Universitat Autònoma de Barcelona
Facultat de Ciències

Modelització i Inferencia
Pràctica 3

Autor:
Gerard Lahuerta Martín
NIU: 1601350

16 de Desembre del 2021

Índex

1 Presentació del problema

A un estudi s'avalua la sensibilitat π d'un test diagnòstic per l'asma que és més econòmic que el test de referència. Cada pacient és testejat repetidament fins que s'obté el primer positiu. Considera que la variable aleatòria X_i és el nombre de tests que s'ha fet el pacient i fins a obtenir el primer positiu i que tots els pacients i els successius tests són independents. Considera, a més a més, que la sensibilitat π és igual per a tots els pacients i tests.

(a) Deriva la funció de massa de probabilitat de X_i.

(b) Escriu la log-versemblança per a una realització $x_1, \cdots, x_n$ d'una mostra aleatòria $X_1, \cdots, X_n$ i calcula l'estimador màxim versemblant de π: $\hat{\pi}$.

(c) Calcula l'error estàndard de $\hat{\pi}$: $se(\hat{\pi})$.

Es realitza l'experiment anterior amb nou pacients i s'obtenen els següents resultats:

$$\{x_1 = 3, x_2 = 5, x_3 = 2, x_4 = 6, x_5 = 9, x_6 = 1, x_7 = 2, x_8 = 2, x_9 = 3\}$$

(d) Donada l'estimació puntual per a π basada en l'estimador màxim versemblant trobat a (b) i construeix l'interval de confiança de *Wald* del 95% per a π. Per què aquest interval de confiança podria ser problemàtic? Justifica la teva resposta.

(e) Considera la següent parametrització:

$$\eta = \log\left(\frac{\pi}{1-\pi}\right)$$

Troba l'estimador i l'estimació màxim versemblant per a $\eta, \hat{\eta}$, i construeix l'interval de confiança de *Wald* del 95% per a η.

(f) Compara els intervals obtinguts als apartats (d) i (e) un cop retransformat per a π.

2 Resolució del primer apartat

Observem que la variable aleatòria X_i, per a $i \in N$, compta el nombre d'intents a fer l'experiment (el test clínic) per a obtenir el primer èxit (positiu en el test). Per tant, la funció de massa de X_i serà el producte entre la probabilitat de que no es tracti d'un exit en els primers $x_i - 1$ intents i que sigui exitos l'intent x_i:

$$P(X_i = x_i) = \pi(1-\pi)^{x_i-1} := f(\pi|x_i)$$

Observem per tant que és la mateixa que la de la variable aleatoria Geomètrica amb paràmetre desconegut π la probabilitat d'exit.

$\Longrightarrow$ $\boxed{\text{La funció de densitat de la variable } X_i \text{ és: } f(\pi|x_i) = \pi(1-\pi)^{x_i-1}}$

3 Resolució del segon apartat

Calculem la log-versemblança per a una realització $x_1, \cdots, x_n$ d'una mostra aleatòria $X_1, \cdots, X_n$, calculant prèviament la funció de màxima versemblança de l'estimador $\hat{\pi}$.

- Funció de màxima versemblança de l'estimador: $L(\pi|x_i)$

$$L(\pi|x_i) = \prod_{i=1}^{n} (F(\pi|x_i)) = \prod_{i=1}^{n} \left(\pi(1-\pi)^{x_i - 1}\right) =$$
$$= \pi^n (1-\pi)^{-n+\sum_{i=1}^{n} x_i}$$

- Funció log-versemblança de l'estimador: $l(\pi|x_i)$

$$\begin{aligned} l(\pi|x_i) = &\quad \log(L(\pi|x_i)) = \log\left(\pi^n (1-\pi)^{-n+\sum_{i=1}^{n} x_i}\right) = \\ = &\quad n\log(\pi) + \left(-n + \sum_{i=1}^{n} x_i\right)\log(1-\pi) \Longrightarrow \end{aligned}$$

$\Longrightarrow$ $\boxed{\text{Log-versemblança: } l(\pi|x_i) = n\log(\pi) + (-n + \sum_{i=1}^{n} x_i)\log(1-\pi)}$

A partir de la funció log-versemblança podem calcular la funció *score* que, trobant el seu màxim, ens permet saber l'estimador $\hat{\pi}$ màxim versemblant.

