x & \longmapsto & 2x \\ & \frac{1}{2\sqrt{y}} & \longleftarrow & y \end{array}$$

Per tant, la densitat $f_\theta(y)$ queda com:

$$f_\theta(y) = 2\theta y \cdot \exp\{-\theta y\} \cdot \frac{1}{2\sqrt{y}} \cdot \mathbb{1}_{(0,+\infty)}(y) = \theta \exp\{-\theta y\} \cdot \mathbb{1}_{(0,+\infty)}(y).$$

Així doncs, queda clar que $Y_i \sim \text{Exp}(\theta)$ i $\sum_i Y_i \sim \text{Gamma}(n,\theta)$. Per tant, substituint:

$$P_{\theta_0}(A_0^c) = P_{\theta_0}\left(\sum_i Y_i \leq k_\alpha\right) = P(\text{Gamma}(6,1) \leq k_\alpha) = \alpha = 0.05,$$

obtenim que $k_\alpha = 2.61301$. Per tant, $A_0^c = \{\sum_i x_i \leq 2.61301\}$ i $\varphi = \mathbb{1}_{A_0^c}$ és el test UMP que buscàvem. ■

Exercici 3.8. *Disposem d'una mostra de mida n d'una llei exponencial de paràmetre λ. Fixem $0 < \lambda_0 < \lambda_1$. Construïu un test per a $H_0 : \lambda = \lambda_0$ contra $H_1 : \lambda = \lambda_1$. Calculeu explícitament la regió quan $\lambda_0 = 1$, $\lambda_1 = 2$ i $n = 5$, $\alpha = 0.05$.*

Demostració. La regió d'acceptació A_0 és $\left\{\frac{L(x,\lambda_1)}{L(x,\lambda_0)} \leq c_\alpha\right\}$, fent càlculs com sempre:

$$\left\{\exp\left\{-\lambda_1 \cdot \sum_i x_i - \lambda_0 \sum_i x_i\right\} \leq c_\alpha \cdot \left(\frac{\lambda_0}{\lambda_1}\right)^n\right\} = \left\{-(\lambda_1 - \lambda_0)\sum_{i=1}^n x_i \leq \log\left(c_\alpha \cdot \left(\frac{\lambda_0}{\lambda_1}\right)^n\right)\right\}.$$
$$= \left\{\sum_{i=1}^n x_i \geq \frac{\log\left(c_\alpha \cdot \left(\frac{\lambda_0}{\lambda_1}\right)^n\right)}{\lambda_0 - \lambda_1}\right\}$$

A l'última igualtat hem usat que $\lambda_1 > \lambda_0$, pel que $\lambda_0 - \lambda_1 < 0$ i s'inverteix la desigualtat. Les altres es poden justificar fàcilment. Prenem $k_\alpha = \frac{\log\left(c_\alpha \cdot \left(\frac{\lambda_0}{\lambda_1}\right)^n\right)}{\lambda_0 - \lambda_1}$. Recordem que com X_i són VAIID i $X_i \sim \text{Exp}(\lambda)$ per a tot i, aleshores $Z = \sum_i X_i \sim \text{Gamma}(\lambda, n)$.

$$P_{\lambda_0}(A_0^c) = P_{\lambda_0}(Z < k_\alpha) = P_{\lambda_0}(Z \leq k_\alpha) = \alpha.$$

Amb els valors concrets de λ_0, n (sabem explícitament la distribució que segueix Z, que serà una $\text{Gamma}(1,5)$) i α, amb una taula o R podríem trobar el valor de k_α i trobaríem la regió. El mateix procediment funcionaria per al cas de λ_1. ■

Exercici 3.9. *En base a una observació d'una variable aleatòria X volem fer un test UMP per a contrastar $H_0 : X \sim \mathcal{N}(0, \frac{1}{2})$ contra $H_1 : X \sim C(0,1)$, amb nivell $\alpha = 0.01$. Quina potència té aquest test?*

Demostració. Abans de començar, val a dir que la densitat d'una llei de Cauchy $C(0,1)$ és $f(x) = \frac{1}{\pi(1+x^2)}$ i la regió d'acceptació queda com:

$$A_0 = \left\{ \frac{\frac{1}{\pi(1+x^2)}}{\frac{1}{\sqrt{\pi}}e^{-x^2}} \leq c_\alpha \right\} = \left\{ \frac{e^{x^2}}{1+x^2} \leq \frac{\pi}{\sqrt{\pi}} c_\alpha \right\} = \left\{ \frac{e^{x^2}}{1+x^2} \leq k_\alpha \right\},\ g(x) = \frac{e^{x^2}}{1+x^2}.$$

Definida aquesta g, podem derivar i $g'(x) = \frac{2e^{x^2}x^3}{(1+x^2)^2}$ té un mínim en $x = 0$, i tal mínim és $(0,1)$. Això ens diu que el cas $k_\alpha < 1$ no té sentit, ja que no intersecaríem amb g. Aleshores, $\{|x| \leq k'_\alpha\}$:

$$P_0(A_0^c) = P(|x| \geq k'_\alpha) = 2 \cdot P(x > k'_\alpha) = 0.01 \implies P(x > k'_\alpha) = 0.005.$$

Així doncs:

$$P_0\left(\frac{x}{\sqrt{\frac{1}{2}}} \leq \frac{k'_\alpha}{\sqrt{\frac{1}{2}}} \right) = 0.995 \implies P(N(0,1) \leq k'_\alpha) = 0.995.$$

