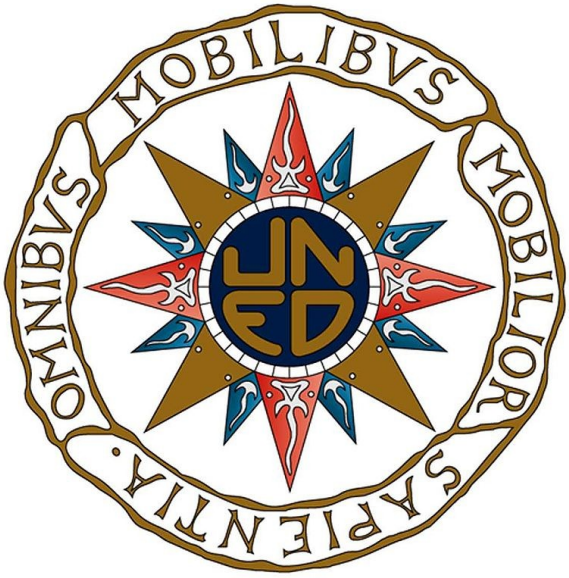

FACULTAD DE CIENCIAS

GRADO EN MATEMÁTICAS

Inferencia Estadística

EXÁMENES

Jose Maria Buades Rubio
Mayo-2021

INFERENCIA ESTADÍSTICA
EXÁMENES

2013 Septiembre - Reserva

Ejercicio 1. Sea X una variable aleatoria discreta con valores naturales y función de probabilidad

$$p_{\theta_0}(x) = (1-\theta)^2\, x\, \theta^{x-1} \qquad x = 1, 2, 3, \ldots$$

siendo $0 < \theta < 1$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

A) Estimar θ por el método de la máxima verosimilitud y por el método de los momentos.
B) Determinar un estadístico suficiente minimal.
C) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = \dfrac{1}{1-\theta}$.

Solución:

A) Se tiene que tomada una muestra de tamaño n, la función de verosimilitud viene dada por

$$L(\theta) = (1-\theta)^{2n}\, \theta^{x_1+\ldots+x_n-n} \left(\prod_{i=1}^{n} x_i \right) \qquad \theta \in (0,1)$$

tomando el logaritmo natural, se tiene que

$$\log L(\theta) = 2n \log(1-\theta) + (x_1 + \ldots + x_n - n) \log \theta + \log \left(\prod_{i=1}^{n} x_i \right)$$

Llamaremos $T = \sum_{i=1}^{n} x_i$. Derivando, se tiene que

$$\frac{\partial \log L}{\partial \theta}(\theta) = -\frac{2n}{1-\theta} + \frac{T-n}{\theta}$$

Por tanto, la función logaritmo-verosimilitud alcanzará su máximo en

$$-\frac{2n}{1-\theta} + \frac{T-n}{\theta} = 0 \;\Rightarrow\; \frac{2n}{1-\theta} = \frac{T-n}{\theta} \;\Rightarrow\; \frac{2n}{T-n} = \frac{1}{\theta} - 1 \;\Rightarrow\; \theta = \frac{T-n}{T+n}$$

El único caso especial es cuando ocurre $T = n$, ya que en ese caso la función logaritmo-verosimilitud es decreciente en todo su dominio $\theta \in (0,1)$, y habría que dar como estimación más verosímil el extremo inferior de este intervalo. Este valor justo coincide con la forma del estimador más verosímil dado para los demás valores de T y particularizado en $T = 0$, por tanto, lo damos como válido para este caso también. Por tanto, concluímos que el estimador de máxima verosimilitud para θ viene dado por

$$\boxed{\hat{\theta}_{\text{verosimilitud}} = \frac{T-n}{T+n}} \qquad \text{siendo } T = \sum_{i=1}^{n} X_i$$

Vamos a estimar ahora el mismo parámetro por el método de los momentos. Para ello, vamos a considerar la variable aleatoria $Y = X - 1$. La función de probabilidad de esta variable aleatoria viene dada por

$$p_{\theta_0}(y) = (1-\theta)^2\, (y+1)\, \theta^y = \binom{y+2-1}{y} (1-\theta)^2 \theta^y \qquad y = 0, 1, 2, \ldots$$

Por tanto, se tiene que Y sigue una distribución binomial negativa $BN(2, 1-\theta)$. Por tanto, se tiene que

$$E_\theta[Y] = 2\frac{1-(1-\theta)}{1-\theta} = \frac{2\theta}{1-\theta}$$

Por tanto, se tiene que será

$$E_\theta[X] = E_\theta[Y+1] = 1 + \frac{2\theta}{1-\theta} = \frac{1+\theta}{1-\theta}$$

Mientras que el momento muestral de primer orden sabemos que es la media muestral $\overline{X} = \frac{1}{n}T$. Por tanto, igualando los momentos poblacionales y muestrales de primer orden, se tiene que

$$\frac{1+\theta}{1-\theta} = \frac{1}{n}T \quad \Rightarrow \quad \theta\left(\frac{2}{n}T\right) = \frac{1}{n}T - 1 \quad \Rightarrow \quad \theta = \frac{T-n}{2T}$$

Por tanto, el estimador dado por el método de los momentos para el parámetro θ viene dado por

$$\boxed{\widehat{\theta}_{\text{momentos}} = \frac{T-n}{2T}} \qquad \text{siendo } T = \sum_{i=1}^{n} X_i$$

B) Tenemos que la función de probabilidad muestral conjunta se puede expresar como

$$p_\theta(x_1, \ldots, x_n) = (1-\theta)^{2n}\,\theta^{T-n}\left(\prod_{i=1}^{n} x_i\right) \qquad x_i = 1, 2, 3, \ldots$$

Por tanto, según el Teorema de Factorización, tenemos que T es un estadístico suficiente. También es suficiente minimal, ya que dadas dos muestras aleatorias simples de tamaño n: $(x_1, \ldots, x_n)$ y $(x_1', \ldots, x_n')$ se tiene que

$$\frac{p_\theta(x_1, \ldots, x_n)}{p_\theta(x_1', \ldots, x_n')} = \frac{(1-\theta)^{2n}\,\theta^{T-n}\left(\prod_{i=1}^{n} x_i\right)}{(1-\theta)^{2n}\,\theta^{T'-n}\left(\prod_{i=1}^{n} x_i'\right)} = \frac{\left(\prod_{i=1}^{n} x_i\right)}{\left(\prod_{i=1}^{n} x_i'\right)}\,\theta^{(T-T')}$$

es independiente de θ si y sólo si $T = T'$. Por tanto, se tiene que T es un estadístico suficiente minimal. Además es completo, ya que tenemos que la función de probabilidad muestral conjunta se puede escribir de forma exponencial tal que

$$p_\theta(x_1, \ldots, x_n) = (1-\theta)^{2n}\,\theta^{-n}\left(\prod_{i=1}^{n} x_i\right)\exp\{T\log\theta\}$$

y, como T es suficiente minimal y la imagen del dominio del parámetro $\theta \in (0,1)$ por la función $f(\theta) = \log\theta$ es $(-\infty, 0)$, que contiene abiertos de $\mathbb{R}$, se tiene que T es completo. (Pág. 273 Texto base).

C) Vamos a calcular la esperanza de T. Tenemos que

$$E_\theta[T] = nE_\theta[X] = n\frac{1+\theta}{1-\theta}$$

Como se tiene que $n\frac{1+\theta}{1-\theta}+n=\frac{2n}{1-\theta}$, entonces se tendrá que

$$E_\theta\left[\frac{T+n}{2n}\right]=\frac{1}{2n}\left(n\frac{1+\theta}{1-\theta}+n\right)=\frac{1}{1-\theta}$$

Por tanto, se tiene que $\frac{T+n}{2n}$ es un estadístico insesgado para $\frac{1}{1-\theta}$, función de un estadístico suficiente y completo. Por tanto, se tiene que

$$\frac{T-n}{2n}$$

es el ECUMV para $g(\theta)=\frac{1}{1-\theta}$.

INFERENCIA ESTADÍSTICA
EXÁMENES

Ejercicio 2. Sea X una variable aleatoria con función de densidad $f_0(x) = \frac{x}{2},\ 0 < x < 2$, bajo la hipótesis nula H_0, y con función de densidad $f_1(x) = 1/2,\ 0 < x < 2$, bajo la hipótesis alternativa H_1. Determinr un constraste de máxima potencia de nivel α para constrastar H_0 frente a H_1, utilizando una muestra aleatoria simple de tamaño n de X. ¿Qué conclusiones obtendría si se observaron los valores $X_1 = 1,\ X_2 = 0'5,\ X_3 = 1'5,\ X_4 = 0'6$ y el nivel de significación era $\alpha = 0'05$?