- Calcul de la funció *score*: $s(\pi|x_i)$

$$s(\pi|x_i) = l'(\pi|x_i) = \left(n\log(\pi) + \left(-n + \sum_{i=1}^{n} x_i\right)\log(1-\pi)\right)' = \frac{n}{\pi} + \frac{n - \sum_{i=1}^{n} x_i}{1-\pi}$$

$$s(\pi|x_i) = 0 \Longleftrightarrow \frac{n}{\pi} + \frac{n - \sum_{i=1}^{n} x_i}{1-\pi} = 0 \Longleftrightarrow \frac{n}{\pi} = \frac{-n + \sum_{i=1}^{n} x_i}{1-\pi} \Longleftrightarrow$$
$$\Longleftrightarrow n - n\pi = -n\pi + \pi\sum_{i=1}^{n} x_i \Longleftrightarrow \pi = \frac{n}{\sum_{i=1}^{n} x_i} \Longrightarrow$$
$$\Longrightarrow \hat{\pi} = \bar{x}^{-1} \text{ on } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ és la mitjana de les mostres}$$

- Trobem si és el màxim estimador:

$$s'(\pi|x_i) = \left(\frac{n}{\pi} + \frac{n - \sum_{i=1}^{n} x_i}{1-\pi}\right)' = \frac{-n}{\pi^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1-\pi)^2} =$$

$$= \frac{-n(1-\pi) - n\pi^2 - \pi^2 \sum_{i=1}^{n} x_i}{\pi^2(1-\pi)^2} = \frac{2n\pi - n - \pi^2 \sum_{i=1}^{n} x_i}{\pi^2(1-\pi)^2} < 2n\pi - n - \pi^2 = 0 \Longleftrightarrow$$

$$\Longleftrightarrow \pi = \frac{-2n \pm \sqrt{4n^2 - 4}}{-2} = n \pm \sqrt{n^2 - 1}$$

com que $n \in N \Rightarrow$ té solució real, però no pertany a $[0, 1] \Longrightarrow \hat{\pi} = \bar{x}^{-1}$és màxim

$\Longrightarrow$ $\boxed{\text{L'estimador màxim versemblant de } \pi \text{ és: } \hat{\pi} = \bar{x}^{-1}}$

4 Resolució del tercer apartat

Calculem l'error estàndard de l'estimador $\hat{\pi}$, $se(\hat{\pi})$, sabent que: $se(\hat{\pi}) = \sqrt{J^{-1}(\pi)}$, on $J(\pi)$ és la informació de Fisher.

- Càlcul de la informació esperada de Fisher: $J(\pi) = E(I(\pi))$, on $I(\pi)$ és la informació observada de Fisher.

$$\begin{aligned} I(\pi) = \quad & -s'(\pi|X_i) = -\left(\frac{-n}{\pi^2} + \frac{n - \sum_{i=1}^n X_i}{(1-\pi)^2}\right) = \frac{n}{\pi^2} - \frac{n - \sum_{i=1}^n X_i}{(1-\pi)^2} = \\ = \quad & \frac{n(1-\pi)^2 - n\pi^2 + \pi^2 \sum_{i=1}^n X_i}{\pi^2(1-\pi)^2} = \frac{n - 2n\pi + \pi^2 \sum_{i=1}^n X_i}{\pi^2(1-\pi)^2} \end{aligned}$$

$$\begin{aligned} J(\pi) = \quad & E\left(\frac{n - 2n\pi + \pi^2 \sum_{i=1}^n X_i}{\pi^2(1-\pi)^2}\right) \underset{\substack{\uparrow \\ (I)}}{=} \frac{1}{\pi^2(1-\pi)^2}\left(n - 2n\pi + n\pi^2 E(X_i)\right) \underset{\substack{\uparrow \\ (II)}}{=} \\ = \quad & \frac{1}{\pi^2(1-\pi)^2}\left(n - 2n\pi + n\pi^2 \frac{1}{\pi}\right) = \frac{1}{\pi^2(1-\pi)^2}(n - 2n\pi + n\pi) = \\ = \quad & \frac{1}{\pi^2(1-\pi)^2}(n - n\pi) = \frac{n}{\pi^2(1-\pi)} \end{aligned}$$

(I): $X_1, \cdots, X_n$ *independents i idènticament distribuïdes*
(II): $E(X_i) = \frac{1}{\pi}$

$$\Longrightarrow se(\hat{\pi}) = \sqrt{\left(\frac{n}{\pi^2(1-\pi)}\right)^{-1}} = \pi\sqrt{\frac{1-\pi}{n}} \Longrightarrow$$