Amb una taula de la normal trobaríem el valor de k'_α. ■

Exemple 3.10. Sigui $L(x,\theta) = \frac{1}{\theta^n}\mathbb{1}_{(0,\theta)}(x_{(n)})$. Podem fer una partició dels reals positius en $I = (0,\theta_0) \cup [\theta_0,\theta_1) \cup [\theta_1,+\infty)$, i $x_{(n)} \in I$. Fem un test per a contrastar $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$, $\theta_0 < \theta_1$. La regió d'acceptació és:

$$A_0 = \{x > 0 \mid L(x,\theta_1) \leq c_\alpha L(x,\theta_0)\} = \left\{ x \in \mathbb{R}^+ \mid \frac{1}{\theta_1^n}\mathbb{1}_{(0,\theta_1)}(x_{(n)}) \leq c_\alpha \frac{1}{\theta_0^n}\mathbb{1}_{(0,\theta_0)}(x_{(n)}) \right\}$$

I ara aquí cal ser una mica curosos. Ara, com $\theta_0 < \theta_1$, $(0,\theta_0) \subset (0,\theta_1)$, i amb la partició que hem fet amb I:

$$\left\{ x_{(n)} \in (0,\theta_0) \mid \frac{\theta_0^n}{\theta_1^n} \leq c_\alpha \right\} \cup \underbrace{\left\{ x_{(n)} \in [\theta_0,\theta_1) \mid \frac{\theta_0^n}{\theta_1^n} \leq 0 \right\}}_{\emptyset} \cup \underbrace{\left\{ x_{(n)} \in [\theta_1,+\infty) \mid \frac{\theta_0^n}{\theta_1^n} \cdot 0 \leq c_\alpha \cdot 0 \right\}}_{\{x_{(n)} \in [\theta_1,+\infty)\}}.$$

Exercici 3.11. *Tenim una mostra de mida n d'una variable aleatòria X amb distribució uniforme en $\{1,\ldots,N\}$ on N és un natural desconegut. Donat un natural N_0, construïu un test per a contrastar $H_0 : N = N_0$ contra $H_1 : N \neq N_0$.*

Demostració. Tenim que $X \sim U\{1,\ldots,N\}$, pel que $P(X = k) = \frac{1}{N}$, per a tot $k \in \{1,\ldots,N\}$. Sigui $\Omega = \mathbb{N}^n$ l'espai mostral, la funció de versemblança $\mathscr{L}(x,N)$ és:

$$\mathscr{L}(x,N) = \prod_{i=1}^{n} \mathscr{L}(x_i,N) = \prod_{i=1}^{n} \frac{1}{N}\mathbb{1}_{1,\ldots,N}(x_i) = \frac{1}{N^n}\mathbb{1}_{1,\ldots,N}(x_{(n)}).$$

On $x_{(n)} = \max(x_1,\ldots,x_n)$. Per a contrastar $H_0 : N = N_0$ contra $H_1 : N \neq N_0$ amb nivell de significació $\alpha \in (0,1)$, usarem el test de la raó de versemblança. Primer observem que $\hat{N} = X_{(n)} = \max(X_1,\ldots,X_n)$ és un estimador de màxima versemblança, ja que la funció de versemblança és decreixent en N. Per tant:

$$\mathscr{L}(x_1,\ldots,x_n;X_{(n)}) = \sup_N \mathscr{L}(x_1,\ldots,x_n;N).$$

L'estadístic de la raó de versemblança $\lambda(x)$ queda determinat per:

$$\lambda(x) = \frac{\mathscr{L}(x, N_0)}{\sup_{N\in\mathbb{N}} \mathscr{L}(x, N)} = \frac{\mathscr{L}(x, N_0)}{\mathscr{L}(x, x_{(n)})} = \frac{\frac{1}{N_0^n}\mathbb{1}_{1,\dots,N_0}(x_{(n)})}{\frac{1}{x_{(n)}^n}\cdot 1} = \frac{x_{(n)}^n}{N_0^n}\mathbb{1}_{\{1,\dots,N_0\}}(x_{(n)}).$$

La regió d'acceptació és $A_0 = \{\lambda(x) \geq c_\alpha\}$ i volem $P_{H_0}(A_0^c) \leq \alpha$. La regió d'acceptació queda:

$$A_0 = \{\lambda(x) \geq c_\alpha\} = \{x_{(n)}^n \mathbb{1}_{\{1,\dots,N_0\}}(x_{(n)}) \geq c'_\alpha\} = \{x_{(n)} \leq N_0,\ x_{(n)} \geq \sqrt[n]{c'_\alpha}\} = \{\sqrt[n]{c'_\alpha} \leq x_{(n)} \leq N_0\}.$$

Per tant:

$$\begin{aligned} P_{H_0}(A_0^c) &= P_{H_0}\left((x_1,\dots,x_n)\in\mathbb{N}^n \mid \lambda(x) < c_\alpha\right) = P_{H_0}\left((x_1,\dots,x_n)\in\mathbb{N}^n \mid x_{(n)}^n \mathbb{1}_{\{1,\dots,N_0\}}(x_{(n)}) < c_\alpha\right) \\ &= P_{H_0}(x_{(n)} > N_0 \text{ o bé } x_{(n)} < \sqrt[n]{c'_\alpha}) = P_{H_0}(x_{(n)} < c'_\alpha) = P_{H_0}(x_1 < c'_\alpha)^n = \left(\frac{c'_\alpha - 1}{N_0}\right)^n \leq \alpha. \end{aligned}$$

Podríem ser molt més estrictes i considerar casos amb $c'_\alpha \in \mathbb{Z}$ i $c'_\alpha \notin \mathbb{Z}$, però no ho farem. Així doncs, cal prendre c'_α tal que:

$$c'_\alpha \leq N_0\sqrt[n]{\alpha} + 1. \text{ Per exemple, } c'_\alpha = N_0\sqrt[n]{\alpha} + 1.$$

Per tant, acceptarem la hipòtesi nul·la quan $x \in A_0$ i la rebutjarem en cas contrari. ■

Observació 3.12. Hem escrit $\alpha \geq P_{H_0}(A_0^c)$ en lloc d'$\alpha = P_{H_0}(A_0^c)$ ja que $\sqrt[n]{\alpha}$ no té per què ser un nombre enter i, si volem que el test sigui «suficientment bo», la probabilitat $P_{H_0}(A_0^c)$ haurà de ser, com a molt, α.