Solución:

Por el Lema de Neyman-Pearson, sabemos que el constraste a determinar tiene la forma

$$\phi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{si } f_1(x_1, \ldots, x_n) > k f_0(x_1, \ldots, x_n) \\ 0 & \text{si } f_1(x_1, \ldots, x_n) \leq k f_0(x_1, \ldots, x_n) \end{cases}$$

donde k viene determinado por

$$P_0\{f_1(x_1, \ldots, x_n) > k f_0(x_1, \ldots, x_n)\} = \alpha$$

se tiene que

$$f_1(x_1, \ldots, x_n) > k f_0(x_1, \ldots, x_n) \iff \frac{1}{2^n} > k \frac{\prod_{i=1}^n x_i}{2^n} \iff \prod_{i=1}^n \frac{x_i}{2} < k' \iff \sum_{i=1}^n -\log \frac{x_i}{2} > k'' \iff$$

$$\iff 4\left(\sum_{i=1}^n -\log \frac{x_i}{2}\right) > c$$

Se tiene que

$$f_0(x) = \frac{x}{2} = e^{\log \frac{x}{2}} = e^{-\left(-\log \frac{x}{2}\right)}$$

Tomando la variable aleatoria $Y = -\log \frac{X}{2}$, se tiene que su función de densidad (bajo H_0) viene dada por

$$2e^{-2y}$$

que es una distribución gamma $\gamma(1, 2)$. Por tanto, se tiene que la variable aleatoria $\sum_{i=1}^n -\log \frac{x_i}{2}$ tiene distribución gamma $\gamma(n, 2)$, por la propiedad reproductiva de la función gamma. A su vez, por la relación entre la distribución gamma y la distribución chi-cuadrado, tenemos que $4\sum_{i=1}^n -\log \frac{x_i}{2}$ sigue una distribución χ^2_{2n}. Por tanto, se tiene que $c = \chi^2_{2n;\alpha}$. Así, se tiene que el contraste viene dado por

$$\phi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{si } \left(\sum_{i=1}^n -\log \frac{x_i}{2}\right) > \frac{\chi^2_{2n;\alpha}}{4} \\ 0 & \text{si } \left(\sum_{i=1}^n -\log \frac{x_i}{2}\right) \leq \frac{\chi^2_{2n;\alpha}}{4} \end{cases}$$

Para los valores dados, se tendría que $\left(\sum_{i=1}^4 -\log \frac{X_i}{2}\right) = 3'571$, mientras que $\frac{\chi^2_{8;0'05}}{4} = \frac{15'51}{4} = 3'8775$. Por tanto, se tiene que se aceptaría la hipótesis nula H_0.

INFERENCIA ESTADÍSTICA
EXÁMENES

Ejercicio 3. El tiempo, en minutos, que esperan los clientes de un determinado banco hasta que son atendidos sigue una distribución normal de media desconocida y desviación típica igual a 3. Los tiempos que esperaron diez clientes elegidos al azar fueron los siguientes: 1'5, 2, 2'5, 3, 1, 5, 5'5, 4'5, 3, 3. Determinar un intervalo de confianza de coeficiente de confianza 0'95, para el tiempo medio de espera.

Solución:

Se tiene que la media muestral, dado que concoemos la desviación típica poblacional $\sigma = 3$, seguirá una distribución normal $N(\mu, \sigma/\sqrt{n})$. Por tanto, se tiene que

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

Por tanto, el intervalo de confianza buscado vendrá dado por

$$\left[\overline{X} - z_{a/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{a/2}\frac{\sigma}{\sqrt{n}} \right]$$

Como se tiene que, para la muestra dada, $\overline{X} = 3'1,\ n = 10,\ a = 0'05,\ z_{0'025} = 1'96,$ se tiene que el intervalo de confianza buscado es

$$[1'24\,,\, 4'95]$$

INFERENCIA ESTADÍSTICA
EXÁMENES

Septiembre 2015 - Reserva

Ejercicio 1. Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta}{1-\theta} x^{(2\theta-1)/(1-\theta)} \qquad \text{si} \quad 0 < x < 1$$

siendo $1/2 < \theta < 1$. Se pide:
A) Determinar el estimador de θ por el método de la máxima verosimilitud y por el método de los momentos.
B) Determinar el estadístico suficiente minimal y, supuesto que es completo, el estimador centrado uniformemente de mínima varianza para θ.

Solución:

A) Suponemos que tomamos una muestra aleatoria simple de tamaño n. Entonces, se tiene que la función de verosimilitud viene dada por

$$L(\theta) = \left(\frac{\theta}{1-\theta}\right)^n \left(\prod_{i=1}^n x_i\right)^{(2\theta-1)/(1-\theta)} \qquad \text{si} \quad 1/2 < \theta < 1$$

Tomando logaritmos, se tiene que

$$\log L(\theta) = n\log\theta - n\log(1-\theta) + \frac{2\theta-1}{1-\theta}\left(\sum_{i=1}^n \log x_i\right)$$

Llamando $T = \sum_{i=1}^n \log x_i$ y ahora, derivando con respecto a θ, se tiene que

$$\frac{\partial \log L}{\partial \theta} = \frac{n}{\theta} + \frac{n}{1-\theta} + \frac{1}{(1-\theta)^2} T$$

e igualando a cero, buscamos el máximo de la función de verosimilitud

$$\frac{n}{\theta} + \frac{n}{1-\theta} + \frac{1}{(1-\theta)^2} T = 0 \quad \Rightarrow \quad n - \theta(n-T) = 0 \quad \Rightarrow \quad \theta = \frac{n}{n-T}$$

Por tanto, el estimador de máxima verosimilitud para θ viene dado por $\boxed{\widehat{\theta}_{\text{verosimilitud}} = \dfrac{n}{n - \sum_{i=1}^n \log X_i}}$.

Vamos a buscar ahora el estimador por el método de los momentos. Se tiene que el momento poblacional de primer orden viene dado por

$$E_\theta[X] = \int_0^1 \frac{\theta}{1-\theta} x^{\theta/(1-\theta)} dx = \theta [x^{1/(1-\theta)}]_0^1 = \theta$$

por tanto, según el método de los momentos, el estimador para θ viene dado por $\boxed{\widehat{\theta}_{\text{momentos}} = \overline{X}}$.

INFERENCIA ESTADÍSTICA
EXÁMENES

B) Tenemos que la función de densidad muestral viene dada por

$$f_\theta(x_1,\ldots,x_n)=\left(\frac{\theta}{1-\theta}\right)^n \exp\left\{\frac{2\theta-1}{1-\theta}\left(\sum_{i=1}^n \log x_i\right)\right\}=\left(\frac{\theta}{1-\theta}\right)^n \exp\left\{\frac{2\theta-1}{1-\theta}T\right\}$$

siendo $0<x_{(1)}\le x_{(n)}<1$. Por el Teorema de Factorización, se tiene que $T=\sum_{i=1}^n \log X_i$ es un estadístico suficiente.

Además, dadas dos muestras de tamaño n: $(x_1,\ldots,x_n)$ y $(x_1',\ldots,x_n')$, se tiene que

$$\frac{f_\theta(x_1,\ldots,x_n)}{f_\theta(x_1',\ldots,x_n')}=\exp\left\{\frac{2\theta-1}{1-\theta}(T-T')\right\}$$

es independiente de θ si y sólo si $T=T'$. Por tanto, se tiene que $T=\sum_{i=1}^n \log X_i$ es un estadístico suficiente minimal.

Como sabemos que $E_\theta[X]=\theta$, se tiene que $E_\theta[X_1]=\theta$. Por tanto, se tiene que X_1 es insesgado para θ. Se tiene entonces que $E_\theta[X_1 \mid T]$ es el ECUMV para θ, según el Teorema de Lehmann-Scheffé (ya que suponemos T completo).

Vamos a estudiar la distribución del estadístico T. Como sabemos, la función de densidad de la variable X viene dada por

$$f_\theta(x)=\frac{\theta}{1-\theta}x^{(2\theta-1)/(1-\theta)}=\frac{\theta}{1-\theta}\exp\left\{\frac{2\theta-1}{1-\theta}\log x\right\}$$

Por tanto, haciendo el cambio de variable $Y=-\log X$, se tiene que la función de densidad de Y viene dada por

$$f_\theta(y)=\frac{\theta}{1-\theta}\exp\left\{-\frac{2\theta-1}{1-\theta}y\right\}\exp\{-y\}=\frac{\theta}{1-\theta}\exp\left\{-\frac{\theta}{1-\theta}y\right\}$$

que es una distribución gamma $\gamma\left(1,\frac{\theta}{1-\theta}\right)$. Por tanto, se tiene que el estadístico $-T=-\sum_{i=1}^n \log X_i$ tiene distribución gamma $\gamma\left(n,\frac{\theta}{1-\theta}\right)$, debido a la propiedad reproductiva de esta distribución. Con este mismo argumento, se tiene también que el estadístico $-T'=-\sum_{i=2}^n \log X_i$ tiene distribución $\gamma\left(n-1,\frac{\theta}{1-\theta}\right)$. Con esto, podemos hallar la densidad del estadístico $X_1 \mid T$. Se tiene que

$$f_\theta(X_1=x \mid T=t)=\frac{f_\theta(x)f_\theta(T=t \mid X_1=x)}{f_\theta(T=t)}=\frac{f_\theta(X_1=x)f_\theta(-T'=-t+\log x)}{f_\theta(-T=-t)}=$$

$$=\frac{\left(\frac{\theta}{1-\theta}\right)x^{\frac{2\theta-1}{1-\theta}}\left(\frac{\theta}{1-\theta}\right)^{n-1}\frac{e^{\frac{\theta}{1-\theta}(t-\log x)}(-t+\log x)^{n-2}}{(n-2)!}}{\left(\frac{\theta}{1-\theta}\right)^n\frac{e^{\frac{\theta}{1-\theta}t}(-t)^{n-1}}{(n-1)!}}=\frac{(n-1)(-t+\log x)^{n-2}}{x(-t)^{n-1}}$$