$\Longrightarrow$ $\boxed{\text{L'error estàndard de l'estimador } se(\hat{\pi}) = \pi\sqrt{\frac{1-\pi}{n}}}$

5 Resolució del quart apartat

Sabent que els resultats de 9 experiments:

$$\{x_1 = 3, x_2 = 5, x_3 = 2, x_4 = 6, x_5 = 9, x_6 = 1, x_7 = 2, x_8 = 2, x_9 = 3\}$$

Calculem ara l'interval de confiança de *Wald* del 95% per al paràmetre π.
Sabem que l'interval de confiança de *Wald* per al paràmetre π (de manera asimptòtica) és: $\pi \in \left[\hat{\pi} \pm z_{1-\frac{\alpha}{2}} \cdot se(\hat{\pi})\right]$ on:

- $\hat{\pi}$ és l'estimador màxim versemblant del paràmetre π trobat a l'apartat 3:

$$\hat{\pi} = \frac{9}{3+5+2+6+9+1+2+2+3} = \frac{9}{33} = \frac{3}{11} = 0.\bar{2}\bar{7}$$

- $z_{1-\frac{\alpha}{2}}$ és el quantil $1-\frac{\alpha}{2}$ corresponent a la distribució Normal amb paràmetres $(1, 0)$, en aquest cas, com volem calcular l'interval de confiança del 95% $\Rightarrow$ $\Rightarrow \alpha = 0,05 \Rightarrow z_{0,975} = 1,96$

- $se(\hat{\pi})$ és l'error estàndard del paràmetre trobat a l'apartat 4:

$$se(\hat{\pi}) = \hat{\pi}\sqrt{\frac{1-\hat{\pi}}{n}} = \frac{3}{11}\sqrt{\frac{1-\frac{3}{11}}{9}} = \frac{1}{11}\sqrt{\frac{8}{11}} = \frac{2}{11}\sqrt{\frac{2}{11}} = 0,07752753322$$

Per tant, l'interval de confiança de *Wald* del 95% per al paràmetre π és $\left[\frac{3}{11} \pm 1,96 \cdot \frac{2}{11}\sqrt{\frac{2}{11}}\right] \simeq$ $\simeq [0,120773308,\ 0,424681238] \Longrightarrow$
$\Longrightarrow$ $\boxed{\text{L'interval de confiança de } \textit{Wald} \text{ del 95\% és } \pi \in [0,121,\ 0,425]}$

Observem que, en aquest cas, interval conté valor que el paràmetre π si pot obtenir, però, l'interval de Confiança de *Wald* no sempre dona un interval que contingui valors que pot obtenir el paràmetre (com valors negatius o majors que 1). Si és el cas, és necessari reparametritzar el paràmetre amb la funció *logit* o log que pot portar la imatge de l'estimador a l'interval $[0, 1]$.

6 Resolució de cinquè apartat

Reparamatritzem el paràmetre π per la funció *logit* i calculem l'interval de confiança de *Wald* del 95% per a la reparamerització.

- Calculem la reparametrització $\eta = \log\left(\frac{\pi}{1-\pi}\right)$:

$$\eta = \log\left(\frac{\pi}{1-\pi}\right) \Rightarrow \frac{\pi}{1-\pi} = e^{\eta} \Rightarrow \pi = (1-\pi)e^{\eta} \Rightarrow \pi = \frac{e^{\eta}}{1+e^{\eta}}$$

- Càlcul de la log-versemblança de η utilitzant els càlculs de l'apartat 3: $l(\eta|x_i)$

$$l(\pi|x_i) = n\log(\pi) + \left(-n + \sum_{i=1}^{n} x_i\right)\log(1-\pi) \Rightarrow$$

$$\Rightarrow l(\eta|x_i) = n\log\left(\frac{e^{\eta}}{1+e^{\eta}}\right) + \left(-n + \sum_{i=1}^{n} x_i\right)\log\left(1 - \frac{e^{\eta}}{1+e^{\eta}}\right) =$$

$$= n\log\left(\frac{e^{\eta}}{1+e^{\eta}}\right) + \left(-n + \sum_{i=1}^{n} x_i\right)\log\left(\frac{1}{1+e^{\eta}}\right) =$$

$$= n\left(\eta - \log(1+e^{\eta})\right) - \left(-n + \sum_{i=1}^{n} x_i\right)\log\left(1+e^{\eta}\right) = n\eta - \log\left(1+e^{\eta}\right)\sum_{i=1}^{n} x_i$$