Exercici 3.13. *Un dau s'ha llançat 200 vegades. Els resultats són els següents:*

Punts	*1*	*2*	*3*	*4*	*5*	*6*
Freqüències	*35*	*42*	*30*	*29*	*40*	*24*

Decidiu, mitjançant un test de la raó de versemblança, si el dau és perfecte o no.

Demostració. Sigui $X = (X_1, \dots, X_{200})$ el vector aleatori associat a l'experiment del dau. Aleshores, X segueix una distribució multinomial de paràmetres $p = (p_1, \dots, p_6)$, $0 < p_k < 1$, on $p_k = P(X_i = k)$, per a tot $i \in \{1, \dots, 200\}$ i per a tot $k \in \{1, \dots, 6\}$. També és clar que $p_1 + \cdots + p_6 = 1$. Tornem a estar en un cas discret, pel que donada la mostra $x = (x_1, \dots, x_{200})$,

$$P(X = x) = P(X_1 = x_1, \dots, X_{200} = x_{200}) = \frac{200!}{n_1!\cdots n_6!}\cdot p_1^{n_1}\cdots p_6^{n_6},\ n_k = \sum_{i=1}^{n} \mathbb{1}_{\{x_i = k\}}.$$

Per a tot $k \in \{1, \dots, 6\}$ i $n_1 + \cdots + n_6 = n = 200$. El nostre objectiu és contrastar les hipòtesis següents:

$$H_0 : p_k = \frac{1}{6},\ \forall k \in \{1,\dots,6\} \text{ contra } H_1 : \exists k \in \{1,\dots,6\} \mid p_k \neq \frac{1}{6},$$

mitjançant un test de la raó de versemblança. Primer, calculem la funció de versemblança per a la nostra $x = (x_1, \dots, x_{100})$ i paràmetre $p = (p_1, \dots, p_6)$, $\mathscr{L}(x, p) = \mathscr{L}((n_1, \dots, n_6), p)$:

$$\lambda(x) = \frac{\mathscr{L}(x, p)}{\mathscr{L}(x, p_0)} = \frac{\frac{n!}{n_1! \cdots n_6!} \frac{1}{6^n}}{\frac{n!}{n_1! \cdots n_6!} \left(\frac{n_1}{n}\right)^{n_1} \cdots \left(\frac{n_6}{n}\right)^{n_6}} = \frac{n^n}{6^n \cdot n_1^{n_1} \cdots n_6^{n_6}}.$$

Com volem aplicar el test de la raó de versemblança, hem de buscar els estimadors de màxima versemblança:

$$\log \mathscr{L}(x, p) = \log\left(\frac{200!}{n_1! \cdots n_6!}\right) + \sum_{k=1}^{6} n_k \log(p_k).$$

Aprofitem per fer la següent observació.

Observació 3.14. Com $p_6 = 1 - (p_1 + \cdots + p_5)$, es pot pensar que p_6 «depèn» de $\{p_1, \dots, p_5\}$, pel que tenim 5 graus de llibertat i no 6. Anàlogament, $n_6 = n - (n_1 + \cdots + n_5)$.

Així doncs, volem trobar quins valors de p_k maximitzen la funció, és a dir, $\partial_{p_k} \log \mathscr{L}(x, p) = 0$.

$$\partial_{p_k} \log \mathscr{L}(x, p) = 0 \iff \frac{n_k}{p_k} - \frac{n_6}{p_6} = 0 \iff n_6 p_k = n_k p_6.$$

Per tant,

$$\sum_{k=1}^{5} = \sum_{k=1}^{5} n_k p_6 \implies n_6(1 - p_6) = (n - n_6) p_6 \implies p_6 = \frac{n_6}{n}.$$

Per tant, $\hat{p}_k = \frac{n_k}{n}$, $k \in \{1, \dots, 5\}$. En el cas unidimensional en tindríem prou amb comprovar que $\partial^2_{p_k} \log \mathscr{L}(x, p) < 0$, però amb el multidimensional necessitaríem la Hessiana (obviem, doncs, el pas de comprovar que és un màxim). Aleshores, $\hat{p} = (\hat{p}_1, \dots, \hat{p}_6)$ és un màxim. Així doncs, la raó de versemblança és:

$$\lambda(x) = \frac{\left(\frac{1}{6}\right)^{200}}{\hat{p}_1^{n_1} \cdots \hat{p}_6^{n_6}} = \left(\frac{1}{6}\right)^{200} \frac{n^n}{n_1^{n_1} \cdots n_6^{n_6}},$$

i la regió d'acceptació és:

$$A_0 = \left\{ x \mid \lambda(x) = \left(\frac{1}{6}\right)^{200} \frac{n^n}{n_1^{n_1} \cdots n_6^{n_6}} \geq c_\alpha \right\}, \ c_\alpha \leq 1.$$

Observem que l'espai de paràmetres $\Theta = \{(p_1, \dots, p_6) \mid p_1 + \cdots + p_6 = 1, 0 < p_i < 1\}$ és un obert de $\mathbb{R}^5$ i, per tant,

$$-2\log(\lambda(x)) \xrightarrow{n \to \infty} \chi^2_{(5)}\,[11].$$

Llavors, com $-2\log(\lambda(x)) = 7.247394$, escollint $\alpha = 0.05$, com $n = 200$ és «suficientment gran» aleshores:

$$0.05 = P_{H_0}(A_0^c) = P(-2\log(\lambda(x)) > -2\log(c_\alpha)) = P(\chi^2_{(5)} > \tilde{c}_\alpha).$$

En la taula de la χ-quadrat podem trobar que $P(\chi^2_{(5)} \leq k) = 0.95 \iff k = 11.0705$, i com $-2\log(\lambda(x)) = 7.247394 < 11.0705$, tenim que $x \in A_0$. Així doncs, no podem rebutjar H_0; és a dir, no tenim raons per suposar que el dau no sigui perfecte i, per tant, podem acceptar que ho és. ∎

[11] Com comentàvem anteriorment, la sisena probabilitat depèn de les cinc anteriors, pel que solament tenim $6 - 1 = 5$ graus de llibertat.