Por tanto, se tiene que

$$E_\theta[X_1 \mid T=t] = \frac{n-1}{(-t)^{n-1}} \int_0^1 (-t+\log x)^{n-2} dx = \frac{n-1}{(-t)^{n-1}} \int_{-\infty}^0 (e^s - t)^{n-2} e^s ds =$$

$$= \frac{\left[(e^s-t)^{n-1}\right]_{-\infty}^0}{(-t)^{n-1}} = \frac{(1-t)^{n-1} - (-t)^{n-1}}{(-t)^{n-1}} = \left(1-\frac{1}{t}\right)^{n-1} - 1$$

Por tanto, tenemos que el estimador centrado uniformemente de mínima varianza para θ viene dado por

$$\boxed{H = \left(1-\frac{1}{T}\right)^{n-1} - 1} \qquad \text{siendo } T = \sum_{i=1}^n \log X_i$$

INFERENCIA ESTADÍSTICA
EXÁMENES

Ejercicio 2. Sea X una variable aleatoria con distribución $N(\theta,\theta)$ (es decir, $E[X]=\theta$, $\sqrt{V(X)}=\theta$). Determinar un estimador suficiente minimal para la familia anterior y analizar si el estimador determinado es completo.

Solución:

En primer lugar, aclaramos que debe ser $\theta>0$ para que tenga sentido la distribución del ejercicio. Se tiene entonces que la función de densidad de la variable aleatoria X viene dada por

$$f_\theta(x)=\frac{1}{\theta\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta)^2}{2\theta^2}\right\} \qquad x\in\mathbb{R}$$

Se tiene así que si tenemos una muestra aleatoria simple de tamaño n, entonces la función de densidad muestral viene dada por

$$f_\theta(x_1,\ldots,x_n)=\left(\frac{1}{\theta\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2\theta^2}\sum_{i=1}^n(x_i-\theta)^2\right\}=$$

$$=\left(\frac{1}{\theta\sqrt{2\pi}}\right)^n\exp\left\{\frac{\sum_{i=1}^n x_i}{\theta}-\frac{\sum_{i=1}^n x_i^2}{2\theta^2}-\frac{n}{2}\right\}$$

Por tanto, se tiene, por el Teorema de Factorización, que $T=\left(\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2\right)$ es un estadístico suficiente. Además, dadas dos muestras distintas de tamaño n: $(x_1,\ldots,x_n)$ y $(x_1',\ldots,x_n')$ se tiene que

$$\frac{f_\theta(x_1,\ldots,x_n)}{f_\theta(x_1',\ldots,x_n')}=\exp\left\{\frac{\left[\sum_{i=1}^n x_i-\sum_{i=1}^n x_i'\right]}{\theta}-\frac{\left[\sum_{i=1}^n x_i^2-\sum_{i=1}^n x_i'^2\right]}{2\theta^2}-\frac{n}{2}\right\}$$

que es independiente de θ si y sólo si $\sum_{i=1}^n x_i=\sum_{i=1}^n x_i'$ y $\sum_{i=1}^n x_i^2=\sum_{i=1}^n x_i'^2$. Es decir, si y sólo si $T=T'$. Por tanto, se tiene que $T=\left(\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2\right)$ es un estadístico suficiente minimal (también puede tomarse el estadístico $T'=\left(\sum_{i=1}^n X_i,\ \prod_{i=1}^n X_i\right)$ que es suficiente minimal).

Si observamos ahora la densidad muestral y llamamos $T_1=\sum_{i=1}^n X_i$, $T_2=\sum_{i=1}^n X_i^2$, $q_1(\theta)=\frac{1}{\theta}$ y $q_2(\theta)=-\frac{1}{2\theta^2}$, se tiene que

$$f_\theta(x_1,\ldots,x_n)=\left(\frac{1}{\theta\sqrt{2\pi}}\right)^n\exp\left\{q_1(\theta)T_1+q_2(\theta)T_2-\frac{n}{2}\right\}$$

que es una distribución de tipo exponencial. Como se tiene que $\mathrm{Im}(q_1,q_2)=((0,+\infty),(-\infty,0))$, que contiene abiertos de $\mathbb{R}^2$ y $T=(T_1,T_2)$ un estadístico minimal suficiente, se tiene directamente que el estadístico T es completo (ver página 273 del texto base).

INFERENCIA ESTADÍSTICA
EXÁMENES

Problema 3. Se quiere averiguar si existe independencia entre la semana de exámenes que elige el alumno para realizar las Pruebas Presenciales y la calificación obtenida en una determinada asignatura. Para ello se eligieron al azar 400 alumnos de la UNED que cursaban la asignatura en cuestión, obteniéndose la siguiente tabla de contingencia.

	Primera Semana	Segunda Semana
$[0,2)$	34	40
$[2,5)$	59	65
$[5,8)$	78	50
$[8,10)$	44	30

¿Qué conclusiones obtendría?

Solución:

Se trata de un test de Independencia de Caracteres. Para ello, tenemos que consideramos como hipótesis nula H_0 : "los caracteres son independientes" y como hipótesis alternativa H_1 : "los caracteres no son independientes".

Tenemos que, suponiendo cierta H_0, la tabla de frecuencias esperadas es

	Primera Semana	Segunda Semana
$[0,2)$	39'775	34'225
$[2,5)$	66'65	57'35
$[5,8)$	68'8	59'2
$[8,10)$	39'775	34'225

y el estadístico de Pearson viene dado por

$$\lambda = 7'341$$

Sabemos que, suponiendo cierta H_0, el estadístico de Pearson sigue una distribución χ^2_3. Por tanto, se tiene que el p-valor viene dado por

$$P\{\chi^2_3 > 7'341\} = 0'062$$

Por tanto, el p -valor aporta un valor que no permite rechazar la hipótesis nula, pero que tampoco la refuerza. Por ejemplo, tendríamos que con nivel de significación $\alpha = 0'1$ se rechazaría H_0 mientras que con nivel de significación $\alpha = 0'05$ sí se rechazaría H_0.

Curso 2012-2013

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria discreta con valores naturales y función de probabilidad

$$p_\theta(x) = \frac{-1}{\log(1-\theta)} \ \frac{\theta^x}{x} \qquad\qquad x = 1, 2, 3, ...$$

siendo $0 < \theta < 1$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

(a) ¿Coinciden el estimador de máxima verosimilitud de θ y el obtenido por el método de los momentos?

2.5 puntos

(b) Determinar el estimador centrado uniformemente de mínima varianza para

$$g(\theta) = \frac{\theta}{-(1-\theta)\log(1-\theta)}.$$

4 puntos

Problema 2

Rutherford y Geiger (1910) anotaron el número de destellos X que se producían en intervalos de 72 segundos, producidos por la desintegración del polonio. Los datos recogidos por ellos aparecen en la siguiente distribución de frecuencias absolutas,

X	0	1	2	3	4	5	6	7	8	9	10
n_i	57	203	383	525	532	408	273	139	45	27	16

Ajustar una distribución de Poisson a estos datos y analizar la bondad del ajuste utilizando las tablas de la distribución de Poisson de la Adenda y considerando un nivel de significación 0'05.

2.5 puntos

Ejercicio

¿Son iguales los contrastes de la χ^2 de Independencia de Caracteres y de Homogeneidad de Varias Muestras, desde un punto de vista conceptual y desde un punto de vista matemático? Razone la respuesta.

1 punto

Problema 1

a) La función de verosimilitud de la muestra (véase PIE, pág. 283) será

$$p_\theta(x_1,...,x_n)=\prod_{i=1}^{n}p_\theta(x_i)=\left(\frac{1}{-\log(1-\theta)}\right)^n\ \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i}.$$

Tomando logaritmos y derivando respecto al parámetro, obtenemos la siguiente ecuación de verosimilitud,

$$\frac{\partial}{\partial\theta}\log p_\theta(x_1,...,x_n)=\frac{n}{(1-\theta)\log(1-\theta)}+\frac{\sum_{i=1}^n x_i}{\theta}=0$$

de donde se deduce que el estimador de máxima verosimilitud, $\hat{\theta}$, debe cumplir la ecuación

$$\frac{\hat{\theta}}{(1-\hat{\theta})\left(-\log(1-\hat{\theta})\right)}=\overline{x}.$$

Por otro lado, al ser

$$E[X]=\sum_{x=1}^{\infty}x\ \frac{1}{-\log(1-\theta)}\ \frac{\theta^x}{x}=\frac{1}{-\log(1-\theta)}\left(\theta+\theta^2+\theta^3+...\right)=\frac{\theta}{(1-\theta)\left(-\log(1-\theta)\right)}$$

la ecuación que proporciona el estimador por el método de los momentos es

$$\frac{\hat{\theta}}{(1-\hat{\theta})\left(-\log(1-\hat{\theta})\right)}=\overline{x}$$

es decir, la misma que proporcionaba el estimador de máxima verosimilitud. Por tanto, ambos estimadores coinciden.

b) Como la distribución modelo es una familia de tipo exponencial (PIE, pág. 174) de la forma

$$p_\theta(x)=\frac{1}{-\log(1-\theta)}\ \frac{1}{x}\ e^{x\ \log\theta}$$

será

$$T(X_1,...,X_n)=\sum_{i=1}^{n}X_i$$

un estadístico minimal suficiente, resultado que también se podría obtener directamente a partir del teorema de factorización (PIE, pág. 167) y del análisis directo de minimalidad (PIE, pág. 171, teorema 5.2).