- Calculem la informació de Fisher per al paràmetre η: $I(\eta)$

$$I(\eta) = -\frac{\partial^2}{\partial^2\eta} l(\eta|x_i) = -\frac{\partial}{\partial\eta}\left(n - \frac{e^{\eta}\sum_{i=1}^{n} x_i}{1+e^{\eta}}\right) =$$

$$= -\frac{\partial}{\partial\eta}\left(n - \frac{e^{\eta}\sum_{i=1}^{n}(x_i) + \sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n}(x_i)}{1+e^{\eta}}\right) =$$

$$= -\frac{\partial}{\partial\eta}\left(n - \sum_{i=1}^{n}(x_i) + \frac{\sum_{i=1}^{n}(x_i)}{1+e^{\eta}}\right) = \frac{e^{\eta}\sum_{i=1}^{n}(x_i)}{(1+e^{\eta})^2} \Longrightarrow$$

$$\Longrightarrow I(\eta) = \frac{e^{\log\left(\frac{\pi}{1-\pi}\right)}\sum_{i=1}^{n}(x_i)}{\left(1 + e^{\log\left(\frac{\pi}{1-\pi}\right)}\right)^2} = \frac{\frac{\pi}{1-\pi}\sum_{i=1}^{n}(x_i)}{\left(1+\frac{\pi}{1-\pi}\right)^2}$$

- Calculem ara el valor de l'estimador $\hat{\eta}$ i de la informació de Fisher $I(\hat{\eta})$:

$$\hat{\eta} = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \log\left(\frac{\frac{3}{11}}{1-\frac{3}{11}}\right) = \log\left(\frac{\frac{3}{11}}{\frac{8}{11}}\right) = \log\left(\frac{3}{8}\right) = -0,980829253$$

$$I(\hat{\eta}) = \frac{\frac{\hat{\pi}}{1-\hat{\pi}}\sum_{i=1}^{n}(x_i)}{\left(1+\frac{\hat{\pi}}{1-\hat{\pi}}\right)^2} = \frac{33\cdot\frac{3}{8}}{\left(1+\frac{3}{8}\right)^2} = \frac{72}{11} = 6.\overline{54}$$

Calculem ara l'interval de confiança de *Wald* del 95% del paràmetre η:

$$\eta \in \left[\hat{\eta} \pm z_{0,975} \cdot \sqrt{I^{-1}(\hat{\eta})}\right] = \left[\log\left(\frac{3}{8}\right) \pm 1,96 \cdot \sqrt{\frac{11}{72}}\right] \simeq \left[-0,981 \pm 1,96 \cdot \sqrt{0,153}\right] =$$

$$= [-1,75, \;\; -0,215] \Longrightarrow$$

$$\Longrightarrow \boxed{\text{L'interval de confiança de } \textit{Wald} \text{ del 95\% reparametritzat per } \eta \text{ és } \eta \in [-1,75, \;\; -0,215]}$$

Recalcar, que l'estimador màxim versemblant de η: $\hat{\eta} = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$, ja que l'estimador és invariant a les transformacions, per la qual cosa si $\hat{\pi}$ és el màxim versemblant sense la transformació, una vegada aplicada, el màxim versemblant de la transformació és la transformació aplicada a $\hat{\pi}$.

7 Resolució del sisè apartat

Observem que els intervals trobats a l'apartat 5 i l'apartat 6 ($[0,121,\ 0,425]$ i $[-1,75,\ -0,215]$ respectivament), difereixen en els valors que contenen. Això és perquè en l'apartat 5 és l'interval en el qual es troba l'estimador real, π, en el 95% dels experiments i en l'apartat 6 és l'interval on es troba (en el 95% dels experiments) l'ordre de magnitud de la proporció exit-fracàs, també anomenat *Odds*, (en base neperiana) de l'experiment; essent (π i $1-\pi$ l'exit i el fracàs del test respectivament). Com que l'interval de l'estimador de η ésta per sota del 0, deduïm que l'orde del valor de $\frac{\pi}{1-\pi}$ és menor que 1, i per això $1-\pi > \pi$; per tant, hi ha més fracassos que èxits.
Concloem que els tests no són bons.