4 Llista 4

Exercici 4.1. *En un experiment sobre la conducta humana, un psicòleg va demanar a quatre homes i quatre dones que s'asseguessin en una taula rectangular. Dues persones havien de ser cap de taula i les altres sis s'havien d'asseure als dos laterals, tres a cada banda. Els dos caps de taula eren especials perquè tenien una posició dominant.*

1. *Si les persones elegeixen el seu lloc a l'atzar, quina és la probabilitat que els caps de taula siguin dos homes? I dues dones? I un home i una dona?*
2. *L'experiment es va realitzar 28 cops i es va observar que en 9 ocasions els caps de taula eren homes, en 5 eren dones i en 14 eren un home i una dona. Contrasteu la hipòtesi que les persones trien els llocs a l'atzar (amb $\alpha = 0.1$).*

Demostració.

1. Utilitzarem $H_2 = \{\text{dos homes}\}$, $D_2 = \{\text{dues dones}\}$, $DH = \{\text{dona i home}\}$. Les probabilitats s'han de calcular com fèiem a ADIP, o a la primera part de Probabilitats:
$$P(H_2) = P(D_2) = \frac{\binom{4}{1}\binom{3}{1}\cdot 6!}{8!} = \frac{3}{14},\ P(DH) = 1 - 2\cdot P(D_2) = \frac{4}{7}.$$
2. Aplicarem el test de la χ^2 per la llei multinomial. Se'ns demana contrastar les hipòtesis:
$$H_0 : P(H_2) = P(D_2) = \frac{\binom{4}{1}\binom{3}{1}\cdot 6!}{8!} = \frac{3}{14},\ P(DH) = 1 - 2\cdot P(D_2) = \frac{4}{7} \text{ contra}$$
$$H_1 : P(H_2) \neq P(D_2) \text{ o bé } P(DH) \neq \frac{3}{14}.$$
Podem ser més formals. Si $X = (X_{D_2}, X_{H_2}, X_{DH})$ on, per exemple, X_{D_2} és la variable que compta quants cops s'han assegut dues dones com a cap de taula, volem contrastar les hipòtesis:
$$H_0 : X \sim M\left(28; \frac{3}{14}, \frac{3}{14}, \frac{4}{7}\right) \text{ i } H_1 : X \nsim M\left(28; \frac{3}{14}, \frac{3}{14}, \frac{4}{7}\right).$$
Per fer-ho, escriurem una taula amb tots els valors que necessitem per emprar l'*estadístic* χ-quadrat de Pearson. Recordem que p_j^0 són les probabilitats sota la hipòtesi nul·la:
$$D_n(x) = \sum_{j=1}^{3} \frac{(n_j - np_j^0)^2}{np_j^0} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{(3-1)} = \chi^2_{(2)}.$$
Per tant, i recordant que $n = n_1 + n_2 + n_3 = 28$:

	n_j	p_j^0	np_j^0	$(n_j - np_j^0)^2$	$\frac{(n_j-np_j^0)^2}{np_j^0}$
H_2	9	$\frac{3}{14}$	6	9	$\frac{9}{6} = 1.5$
D_2	5	$\frac{3}{14}$	6	1	$\frac{1}{6} = 0.167$
DH	14	$\frac{4}{7}$	16	4	$\frac{4}{16} = 0.25$
	28	1	28	14	**1.917**

Així doncs, $D_n(x) = 1.917$. Recordem que la regió d'acceptació és $A_0 = \{D_n \leq d_\alpha\}$, tal que $P_{H_0}(A_0^c) = \alpha$. Amb el comportament asimptòtic de D_n, i per $\alpha = 0.1$:

$$P_{H_0}(A_0^c) = P_{H_0}(D_n > d_\alpha) = P(\chi^2_{(2)} > d_\alpha) = 0.1 \implies d_\alpha = 4.605.$$

Com $A_0 = \{D_n \leq d_\alpha\}$, $D_n = 4.605$ i $d_\alpha = 4.605$, acceptem la hipòtesi nul·la; és a dir, no tenim prou evidències per a rebutjar la hipòtesi que les persones trien els llocs a l'atzar. ■

Exercici 4.2. *Una formatgeria rep llet de dues vaqueries A i B. El formatger desitja estudiar els valors dels productes rebuts. Per això extrau dues mostres que són analitzades per tal d'obtenir el contingut de matèria grassa. Els resultats de les anàlisis són:*

A	0.360	0.405	0.395	0.340	0.380	0.395	0.365	0.400	0.385	0.395	0.365	0.375	0.340
B	0.430	0.425	0.350	0.335	0.430	0.375	0.355	0.410	0.395	0.455	0.425	0.375	0.385

Contrasteu la igualtat de variàncies. Suposeu normalitat i feu un test de la raó de versemblança.