Para determinar el ECUMV de cualquier función del parámetro, lo primero que necesitamos es determinar un estadístico suficiente y completo para la familia de distribuciones $p_\theta(x)$. Al ser la distribución modelo una familia de tipo exponencial el estadístico $T = \sum_{i=1}^{n} X_i$ es suficiente minimal. Además (PIE, pág. 273), T será completo si la imagen de $q(\theta) = \log\theta$ contiene un abierto de $\mathbb{R}$; como es $0 < \theta < 1$ la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, T será suficiente y completo. Ahora sólo necesitamos determinar una función de T que sea insesgada para $g(\theta)$, ya que entonces, por el teorema de Lehmann-Scheffé (PIE, pág. 218), ese estadístico será el ECUMV buscado.

Por la linealidad de la esperanza, será

$$E[T] = E\left[\sum_{i=1}^{n} X_i\right] = nE[X] = n\,\frac{\theta}{(1-\theta)\,(-\log(1-\theta))}$$

con lo que el estimador insesgado para $g(\theta)$, función del suficiente y completo, será $T/n = \overline{x}$.

Problema 2

Se trata de un contraste de la chi-cuadrado de bondad del ajuste (PIE-sección 10.3 ó CB-sección 12.2). Como el parámetro de la distribución a ajustar no se conoce, debemos de estimarlo y sabemos que el mejor ajuste se va a obtener cuando el parámetro se estime por el método de la máxima verosimilitud. En el caso de la distribución de Poisson, el estimador de máxima verosimilitud es la media muestral (PIE-ejemplo 7.2 ó CB-ejemplo 5.3). De la distribución de frecuencias absolutas del enunciado se obtiene

X	n_i	$X \cdot n_i$
0	57	0
1	203	203
2	383	766
3	525	1575
4	532	2128
5	408	2040
6	273	1638
7	139	973
8	45	360
9	27	243
10	16	160
	2608	10086

Es decir, una media muestra igual a $\overline{x} = 10086/2608 = 3'867$.

La distribución de Poisson de la que tenemos que analizar su bondad de ajuste es, por tanto, la de parámetro $3'87$ que redondearemos en $3'8$ porque dice el enunciado que utilicemos las tablas de la Adenda (ADD). De la Tabla 2 de la distribución de Poisson obtenemos las siguientes probabilidades que esta distribución asigna a los valores X observados, a la que hemos añadido las frecuencias esperadas

X	n_i	$P\{X=x\}$	$2608 \cdot P\{X=x\}$
0	57	$0'0224$	$58'42$
1	203	$0'0850$	$221'68$
2	383	$0'1615$	$421'19$
3	525	$0'2046$	$533'60$
4	532	$0'1944$	$507'00$
5	408	$0'1477$	$385'20$
6	273	$0'0936$	$244'11$
7	139	$0'0508$	$132'49$
8	45	$0'0241$	$62'85$
9	27	$0'0102$	$26'60$
≥ 10	16	$0'0057$	$14'86$
	2608	1	2608

El estadístico λ de Pearson de bondad del ajuste seguirá una distribución χ^2 con 11-1-1 grados de libertad porque hemos estimado un parámetro. Tomará el valor

$$\lambda = \sum_i \frac{(n_i - np_i)^2}{np_i} = \sum_i \left(\frac{n_i^2}{np_i}\right) - n = 2624'69 - 2608 = 16'69$$

y el p-valor será, a partir de la Tabla 4 de la distribución χ^2

$$P\{\chi^2_9 > 16'69\} > P\{\chi^2_9 > 16'92\} = 0'05$$

por lo que a nivel $0'05$ debemos aceptar la hipótesis nula de que se ajusta bien esta distribución a los datos.

Ejercicio

Los contrastes de la χ^2 de homogeneidad de varias muestras e independencia de caracteres, coinciden desde el punto de vista técnico en el resultado final, pero no conceptualmente. Una de sus mayores diferencias consiste en que, en el de independencia de dos caracteres, se fija el tamaño de muestra n y se clasifican los n individuos seleccionados según las clases que forman las dos variables.

Por contra, en el de homogeneidad de k muestras, se fijan los tamaños muestrales marginales n_i, $i = 1, ..., k$, a seleccionar de cada una de las k poblaciones a comparar, siendo n, simplemente el resultado de sumar $n_1 + ... + n_k = n$. Véase PIE, página 445 y siguientes.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria con distribución de Poisson de parámetro λ. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

(a) Estimar λ por el método de la máxima verosimilitud.

1 punto

(b) El estimador de máxima verosimilitud obtenido en el apartado anterior, ¿es insesgado para λ? ¿Lo es la cuasivarianza muestral S^2? ¿Cuál de estos dos estimadores de λ es mejor desde el punto de vista de su varianza?

2 puntos

Problema 2

Sea X una variable aleatoria discreta con función de probabilidad bajo la hipótesis nula H_0

$$p_0(x) = \left(\frac{2}{3}\right)^x \ \frac{1}{3} \qquad x = 0, 1, 2, 3, ...$$

y con función de probabilidad

$$p_1(x) = \left(\frac{1}{2}\right)^{x+1} \qquad x = 0, 1, 2, 3, ...$$

bajo la hipótesis alternativa H_1. Determinar un contraste de máxima potencia de nivel α para contrastar H_0 frente a H_1, utilizando una muestra aleatoria simple de tamaño n de X. ¿Qué conclusión sacaría si, en una muestra de tamaño $n = 100$, se obtuvo una media muestral de $2'1$ y el nivel de significación era $\alpha = 0'05$?

4 puntos

Problema 3

Las coces de caballo fueron una causa de muerte tenida en cuenta en el ejército Prusiano. El ruso nacido en San Petersburgo, Ladislaus Josephovich Bortkiewicz, escribió un libro en alemán en 1898 sobre la distribución de Poisson en el que estudió el número de muertos X por esta causa entre los años 1875 y 1894, obteniendo la siguiente distribución de frecuencias absolutas (Preece, Ross y Kirby, 1988)

X	0	1	2	3	4	≥ 5
n_i	109	65	22	3	1	0

Ajustar una distribución de Poisson a estos datos y analizar la bondad del ajuste utilizando las tablas de la distribución de Poisson de la Adenda.

2 puntos

Ejercicio

Explique brevemente qué es y para qué se utiliza la noción de Completitud. Ponga algún ejemplo.

1 punto

Problema 1

a) La función de verosimilitud de la muestra (véase PIE, pág. 283) será

$$L(\lambda) = f_\lambda(x_1, ..., x_n) = \prod_{i=1}^{n} f_\lambda(x_i) = \frac{e^{-n\,\lambda}\ \lambda^{\sum x_i}}{\prod x_i!}$$

con lo que la derivada de su logaritmo, igualada a cero será

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

obteniéndose como estimador máximo-verosímil, la media muestral,

$$\hat{\lambda} = \overline{x}.$$

b) Claramente la media muestral es insesgado para λ, puesto que la esperanza de la media muestral es la media poblacional, en este caso, λ.

Respecto a la cuasivarianza muestral, su esperanza es (PIE, pág. 41) la varianza poblacional que en el caso de una población de Poisson es λ. Por tanto, tanto $\overline{x}$ como S^2 son insesgados para λ.