Demostració. Siguin $X_1, \dots, X_n$ VAIID tals que $A \sim N(\mu_A, \sigma_A^2)$, i $Y_1, \dots, Y_n$ tal que $B \sim N(\mu_B, \sigma_B^2)$. Volem testejar $H_0 : \sigma_A^2 = \sigma_B^2$ contra $H_1 : \sigma_A^2 \neq \sigma_B^2$.

$$\Theta = \{(\mu_A, \mu_B, \sigma_A^2, \sigma_B^2) \in \mathbb{R}^2 \times (0, +\infty)^2\},$$
$$\Theta_0 = \{(\mu_A, \mu_B, \sigma_A^2, \sigma_B^2) \in \mathbb{R}^2 \times (0, +\infty)^2 \mid \sigma_A^2 = \sigma_B^2\}.$$

El model estadístic seria quelcom tal que $(\Omega, \mathscr{A}, \mathscr{P})$, $\Omega = (0, +\infty)^{26}$, $\mathscr{A} = \mathscr{B}((0,+\infty)^{26})$ i $\mathscr{P} = \{P_A \times P_B \mid P_A \sim N(\mu_A, \sigma_A^2), P_B \sim N(\mu_B, \sigma_B^2)\}$. Posem:

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0} \mathscr{L}(x,\theta)}{\sup_{\theta\in\Theta} \mathscr{L}(x,\theta)}$$

Necessitem, doncs, la funció de versemblança. Anem a calcular-la:

$$\begin{aligned}
\mathscr{L}(x,\theta) &= \prod_{i=1}^n \mathscr{L}((\mu_A, \sigma_A^2), x_i) \prod_{i=1}^n \mathscr{L}((\mu_B, \sigma_B^2), y_i) \\
&= \left(\frac{1}{2\pi\sigma_A^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_A^2}\sum_{i=1}^n (x_i - \mu_A)^2\right\} \cdot \left(\frac{1}{2\pi\sigma_B^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_B^2}\sum_{i=1}^n (y_i - \mu_B)^2\right\} \\
&= \frac{1}{(2\pi)^n} \frac{1}{(\sigma_A^2\sigma_B^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma_A^2}\sum_{i=1}^n (x_i - \mu_A)^2 - \frac{1}{2\sigma_B^2}\sum_{i=1}^n (y_i - \mu_B)^2\right\}.
\end{aligned}$$

Sabem que $\hat{\mu}_A = \bar{X}$ i $\hat{\mu}_B = \bar{Y}$. Per definició,

$$\sigma_A^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2, \ \ \sigma_B^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Pel que $\hat{\sigma}_A^2 = S_A^2$ i $\hat{\sigma}_B = S_B^2$ són els respectius estimadors de màxima versemblança. En total, $\hat{\theta} = (\bar{X}, \bar{Y}, S_A^2, S_B^2)$. S'hauria de comprovar que els estimadors maximitzen la funció, i per fer-ho s'han de fer els càlculs $\partial_{\mu_A}, \partial_{\mu_B}, \partial_{\sigma_A^2}, \partial_{\sigma_B^2}$ i calcular la matriu Hessiana. Sota H_0 ($\sigma = \sigma_A = \sigma_B$),

$$\hat{\mu}_A = \bar{X},\ \hat{\mu}_B = \bar{Y},\ \hat{\sigma}^2 = \frac{S_A^2 + S_B^2}{2}.$$

i igual que abans, s'ha de comprovar que són, efectivament, estimadors de màxima versemblança. En total, $\hat{\theta}_0 = (\bar{X}, \bar{Y}, \frac{S_A^2+S_B^2}{2}, \frac{S_A^2+S_B^2}{2})$. A partir del segon apartat,

$$\begin{aligned}\lambda(x) = \frac{\mathscr{L}(x,\hat{\theta}_0)}{\mathscr{L}(x,\hat{\theta})} &= \frac{\frac{1}{(2\pi)^n}\frac{1}{\hat{\sigma}^n}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i-\bar{X})^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (y_i-\bar{Y})^2\right\}}{\frac{1}{(2\pi)^n}\frac{1}{(\hat{\sigma}_A^2\hat{\sigma}_B^2)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\hat{\sigma}_A^2}\sum_{i=1}^n (x_i-\bar{X})^2 - \frac{1}{2\hat{\sigma}_B^2}\sum_{i=1}^n (y_i-\bar{Y})^2\right\}} \\ &= 2^n\left(\frac{S_A^2}{S_A^2+S_B^2}\right)^{\frac{n}{2}}\cdot\left(\frac{S_B^2}{S_A^2+S_B^2}\right)^{\frac{n}{2}}.\end{aligned}$$

A partir d'ara es donarà un esbós de la solució. Una de les conseqüències del teorema de Fisher és la següent:

$$\frac{n_A S_{n_A}^2 (n_B-1)\sigma_B^2}{n_B S_{n_B}^2 (n_A-1)\sigma_A^2} \sim F_{(n_A-1, n_B-1)},$$

on $F_{(n_A-1,n_B-1)}$ és la distribució F de Fisher. Sota la hipòtesi nul·la, $\frac{\sigma_A^2}{\sigma_B^2} = 1$ i, per tant, una regió d'acceptació de nivell 0.1 vindrà donada per:

$$A_0 = \left\{x \in \Omega \;\middle|\; F_{(12,12)}(0.95) < \frac{\sum_i^{n_A}(x_i-\bar{X})^2}{\sum_i^{n_B}(y_i-\bar{Y})^2} < F_{(12,12)}(0.05)\right\} \implies \frac{\sum_i^{n_A}(x_i-\bar{X})^2}{\sum_i^{n_B}(y_i-\bar{Y})^2} = 0.35$$

i, a més, $F_{(12,12)}(0.95) = 0.37$, $F_{(12,12)}(0.05) = 2.69$. Per tant, estem fora de la regió d'acceptació i refusem hipòtesi nul·la. ■

Exercici 4.3. *Segons un model genètic, tres caràcters mútuament excloents A, B, C d'una població es presenten amb probabilitats $P(A) = p^2$, $P(B) = pq$, $P(C) = q$, on p, q són dos paràmetres tals que $p + q = 1$. D'una mostra de 90 individus es va observar que 12 tenien el caràcter A, 15 el B i 63 el C. Podem acceptar que les dades s'ajusten al model teòric amb $\alpha = 0.05$?*