Al ser la distribución de Poisson una familia de tipo exponencial,

$$f_\lambda(x) = \frac{e^{-\lambda}\,\lambda^x}{x!} = \frac{e^{-\lambda}}{x!}\ \exp\{x\,\log\lambda\}$$

y contener la imagen de $q(\lambda) = \log\lambda$ un abierto de $\mathbb{R}$, claramente la media muestral es suficiente minimal y completo. Como es insesgado para λ, la media muestral es **el** ECUMV suyo. Por tanto, S^2 deber tener mayor varianza que $\overline{x}$, al ser ambos estimadores insesgados. De hecho, es

$$V(S^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1} > \frac{\lambda}{n} = V(\overline{x})$$

Problema 2

El lema de Neyman-Pearson nos dice (PIE, p g. 338) que el test de máxima potencia, para contrastar una hipótesis nula simple frente a una alternativa simple, como las aquí planteadas, tiene como región crítica la dada por

$$p_1(x_1, ..., x_n) > a\ p_0(x_1, ..., x_n)$$

es decir,

$$\left(\frac{1}{2}\right)^{\sum x_i + n} > a \left(\frac{2}{3}\right)^{\sum x_i} \left(\frac{1}{3}\right)^n$$

o lo que es lo mismo,

$$-0'2875 \sum_{i=1}^{n} x_i > d$$

o equivalentemente,

$$\sum_{i=1}^{n} x_i < c$$

con lo que el test de máxima potencia buscado será el test aleatorizado

$$\varphi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad \sum x_i < c \\ \gamma & \text{si} \quad \sum x_i = c \\ 0 & \text{si} \quad \sum x_i > c \end{cases}$$

siendo c la constante para la que el test tiene tamaño α; es decir, tal que

$$P_0\left\{\sum_{i=1}^{n} X_i < c\right\} + \gamma P_0\left\{\sum_{i=1}^{n} X_i = c\right\} = \alpha.$$

Para determinar la constante c necesitamos la distribución de $T = \sum_{i=1}^{n} X_i$ bajo H_0. Para ello, observamos que las funciones de probabilidad dadas en el enunciado corresponden a distribuciones binomiales negativas (véase LCP, pág. 293), $BN(1, 1/3)$ bajo la hipótesis nula H_0 y $BN(1, 1/2)$ bajo la hipótesis alternativa.

Como esta distribución de probabilidad es reproductiva respecto al primer parámetro, bajo la hipótesis nula será $T \rightsquigarrow BN(n, 1/3)$.

Si el tamaño muestral es suficientemente grande, al ser T una suma de variable aleatorias independientes e idénticamente distribuidas, podremos aplicar el teorema central del límite y utilizar un test no aleatorizado en el que se rechaza la hipótesis nula cuando y sólo cuando sea $T < c$ en donde c se calcula utilizando la aproximación normal,

$$\overline{x} = T/n \rightsquigarrow N\left(2, \sqrt{6/n}\right)$$

de donde, de la condición del nivel, será

$$\alpha = P\{T < c\} = P\{\overline{x} < k\} = P\left\{Z < (k-2)/\sqrt{6/n}\right\}$$

es decir,

$$k = 2 + z_{1-\alpha}\sqrt{\frac{6}{n}} = 2 - z_{\alpha}\sqrt{\frac{6}{n}}.$$

Para los datos del enunciado es $z_\alpha = z_{0'05} = 1'645$ y como es

$$\overline{x} = 2'1 > 1'597 = 2 - 1'645\sqrt{\frac{6}{100}}$$

aceptaremos la hipótesis nula.

Problema 3

Se trata de un contraste de la chi-cuadrado de bondad del ajuste (PIE-sección 10.3 ó CB-sección 12.2). Como el parámetro de la distribución a ajustar no se conoce, debemos de estimarlo y sabemos que el mejor ajuste se va a obtener cuando el parámetro se estime por el método de la máxima verosimilitud. En el caso de la distribución de Poisson, el estimador de máxima verosimilitud es la media muestral (PIE-ejemplo 7.2 ó CB-ejemplo 5.3). De la distribución de frecuencias absolutas del enunciado se obtiene

X	n_i	$X \cdot n_i$
0	109	0
1	65	65
2	22	44
3	3	9
4	1	4
≥ 5	0	0
	200	122

La media muestra es igual a $\overline{x} = 122/200 = 0'61$.

La distribución de Poisson de la que tenemos que analizar su bondad de ajuste es, por tanto, la de parámetro $0'61$ que redondearemos en $0'6$ porque dice el enunciado que utilicemos las tablas de la Adenda (ADD). De la Tabla 2 de la distribución de Poisson obtenemos las siguientes probabilidades que esta distribución asigna a los valores X observados, a la que hemos añadido las frecuencias esperadas

X	n_i	$P\{X = x\}$	$200 \cdot P\{X = x\}$
0	109	$0'5488$	$109'76$
1	65	$0'3293$	$65'86$
2	22	$0'0988$	$19'76$
3	3	$0'0198$	$3'96$
≥ 4	1	$0'0033$	$0'66$
	200	1	200

El estadístico λ de Pearson de bondad del ajuste seguirá una distribución χ^2 con 5-1-1 =3 grados de libertad porque hemos estimado un parámetro. Tomará el valor

$$\lambda = \sum_i \frac{(n_i - np_i)^2}{np_i} = \sum_i \left(\frac{n_i^2}{np_i}\right) - n = 200'678 - 200 = 0'678$$

y el p-valor $P\{\chi_3^2 > 0'678\}$, estará, a partir de la Tabla 4 de la distribución χ^2, entre los valores

$$P\{\chi_3^2 > 1'424\} < P\{\chi_3^2 > 0'678\} < P\{\chi_3^2 > 0'584\}$$

es decir, entre

$$0'7 < P\{\chi_3^2 > 0'678\} < 0'9$$

suficientemente grande como para aceptar la hipótesis nula. Es cierto que las últimas frecuencias esperadas no son mayores que 5, pero es tan claro es resultado que no sería necesario agrupar estas últimas clases.

Ejercicio

La completitud se introduce, fundamentalmente, para asegurar la unicidad del estimador centrado, función de un suficiente. Véase PIE, página 216 y siguientes.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2012-2013.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta^3}{2}\,(x-2)^2\,e^{-\theta\,x}\,e^{2\,\theta} \qquad\qquad x > 2$$

siendo $\theta > 0$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
(a) Estimar θ por el método de la máxima verosimilitud y por el método de los momentos.

1.5 puntos

(b) Determinar el estimador centrado uniformemente de mínima varianza para $g_1(\theta) = 1/\theta$ y para $g_2(\theta) = \theta$.

2 puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con función de densidad $f_0(x) = 1\,,\ 0 < x < 1$ bajo la hipótesis nula H_0 y con función de densidad $f_1(x) = 2\,x\,,\ 0 < x < 1\,,$ bajo la hipótesis alternativa H_1. Determinar un contraste de máxima potencia de nivel α para contrastar H_0 frente a H_1, utilizando una muestra aleatoria simple de tamaño n de X. Calcular la potencia del contraste obtenido si es $\alpha = 0'05$ y $n = 3$.

3 puntos

Problema 3

En el año 1909, el famoso estadístico Karl Pearson estudió la posible relación entre el tipo de delito cometido y el consumo de alcohol, anotando el número de personas condenadas por los crímenes de la siguiente tabla y si el delincuente era abstemio o consumidor habitual de bebidas alcohólicas. Los resultados obtenidos por Pearson fueron los siguientes:

Tipo de delito	Bebedor	Abstemio
Incendio provocado	**50**	**43**
Violación	**88**	**62**
Violencia callejera	**155**	**110**
Robo	**379**	**300**
Falsificación de moneda	**18**	**14**
Estafa	**63**	**144**

Utilizar el test inventado por el autor del estudio para analizar si existía o no relación entre ambas características. 2.5 puntos

Ejercicio

Diga en qué consiste el Método de los Momentos y ponga algún ejemplo de su aplicación.

1 punto

Problema 1

a) La función de verosimilitud de la muestra (véase PIE, pág. 283) será

$$f_\theta(x_1,...,x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{\theta^{3n}}{2^n} \left(\prod_{i=1}^{n} (x_i - 2)^2 \right) e^{2n\theta} \exp\{-\theta \sum_{i=1}^{n} x_i\}.$$

Tomando logaritmos y derivando respecto al parámetro, obtenemos la siguiente ecuación de verosimilitud,

$$\frac{\partial}{\partial \theta} \log f_\theta(x_1,...,x_n) = \frac{3n}{\theta} - \sum_{i=1}^{n} x_i + 2n = 0$$

de donde se deduce que el estimador de máxima verosimilitud, $\hat{\theta}$, será

$$\hat{\theta} = \frac{3}{\overline{x} - 2}.$$

Por otro lado, al ser $X = Y + 2$ con $Y \rightsquigarrow \gamma(3, \theta)$, será

$$E[X] = E[Y + 2] = E[Y] + 2 = \frac{3}{\theta} + 2$$

(resultado que se podría obtener también directamente sin conocer la relación de X con Y), con lo que la ecuación que proporciona el estimador por el método de los momentos será

$$\frac{3}{\theta} + 2 = \overline{x}$$

de donde se obtiene el estimador de θ por el método de los momentos,

$$\hat{\theta} = \frac{3}{\overline{x} - 2}$$

es decir, el de máxima verosimilitud.

b) Como la distribución modelo es una familia de tipo exponencial (PIE, pág. 174) de la forma

$$f_\theta(x) = \frac{\theta^3}{2} e^{2\theta} (x-2)^2 e^{-\theta x}$$

será $S(X_1,...,X_n) = \sum_{i=1}^{n} X_i$ un estadístico minimal suficiente. Por tanto, (PIE, pág. 273), S será además completo si la imagen de $q(\theta) = -\theta$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$, la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$. En definitiva, S será suficiente y completo.

Basta ahora con encontrar un estimador, función del suficiente y completo S, que sea insesgado para

$$g_1(\theta) = \frac{1}{\theta}$$

y ese será el ECUMV para $g_1(\theta)$. Comencemos calculando la esperanza de S.

$$E[S(X_1, ..., X_n)] = E\left[\sum_{i=1}^{n} X_i\right] = n\,E[X] = n\left(\frac{3}{\theta} + 2\right).$$

Por tanto, el ECUMV para $g_1(\theta) = 1/\theta$ será

$$\frac{1}{3}\left(\frac{S}{n} - 2\right) = \frac{1}{3}\left(\overline{x} - 2\right).$$

Tanto el estimador de máxima verosimilitud, como el de los momentos, como el resultado anterior, sugieren que

$$\hat{\theta} = \frac{3}{\overline{x} - 2} = \frac{3n}{\sum_{i=1}^{n} X_i - 2n}$$

sea el ECUMV para $g_2(\theta) = \theta$. Sabemos que es función del suficiente y completo. Calculemos su esperanza.

Como es $X - 2 \rightsquigarrow \gamma(3, \theta)$, al ser esta distribución reproductiva respecto al primer parámetro (LCP, pág. 311), será

$$\sum_{i=1}^{n} X_i - 2n = \sum_{i=1}^{n} (X_i - 2) \rightsquigarrow \gamma(3n, \theta)$$

por lo que

$$E[\hat{\theta}] = 3n \int \frac{1}{y} \frac{1}{\Gamma(3n)}\, \theta^{3n}\, e^{-\theta y}\, y^{3n-1}\, dy =$$

$$= 3n\,\frac{\theta}{3n-1} \int \frac{1}{\Gamma(3n-1)}\, \theta^{3n-1}\, e^{-\theta y}\, y^{(3n-1)-1}\, dy = \frac{3n\theta}{3n-1}$$

que conduce a que el ECUMV para $g_2(\theta) = \theta$ sea

$$\frac{3n-1}{\sum_{i=1}^{n} X_i - 2n}.$$

Problema 2

El lema de Neyman-Pearson nos dice (PIE, pág. 338) que el test de máxima potencia, para contrastar una hipótesis nula simple frente a una alternativa simple, como las aquí planteadas, tiene por región crítica la siguiente

$$f_1(x_1, ..., x_n) > k f_0(x_1, ..., x_n)$$

es decir, si $0 < x_1, ..., x_n < 1$,

$$2^n \prod_{i=1}^{n} x_i > k$$

o lo que es lo mismo,

$$T_n = \sum_{i=1}^{n} \log x_i > c$$

en donde la constante c es tal que

$$P_{H_0}\{T_n > c\} = \alpha.$$

Como, bajo H_0, es $X \rightsquigarrow U(0,1)$, un sencillo cambio de variable permite determinar que la función de densidad de la variable $Y = -\log X$ es

$$g(y) = e^{-y} \qquad y > 0$$

es decir, una gamma $\gamma(1,1)$, por lo que, al ser reproductiva respecto al primer parámetro, será

$$\alpha = P_{H_0}\{T_n > c\} = P_{H_0}\{\sum_{i=1}^{n} \log X_i > c\} = P_{H_0}\{\sum_{i=1}^{n} -\log X_i < -c\} =$$

$$= P\{\gamma(n,1) < -c\} = P\{\chi^2_{2n} < -2c\}$$

en donde la última igualdad se obtiene por la conocida propiedad de que si es $W \rightsquigarrow \gamma(p,a)$ entonces es $2aW \rightsquigarrow \chi^2_{2p}$.

Por tanto, será

$$c = -0'5\, \chi^2_{2n;1-\alpha}$$

Para calcular la potencia del test, debemos determinar la distribución del estadístico de contraste, ahora bajo la hipótesis alternativa. Bajo H_1, es $Y = -\log X \rightsquigarrow \gamma(1,2)$ por lo que

$$\sum_{i=1}^{n} -\log X_i \rightsquigarrow \gamma(n,2).$$

La potencia del test será, por tanto,

$$P_{H_1}\{T_n > -0'5\,\chi^2_{2n;1-\alpha}\} = P_{H_1}\left\{\sum_{i=1}^{n} -\log X_i < 0'5\,\chi^2_{2n;1-\alpha}\right\} = P\{\gamma(n,2) < 0'5\,\chi^2_{2n;1-\alpha}\} =$$

$$= P\{\chi^2_{2n} < 2\,\chi^2_{2n;1-\alpha}\} = P\{\chi^2_6 < 2\,\chi^2_{6;0'95}\} = P\{\chi^2_6 < 3'27\} \approx 0'3.$$

Problema 3

Se trataría de un test de chi-cuadrado de independencia de caracteres (PIE-sección 10.4 ó CB-sección 12.4) en donde la hipótesis nula es que ambas variables son independientes frente a la alternativa de que están relacionadas.

El estadístico de contraste toma el valor

$$\lambda = \sum_{i=1}^{6}\sum_{j=1}^{2} \frac{(n_{ij} - n\,p_i\,q_j)^2}{n\,p_i\,q_j} = 49'73$$

siendo, por tanto, el p-valor del test

$$P\{\chi^2_5 > 49'73\} \approx 0$$

lo que indica un claro rechazo de la hipótesis nula de independencia entre ambas características.

Ejercicio

El Método de los momentos consiste en igualar los primeros momentos respecto a la media poblacionales con los correspondientes muestrales, hasta conseguir un número suficiente de ecuaciones de donde despejar los parámetros. Véase PIE, páginas 279 y siguientes.

– –

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

PIE: **Principios de Inferencia Estadística**, 1993. R. Vélez y A. García Pérez. Editorial UNED.

CB: **Estadística Aplicada: Conceptos Básicos** (Segunda Edición), 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código:0184011EP01A02).

ADD: **Fórmulas y Tablas Estadísticas** (1998). Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

Curso 2013-2014

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2013-2014.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Dada una muestra aleatoria simple de tamaño n de una distribución con función de densidad

$$f_\theta(x) = \frac{1}{2\sqrt{\theta}}\, x^{-3/2} \qquad\qquad x \geq \frac{1}{\theta}$$

siendo θ un parámetro positivo desconocido, se pide:

a) Determinar el estimador de máxima verosimilitud de θ y estudiar si es suficiente.

2 puntos

b) Determinar, por el método de la cantidad pivotal, el intervalo de confianza para θ de nivel de confianza $1-\alpha$ que tenga longitud mínima.

2 puntos

c) Determinar un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

2.5 puntos

Problema 2

En un estudio realizado por Sterling y Weinkam (1976) se clasificó a 31.315 trabajadoras estadounidenses según su Tipo de Trabajo y sus Hábitos Fumadores. Como clases para el Tipo de Trabajo se consideraron: Trabajadoras Manuales, TM (denominadas en USA *blue-collar workers*), Profesionales, P (*white-collar workers*) y Otras, O. Y, como clases de los Hábitos Fumadores se consideraron las clases: 20 o más cigarrillos al día, entre 10 y 19, entre 1 y 9, Antigua fumadora, AF, y Nunca había fumado, N. Los datos obtenidos fueron los dados por la siguiente tabla:

	TM	P	O
20+	3947	555	4449
10-19	1106	150	1333
1-9	607	119	846
AF	2629	936	4944
N	2937	1105	5652

A la vista de estos datos, ¿cree que hay una relación significativa entre el tipo de trabajo y los hábitos fumadores?

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Esto tiene especial sentido cuando hablamos de la resolución con el paquete R de un ejercicio, dado que en examen no se puede utilizar un ordenador.

Problema 1

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n) = \frac{1}{2^n}\,\frac{1}{\theta^{n/2}}\,\prod_{i=1}^{n} x_i^{-3/2}\, I_{\{\theta \geq 1/x_{(1)}\}}$$

la cual no es derivable respecto a θ. Sin embargo, es claro que se anula para $\theta < 1/x_{(1)}$ y que decrece a medida que θ avanza por el intervalo $[1/x_{(1)}, \infty)$. El estimador de máxima verosimilitud será, por tanto,

$$\hat{\theta} = \frac{1}{X_{(1)}}.$$

Como la función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1,...,x_n) = h(x_1,...,x_n)\cdot g_\theta(T(x_1,...,x_n))$$

con

$$h(x_1,...,x_n) = \frac{1}{2^n}\,\prod_{i=1}^{n} x_i^{-3/2}$$

y

$$g_\theta(T(x_1,...,x_n)) = \frac{1}{\theta^{n/2}}\, I_{\{\theta \geq 1/x_{(1)}\}}$$

es decir, mediante el producto de una función que no depende del parámetro, h, y una función que depende del parámetro y de la muestra sólo a través del estadístico, $X_{(1)}$, por el teorema de factorización (*PIE*, pag. 167), éste será suficiente para la familia y, por tanto, también el estimador de máxima verosimilitud antes determinado.