Demostració. Per saber si podem acceptar que els dades s'ajusten al model teòric amb $\alpha = 0.05$, usarem un test de la χ^2 aplicat al model multinomial parametritzat de l'exercici per a contrastar H_0 contra H_1:

$$H_0 : p_A = p^2, p_B = p(1-p) \text{ i } p_C = 1-p,$$
$$H_1 : p_A \neq p^2 \text{ o } p_B \neq p(1-p) \text{ o } p_C \neq 1-p.$$

Primer anem a calcular la funció de versemblança per la nostra $x = (x_1, \ldots, x_{90}) \in \{A, B, C\}^n$, usant $\mathscr{L}(x,p) = \prod_{i\in\{A,B,C\}} p_i^{n_i} = p_A^{n_A}\cdot p_B^{n_B}\cdot p_C^{n_C}$[12]:

$$\mathscr{L}(x,p) = (p^2)^{12}(p(1-p))^{15}(1-p)^{63} = p^{39}(1-p)^{78} \implies \log\mathscr{L}(x,p) = 39\log p + 78\log(1-p).$$

[12] En un model multinomial parametritzat, suposant $(n_1, \ldots, n_m)$ el vector de freqüències absolutes, la funció de versemblança es pot escriure com $\mathscr{L}(x,\theta) = \prod_{i=1}^m p_i(\theta)^{n_i}$.

Calculem un estimador de màxima versemblança (EMV):

$$\partial_p \log \mathscr{L}(x,p) = 0 \iff \frac{39}{n} - \frac{78}{1-p} = 0 \iff p = \frac{39}{117} = \frac{1}{3}.$$

$$\partial_p^2 \log \mathscr{L}(x,p) = -\left(\frac{39}{p^2} + \frac{78}{(1-p)^2}\right) < 0.$$

Així doncs, obtenim que $\hat{p} = \frac{1}{3}$ és un estimador de màxima versemblança. Ara bé, si considerem l'estadístic de la χ-quadrat de Pearson:

$$D_n = \sum_{i=1}^{n} \frac{(n_i - n p_i(\lambda))^2}{n p_i(\lambda)} \implies D_{90} = \frac{(12-90\hat{p}^2)^2}{90\hat{p}^2} + \frac{(15-90\hat{p}(1-\hat{p}))^2}{90\hat{p}(1-\hat{p})} + \frac{(63-90(1-\hat{p}))^2}{90(1-\hat{p})} = \frac{9}{5},$$

tenim que la regió d'acceptació és $A_0 = \{x \mid D_{90} \leq c_\alpha\}$. Cramer va demostrar que sota H_0 (i bones condicions),

$$D_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{(m-d-1)},$$

on m és el nombre de variables ($m = 3$ en el nostre cas) i d el nombre de paràmetres ($d = 1$ en el nostre cas). Així doncs, per $n = 90$, podem suposar que $D_{90} \sim \chi^2_{(1)}$ i, llavors, trobem c_α imposant que:

$$0.05 = P_{H_0}(A_0^c) = 1 - P(D_{90} < c_\alpha) \iff P(D_{90} < c_\alpha) = P(\chi^2_{(1)} < c_\alpha) = 0.05.$$

Amb una taula de la $\chi^2_{(1)}$, per exemple, trobem que $c_\alpha = 3.841459 > D_{90} = 1.8$; per tant, la mostra cau dins d'A_0. Aleshores, no podem rebutjar H_0 i no podem dir que no se segueixen les probabilitats proposades. ■

Exercici 4.4. *Tenim una mostra de mida $n = 10$ d'una distribució de Poisson de paràmetre λ desconegut.*

1. *Es vol contrastar $H_0 : \lambda = 0.3$ contra $H_1 : \lambda = 0.4$. Construïu un test amb nivell $\alpha = 0.05$. Calculeu la potència d'aquest test. És UMP?*
2. *EL test de l'apartat anterior es pot estendre per a contrastar $H_0 : \lambda = 0.3$ contra $H_1 : \lambda > 0.3$? Si la resposta és positiva, té alguna propietat bona?*
3. *En una mostra d'una distribució de Poisson s'ha observat:*

Dades	*0*	*1*	*2*	*3*	*4*
Freqüència	*38*	*36*	*17*	*8*	*1*

Contrasteu $H_0 : \lambda = 0.3$ contra $H_1 : \lambda \neq 0.3$.

Exercici 4.5. *Les dades següents representen la durada en hores de 20 vàlvules elèctriques:*

7.2	*37.8*	*49.6*	*21.4*	*67.2*	*41.1*	*7.4*	*3.8*	*8.1*	*23.2*
72.1	*11.4*	*17.5*	*33.4*	*29.8*	*57.8*	*84.6*	*12.8*	*2.9*	*42.7*

Podem acceptar amb un error de primera espècie menor que 0.01 que la durada d'aquestes vàlvules segueix una distribució exponencial?

Demostració. **Donarem una idea de la resolució**. Utilitzarem el test d'ajustament de la χ^2 per un model multinomial parametritzat amb un nivell de significació menor que 0.01, agrupant les dades en intervals disjunts $I_1, \dots, I_m$ (ja veurem quin valor pren $m \in \mathbb{N}$) que recobreixin tots els valors possibles de la distribució. Volem contrastar les hipòtesis:

$$H_0 : (p_1, \dots, p_m) \in \{p_1(\lambda), \dots, p_m(\lambda)\} \text{ contra } H_1 : (p_1, \dots, p_m) \notin \{p_1(\lambda), \dots, p_m(\lambda)\},$$

on $p_j(\lambda)$ representa la probabilitat d'una distribució exponencial de paràmetre $\lambda > 0$[13]. Així doncs, haurem de calcular la regió d'acceptació per a $\alpha < 0.01$, $A_0 = \{x \mid D_n(x) \leq c_\alpha\}$, on:

$$D_n(x) = \sum_{i=1}^{m} \frac{(n_i - np_i(\hat{\lambda}))^2}{np_i(\hat{\lambda})} \tag{4.1}$$

és l'estadístic de la χ-quadrat de Pearson, $(n_1, \dots, n_m)$ el vector de freqüències absolutes. Recordem que el test d'ajustament de la χ-quadrat es basa en l'aproximació:

$$D_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{(m-d-1)}$$

i, així, sota H_0, imposem:

$$0.01 > \alpha = P_{H_0}(D_n > c_\alpha) = P_{H_0}(\chi^2_{(m-d-1)} > c_\alpha) \implies P_{H_0}(\chi^2_{(m-d-1)} \leq c_\alpha) > 0.9,$$

on en el nostre cas tenim $d = 1$ paràmetres. Necessitem conèixer el valor de m i $\hat{\lambda}$ per poder fer el test. Ara bé, si escrivim els intervals com $I_i = [a_i, b_i)$, $a_i < b_i$ $(i = 1, \dots, m)$, aleshores:

$$\mathscr{L}(x, \lambda) = \prod_{i=1}^{m} p_i(\lambda)^{n_i} = \prod_{i=1}^{m} (e^{-\lambda a_i} - e^{-\lambda b_i})^{n_i},$$

el qual és molt complicat de maximitzar (caldria usar mètodes numèrics).

Teorema 4.6. *Siguin $\hat{\lambda}_1, \dots, \hat{\lambda}_d$ els estimadors de màxima versemblança de la mostra completa. Aleshores, l'estadístic (4.1) sota les condicions $np_i(\lambda) > 0$ i $n_i \geq 5$ per a tot i, segueix asimptòticament una llei amb funció de distribució F tal que:*

$$F_{\chi^2_{(m-1)}}(x) \leq F(x) \leq F_{\chi^2_{(m-1)}}(x).$$

Per tant, imposant:

$$0.01 > \alpha = P_{H_0}(\chi^2_{(m-1)} > C_\alpha) \implies P_{H_0}(\chi_{(m-1)})^2 \leq C_\alpha) > 0.9,$$

podem establir la *regla de decisió* següent: sigui D_n el valor que pren l'estadístic χ-quadrat de Pearson $D_n(x)$ per a la mostra,

[13] En altres paraules, a la primera hipòtesi demanem que tota probabilitat segueixi una exponencial λ, mentre que en la segona ha d'*existir* una (o més) tal que no segueixi una exponencial.

1. Si $D_n > C_\alpha$ rebutgem la hipòtesi nul·la, ja que aleshores:

$$\begin{aligned} p\text{-valor}^{14} = P_{H_0}(D_n(x) > D_n) &= 1 - P_{H_0}(D_n(x) \leq D_n) = 1 - F(D_n) \leq 1 - F_{\chi^2_{(m-1)}}(D_n) \\ &< 1 - F_{\chi^2_{(m-1)}}(C_\alpha) = 1 - P_{H_0}(\chi^2_{(m-1)} \leq C_\alpha) = P_{H_0}(\chi^2_{(m-1)} > C_\alpha) = \alpha. \end{aligned}$$

2. Si $D_\alpha < c_\alpha$ no rebutgem la hipòtesi nul·la, ja que aleshores:

$$\begin{aligned} p\text{-valor} = P_{H_0}(D_n(x) > D_n) &= 1 - P_{H_0}(D_n(x) \leq D_n) = 1 - F(D_n) \leq 1 - F_{\chi^2_{(m-d-1)}}(D_n) \\ &> 1 - F_{\chi^2_{(m-d-1)}}(c_\alpha) = 1 - P_{H_0}(\chi^2_{(m-d-1)} \leq c_\alpha) = P_{H_0}(\chi^2_{(m-d-1)} > c_\alpha) = \alpha. \end{aligned}$$

3. Per últim, si $c_\alpha < D_\alpha < C_\alpha$ no prenem cap decisió.

Recordem que l'objectiu és aplicar el test d'ajustament de la χ^2 per un model multinomial. Per tant, si recordem (4.1), necessitem $p_i(\lambda)$ (en particular, l'EMV). Així doncs, anem a calcular la funció de màxima versemblança per la funció de versemblança de la mostra completa següent:

$$\mathscr{L}(x, \lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda x_i\} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\}.$$

Amb aquest objectiu, trobem $\hat{\lambda}$ tal que $\partial_\lambda \log \mathscr{L}(x, \hat{\lambda}) = 0$:

$$\partial_\lambda \log \mathscr{L}(x, \lambda) = 0 \iff \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \iff \hat{\lambda} = \frac{n}{\sum_i x_i},$$

i comprovem que, efectivament, $\hat{\lambda}$ és un estimador de màxima versemblança ja que $\partial^2_\lambda \log \mathscr{L}(x, \lambda) = -\frac{n}{\lambda^2} < 0$ per a tot $\lambda > 0$. Ens queda que:

$$\hat{\lambda} = \frac{n}{\sum_i x_i} = \frac{20}{631.8} = 0.03165559.$$

Hem de dividir en grups, independentment de com sigui la mostra, de manera que cada grup satisfaci les propietats $np_i(\lambda) > 0$ i $n_i \geq 5$, per a tot $i \in \{1, \dots, m\}$. A partir de la seqüència $\{0, 10, 25, 45, 60, \dots\}$, que és independent de la mostra, prenem els intervals $[0, 10)$, $[10, 25)$, $[25, 45)$, $[45, +\infty)$. Aleshores, per a cada interval (a, b) tenim:

$$p_k(\hat{\lambda}) = \int_a^b \hat{\lambda} e^{-\lambda x}\, dx = e^{-\hat{\lambda} a} - e^{-\hat{\lambda} b} > 0,$$

de manera que podem fer la taula següent (omplint els valors corresponents amb els càlculs):

	$[0, 10)$	$[10, 25)$	$[25, 45)$	$[45, +\infty)$
n_i	–	–	–	–
$p_i(\hat{\lambda})$	–	–	–	–
$np_i(\hat{\lambda})$	–	–	–	–

14 El p-valor s'acostuma a utilitzar en els contrastos d'hipòtesi i és el valor que s'usa com a referència per a acceptar o rebutjar una hipòtesi. S'accepta quan és major a 0.05 (llavors diem que el p-valor no és significatiu, és a dir, és major que el valor de significació) i es rebutja en cas contrari.