b) Con objeto de obtener una cantidad pivotal utilizaremos el estadístico $X_{(1)}$. Como la función de distribución poblacional es

$$F_\theta(x) = \begin{cases} 0 & \text{si} \quad x < 1/\theta \\ 1 - \dfrac{1}{\sqrt{\theta\, x}} & \text{si} \quad x \geq 1/\theta \end{cases}$$

la función de distribución del mínimo será

$$G_\theta(y) = 1 - (1 - F_\theta(y))^n = \begin{cases} 0 & \text{si} \quad y < 1/\theta \\ 1 - \dfrac{1}{(\theta\, y)^{n/2}} & \text{si} \quad y \geq 1/\theta \end{cases}$$

Por tanto, siguiendo el mismo razonamiento que se sigue en *PIE* pag. 97, analizaremos si $Z = \theta\, X_{(1)}$ es una cantidad pivotal. Su función de distribución es

$$\mathcal{P}_\theta \left\{ X_{(1)} \leq z/\theta \right\} = G_\theta(z/\theta) = \begin{cases} 0 & \text{si} \quad z < 1 \\ 1 - \dfrac{1}{z^{n/2}} & \text{si} \quad z \geq 1 \end{cases}$$

es decir, no dependiente de θ. Por tanto, $\theta\, X_{(1)}$ es una cantidad pivotal y fijado el nivel de confianza $1-\alpha$, si $\alpha_1, \alpha_2 > 0$ cumplen $\alpha_1 + \alpha_2 = \alpha$ de la distribución anterior se obtiene que

$$P\left\{ \frac{1}{(1-\alpha_1)^{2/n}} < \theta\, X_{(1)} < \frac{1}{\alpha_2^{2/n}} \right\} = 1 - \alpha_1 - \alpha_2 = 1 - \alpha.$$

es decir,

$$P\left\{ \frac{1}{X_{(1)}\,(1-\alpha_1)^{2/n}} < \theta < \frac{1}{X_{(1)}\,\alpha_2^{2/n}} \right\} = 1 - \alpha$$

y, por tanto,

$$\left[\frac{1}{X_{(1)}\,(1-\alpha_1)^{2/n}} \,,\; \frac{1}{X_{(1)}\,\alpha_2^{2/n}} \right]$$

ser un intervalo de confianza para θ, de nivel de confianza $1 - \alpha$, cualquiera que sean $\alpha_1, \alpha_2 > 0$ con $\alpha_1 + \alpha_2 = \alpha$.

Como su longitud

$$Long(\alpha_1) = \frac{1}{X_{(1)}} \left[\frac{1}{\alpha_2^{2/n}} - \frac{1}{(1-\alpha_1)^{2/n}} \right] = \frac{1}{X_{(1)}} \left[\frac{1}{(\alpha - \alpha_1)^{2/n}} - \frac{1}{(1-\alpha_1)^{2/n}} \right]$$

es una función creciente de α_1, la longitud mínima se obtiene para $\alpha_1 = 0$, $\alpha_2 = \alpha$. Por tanto, el intervalo de confianza de longitud mínima buscado será

$$\left[\frac{1}{X_{(1)}} \,,\; \frac{1}{X_{(1)}\,\alpha^{2/n}} \right].$$

c) Para $\theta < \theta'$, la razón de verosimilitudes

$$\frac{f_{\theta'}(x_1,...,x_n)}{f_{\theta}(x_1,...,x_n)} = \left(\frac{\theta}{\theta'}\right)^{n/2} \frac{I_{\{x_{(1)}\geq 1/\theta'\}}}{I_{\{x_{(1)}\geq 1/\theta\}}} = \left(\frac{\theta}{\theta'}\right)^{n/2} \frac{I_{\{-x_{(1)}\leq -1/\theta'\}}}{I_{\{-x_{(1)}\leq -1/\theta\}}}$$

sólo está definida para $-x_{(1)} \leq -1/\theta'$; pero es función creciente de $-x_{(1)}$, puesto que toma el valor constante $(\theta/\theta')^{n/2}$ si $-x_{(1)} \leq -1/\theta$ y vale $+\infty$ para $-x_{(1)} \in (-1/\theta\,, -1/\theta']$. El teorema de Karlin-Rubin, (*PIE*, pag. 349) nos garantiza la existencia de un contraste uniformemente de máxima potencia, ϕ, para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual tiene región crítica de la forma $-X_{(1)} > c$, en donde c se elige para que ϕ tenga tamaño α; es decir, tal que

$$\alpha = \mathcal{P}_{\theta_0}\left\{X_{(1)} < -c\right\} = 1 - \frac{1}{(-c\,\theta_0)^{n/2}}$$

de donde se obtiene el valor

$$c = \frac{-1}{\theta_0(1-\alpha)^{2/n}}$$

El contraste buscado será, por tanto,

$$\phi(x_1,...,x_n) = \begin{cases} 1 & \text{si} \quad x_{(1)} < \dfrac{1}{\theta_0(1-\alpha)^{2/n}} \\ 0 & \text{si} \quad x_{(1)} \geq \dfrac{1}{\theta_0(1-\alpha)^{2/n}} \end{cases}$$

Problema 2

Nos piden un test de la χ^2 de *independencia de caracteres* (CB-sección 12.4), en donde la hipótesis nula que se establece es que ambas variables son independientes. Esta hipótesis nula se rechazará cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n_{i.}\,m_{.j}/n)^2}{n_{i.}\,m_{.j}/n}$$

el estadístico de Pearson. Si queremos resolverlo con R (EAR-sección 7.4), ejecutaríamos la siguiente secuencia de instrucciones. Con `(1)` incluimos los

datos, que tienen que venir en forma de matriz. Recordemos que, por defecto, los incorpora por columnas. Las sentencias (2) y (3) son opcionales y sirven para poner nombre a las filas y a las columnas de la tabla. Con (4) comprobamos que hemos incorporado bien los datos a R. Ejecutando (5) es como le pedimos que haga el test χ^2.

```
> habitos<-matrix(c(3947,1106,607,2629,2937,555,150,119,936,1105,4449,1333,
+ 846,4944,5652),ncol=3)                                                    (1)
> colnames(habitos)<-c("TM","P","O")                                        (2)
> rownames(habitos)<-c("20 o más","10 a 19","1 a 9","AF","N")               (3)

> habitos                                                                   (4)
           TM    P    O
20 o más 3947  555 4449
10 a 19  1106  150 1333
1 a 9     607  119  846
AF       2629  936 4944
N        2937 1105 5652

> chisq.test(habitos)                                                       (5)

        Pearson's Chi-squared test

data:  habitos
X-squared = 641.5037, df = 8, p-value < 2.2e-16                             (6)
```

En (6) vemos el valor del estadístico de Pearson, $\lambda = 641'5$ y el p-valor del test es muy pequeño, suficiente como para concluir que puede rechazarse la hipótesis nula de independencia de ambas variables. Es decir, puede concluirse que sí existe una relación significativa entre los hábitos fumadores de las mujeres y el tipo de trabajo que desempeñan.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2013-2014.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{2\,x}{\theta^2} \qquad \text{si} \quad 0 < x < \theta$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar un estadístico suficiente minimal para la familia de densidades dada. ¿Es completo?

$1'5$ puntos

b) Determinar, si existe, el estimador centrado uniformemente de mínima varianza para $g_1(\theta) = \theta$ y para $g_2(\theta) = 1/\theta$.

$1'5$ puntos

c) Determinar el estimador de máxima verosimilitud de θ.

1 punto

Problema 2

Supongamos que se toma una muestra aleatoria simple de tamaño uno de la distribución con densidad

$$f_\theta(x) = 2\,\theta\,x + 2\,(1-\theta)\,(1-x) \qquad \text{si} \quad 0 < x < 1$$

siendo $\theta > 0$. Se pide:

a) Determinar un contraste de máxima potencia de nivel α para contratar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, siendo $\theta_1 < \theta_0$.

$1'5$ puntos

b) El contraste determinado en el apartado anterior, ¿es uniformemente de máxima potencia para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta < \theta_0$? ¿Lo es para contrastar $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$?

$1'5$ puntos

Problema 3

En un artículo de Price et al. (1998) se establece cómo mejorar la calidad de los sustratos de CdZnTe usados para producir un sensor en los montajes de obleas en fábricas de productos electrónicos. La longitud de onda de corte en μm de 11 obleas fueron las siguientes:

6′06	**6′16**	**6′57**	**6′67**	**6′98**	**6′17**	**6′17**	**6′93**	**6′73**	**6′87**	**6′76**

Con estos datos obtenidos y suponiendo que la longitud de onda de corte siga una distribución normal, ¿hay evidencia para concluir que la media de la longitud de onda de corte sea distinta de 6′5 μm?

3 puntos

Problema 1

a) La densidad de la muestra es

$$f_\theta(x_1,...,x_n) = \left(\frac{2}{\theta^2}\right)^n I_{\{x_{(n)}<\theta\}} I_{\{x_{(1)}>0\}} \prod_{i=1}^n x_i.$$

Como ésta se puede factorizar de la forma

$$f_\theta(x_1,...,x_n) = h(x_1,...,x_n) \cdot g_\theta(T(x_1,...,x_n))$$

con

$$h(x_1,...,x_n) = I_{\{x_{(1)}>0\}} \prod_{i=1}^n x_i$$

y

$$g_\theta(T(x_1,...,x_n)) = \left(\frac{2}{\theta}\right)^n I_{\{x_{(n)}<\theta\}}$$

por el teorema de factorización (véase *PIE*, pag. 167) un estadístico suficiente será $T = X_{(n)}$.