Per tant,

$$D_4 = \sum_{i=1}^{4} \frac{(n_i - np_i(\lambda))^2}{np_i(\lambda)} = 0.2195687.$$

Només ens quedaria calcular c_α i C_α tals que:

$$P_{H_0}(\chi_{(4-1-1)} \leq c_\alpha) > 0.9 \text{ i } P_{H_0}(\chi_{(4-1)} \leq C_\alpha) > 0.9.$$

Recordem que tenim $d = 1$ paràmetres, el qual és λ (restem un grau de llibertat a l'estimar) i $m = 4$ intervals. Amb R o similar trobem que cal prendre $c_\alpha > 4.60517$ ($\alpha < 0.01$). Observem que sempre podem prendre $\alpha < 0.01$ tal que $D_4 = 0.2195687 < 4.60517 < c_\alpha$. Per exemple, podem prendre $\alpha = 0.05$ i llavors $c_\alpha = 5.991465$. En aquest cas, no ens cal calcular C_α i observem que fallem en rebutjar la hipòtesi nul·la, ja que no tenim prou evidències per a no acceptar amb un error de primera espècie menor que 0.01 que la durada d'aquestes vàlvules segueix una distribució exponencial (en aquest cas de paràmetre $\hat{\lambda}$). ■

Exercici 4.7. *Repetiu el problema, ara usant el test de Kolmogorov.*

Demostració. En aquest problema ens demanen que contrastem:

$$H_0 : X \sim \text{Exp}(0.03) \quad \text{i} \quad H_1 : X \nsim \text{Exp}(0.03).$$

Usarem el test de Kolmogorov, en particular l'estadístic

$$\Delta_n = \sup_{y \in \mathbb{R}} \left| F_0(y) - F_n(y) \right|,$$

on F_0 en el nostre cas és la funció de distribució d'una Exp(0.03) i F_n és la funció de distribució empírica. Com que la funció F_0 és contínua i creixent i F_n és creixent i esglaonada, les diferències més grans es troben en els salts o just abans. Per tant,

$$\Delta_n = \sup_{1 \leq j \leq n} \left\{ \left| F_0(x_{(j)}) - \frac{j}{n} \right|, \left| F_0(x_{(j)}) - \frac{j-1}{n} \right| \right\},$$

on $x_{(j)}$ és la dada del problema que ocupa el lloc j-èsim quan les ordenem. Per resoldre el problema, és recomanable elaborar una taula: a la primera columna escriurem les dades ordenades $x_{(j)}$, a la segona columna posarem $F_n(x_{(j)})$, a la tercera columna donarem $F_0(x_{(j)})$ $\left(F_0 \sim \text{Exp}(0, 03)\right)$ i a la quarta columna indicarem el valor de

$$\Delta_{j,n} = \max \left\{ \left| F_0(x_{(j)}) - \frac{j}{n} \right|, \left| F_0(x_{(j)}) - \frac{j-1}{n} \right| \right\}.$$

Fem-ho:

$x_{(j)}$	$F_n(x_{(j)}) = \frac{j}{n}$	$F_0(x_{(j)}) = 1 - e^{-0.03x_{(j)}}$	$\Delta_{j,n}$
2.9	0.05	0.0833	0.0833
3.8	0.10	0.1077	0.0577
7.2	0.15	0.1943	0.0943
7.4	0.20	0.1991	0.0491
8.1	0.25	0.2157	0.0343
11.4	0.30	0.2897	0.0397
12.8	0.35	0.3189	0.0311
17.5	0.40	0.4084	0.0584
21.4	0.45	0.4738	0.0738
23.2	0.50	0.5014	0.0514
29.8	0.55	0.5910	0.0910
33.4	0.60	0.6329	0.0829
37.8	0.65	0.6783	0.0783
41.1	0.70	0.7086	0.0586
42.7	0.75	0.7222	0.0278
49.6	0.80	0.7742	0.0258
57.8	0.85	0.8234	0.0266
67.2	0.90	0.8668	0.0332
72.1	0.95	0.8850	0.0650
84.6	1.00	0.9210	0.0790

El valor empíric de Δ_n és 0.0943 i la regió d'acceptació és $A_0 = \{\Delta_n \leq c_{\alpha,n}\}$, satisfent $P_0(A_0^c) = 0.01$. Buscant a les taules, observarem que $c_{\alpha,n} = 0.3524$. Així, com $0.0943 \leq 0.3524$, no podem rebutjar que la durada de les vàlvules segueix una exponencial Exp(0.03). ∎

Referències

[San99] Marta Sanz-Solé. *Probabilitats*. cat. UB; 28. Barcelona: Edicions Universitat de Barcelona, 1999. isbn: 8483380919.

[JM10] Olga Julià de Ferran i David Márquez-Carreras. *Un Primer curs d'estadística*. cat. Universitat ; 48. Barcelona: Publicacions i Edicions Universitat de Barcelona, 2010. isbn: 9788447534838.

[CV21] Eloi Castaño Camps i Víctor Valldoseria i Socías. *Estadística*. Català. Barcelona, maig de 2021.

Apunts de l'assignatura d'Estadística, impartida per David Márquez durant la primavera de 2021.

[JMR21] Olga Julià de Ferran, David Márquez-Carreras i Carles Rovira Escofet. *Estadística: problemes i més problemes*. cat. Textos docents; 431. Barcelona: Edicions de la Universitat de Barcelona, 2021. isbn: 9788491686453.