Como es

$$\frac{f_\theta(x_1,...,x_n)}{f_\theta(x'_1,...,x'_n)} = \left(\prod_{i=1}^n \frac{x_i}{x'_i}\right) \cdot \frac{I_{\{x_{(n)}<\theta\}}}{I_{\{x'_{(n)}<\theta\}}} \cdot \frac{I_{\{x_{(1)}>0\}}}{I_{\{x'_{(1)}>0\}}}$$

este cociente no depender de θ cuando y sólo cuando sea $X_{(n)} = X'_{(n)}$, con lo que $X_{(n)}$ será suficiente minimal (*PIE*, pag. 170 y siguientes).

La completitud de $X_{(n)}$ viene explicada y probada en el Ejemplo 6.5 de *PIE*, pag. 217 ($k = 2$).

b) Para determinar el ECUMV de $g_1(\theta) = \theta$, buscaremos un estimador insesgado de θ función del suficiente y completo $X_{(n)}$ (Teorema de Lehmann-Scheffé, *PIE*, pag. 218).

Como es $E[X_{(n)}] = 2n/(2n+1)\theta$, el ECUMV para θ será $(2n+1)/(2n)X_{(n)}$ al ser insesgado para el parámetro y función del suficiente y completo.

Respecto al ECUMV de $g_2(\theta) = 1/\theta$, es razonable pensar en $1/X_{(n)}$. Como es $E[1/X_{(n)}] = 2n/[(2n-1)\theta]$, el estimador $(2n-1)/[2nX_{(n)}]$ será el ECUMV para $1/\theta$.

c) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n) = \left(\frac{2}{\theta^2}\right)^n I_{\{x_{(1)}>0\}}\, I_{\{x_{(n)}<\theta\}} \prod_{i=1}^{n} x_i$$

la cual no es derivable respecto a θ. Sin embargo, es claro que se anula para $\theta < x_{(n)}$ y que decrece a medida que θ avanza por el intervalo $[x_{(n)}, \infty)$. El estimador de máxima verosimilitud será, por tanto,

$$\hat{\theta} = X_{(n)}.$$

Problema 2

a) El lema de Neyman-Pearson nos dice (*PIE*, pag. 338) que el test de máxima potencia tiene como región crítica la dada por

$$f_{\theta_1}(x_1,...,x_n) > k f_{\theta_0}(x_1,...,x_n)$$

es decir, supuesto que $0 < x < 1$

$$\frac{2\theta_1 x + 2(1-\theta_1)(1-x)}{2\theta_0 x + 2(1-\theta_0)(1-x)} > k.$$

Como la función

$$h(x) = \frac{\theta_1 x + (1-\theta_1)(1-x)}{\theta_0 x + (1-\theta_0)(1-x)}$$

es decreciente (por ser $\theta_1 < \theta_0$), será $h(x) > k$ cuando y sólo cuando sea $x < c$.

Por tanto la región crítica será la dada por

$$x < c$$

siendo c la constante para la que el test tiene tamaño α; es decir, tal que

$$\mathcal{P}_{\theta=\theta_0}\{X < c\} = \alpha$$

o bien,

$$\alpha = \mathcal{P}_{\theta=\theta_0}\{X < c\} = \int_0^c [2\theta_0 x + 2(1-\theta_0)(1-x)]\, dx = (2\theta_0 - 1)c^2 + 2(1-\theta_0)c$$

es decir,

$$c = \frac{1 - \theta_0 - \sqrt{(1-\theta_0)^2 + (2\theta_0 - 1)\alpha}}{1 - 2\theta_0}$$

b) Como el contraste ϕ obtenido en el apartado anterior no depende del valor de la alternativa θ_1, sino simplemente de que sea menor que θ_0 (es necesario el signo de su diferencia con θ_0 en la determinación de la región crítica), será (*PIE*, pag. 348) uniformemente de máxima potencia para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta < \theta_0$.

La función de potencia del test

$$\beta_\phi(\theta) = \mathcal{P}_\theta\{X < c\} = \theta[c^2 + (1-c)^2 - 1] + [1 - (1-c)^2]$$

es decreciente, por lo que (*PIE*, pag. 348) si era $\beta_\phi(\theta_0) = \alpha$, se cumplirá también la restricción $\beta_\phi(\theta) \leq \alpha$ para $\theta > \theta_0$, por lo que será uniformemente de máxima potencia para contrastar $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$.

Problema 3

Estamos en un caso de un contraste para la media de una población normal de varianza desconocida (CB-sección 7.2) en donde rechazaremos la hipótesis nula de ser $H_0 : \mu = 6'5$ y aceptaremos la hipótesis alternativa $H_0 : \mu \neq 6'5$ cuando y sólo cuando sea

$$\frac{|\overline{x} - 6'5|}{S/\sqrt{11}} > t_{n-1;\alpha/2}$$

Como es $\overline{x} = 6'55$ y $S = 0'347$ será

$$\frac{|\overline{x} - 6'5|}{S/\sqrt{11}} = 0'4779.$$

El p-valor del test será $2 \cdot P\{t_{10} > 0'4779\}$ y, a partir de la Tabla 5 de la Adenda AD, será

$$2 \cdot 0'3 < 2 \cdot P\{t_{10} > 0'4779\} < 2 \cdot 0'4$$

es decir, un p-valore entre $0'6$ y $0'8$, suficientemente grande como para aceptar la hipótesis nula.

Si queremos hacer este ejercicio con R, ejecutaríamos (EAR-sección 4.2)

```
> x<-c(6.06,6.16,6.57,6.67,6.98,6.17,6.17,6.93,6.73,6.87,6.76)

>  t.test(x,mu=6.5)
        One Sample t-test
data:  x
t = 0.4953, df = 10, p-value = 0.6311
alternative hypothesis: true mean is not equal to 6.5
95 percent confidence interval:
 6.318713 6.784924
sample estimates:
```

```
mean of x
 6.551818
```

obteniendo, lógicamente, las mismas conclusiones.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2013-2014.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta}{e^{\theta^2} - 1}\, e^{\theta x} \qquad \text{si} \quad 0 \leq x \leq \theta$$

siendo $\theta > 0$. Determinar, utilizando una muestra aleatoria simple de tamaño n de X, un estadístico suficiente minimal para la familia de densidades dada.

$1'5$ puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = e^{-(x-\theta)} \cdot \exp\{-e^{-(x-\theta)}\} \qquad -\infty < x < +\infty$$

siendo $\theta \in \mathbb{R}$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar la cota de Frechet-Cramer-Rao para los estimadores insesgados de θ.

2 puntos

b) Determinar, si existe, el estimador centrado uniformemente de mínima varianza para $g(\theta) = \theta$.

3 puntos

Problema 3

En Graham y Shotz (1979) aparecen los siguientes datos sobre cáncer de cuello de útero en donde los individuos elegidos al azar eran mujeres entre 50 y 59 años las cuales se clasificaron en dos grupos dependiendo de si padecían o no la mencionada enfermedad y de la Edad que tenían en el primer embarazo, con objeto de analizar la incidencia de esta enfermedad en mujeres con primer embarazo tardío.

	Cáncer de cuello de útero	
	SÍ	**NO**
Edad al primer embarazo		
Menor o igual que 25	**42**	**203**
Mayor que 25	**7**	**114**

A la vista de estos datos, ¿cabe pensar en una homogeneidad de ambos grupos de edades al primer embarazo en esta enfermedad?

3′5 puntos

Problema 1

La densidad de la muestra es

$$f_\theta(x_1,...,x_n) = \left(\frac{\theta}{e^{\theta^2}-1}\right)^n e^{\theta\sum_{i=1}^n x_i} I_{\{x_{(n)}<\theta\}} I_{\{x_{(1)}>0\}}.$$

Como ésta se puede factorizar de la forma

$$f_\theta(x_1,...,x_n) = h(x_1,...,x_n)\cdot g_\theta(T(x_1,...,x_n))$$

con

$$h(x_1,...,x_n) = I_{\{x_{(1)}>0\}}$$

y

$$g_\theta(T(x_1,...,x_n)) = \left(\frac{\theta}{e^{\theta^2}-1}\right)^n e^{\theta\sum_{i=1}^n x_i} I_{\{x_{(n)}<\theta\}}$$

por el teorema de factorizaciín (*PIE*, pag. 167) un estadístico suficiente será $T = (\sum_{i=1}^n X_i\,,\,\text{máx}\,X_i)$.

Como es

$$\frac{f_\theta(x_1,...,x_n)}{f_\theta(x'_1,...,x'_n)} = e^{\theta(\sum_{i=1}^n x_i - \sum_{i=1}^n x'_i)} \frac{I_{\{x_{(n)}<\theta\}}}{I_{\{x'_{(n)}<\theta\}}} \frac{I_{\{x_{(1)}>0\}}}{I_{\{x'_{(1)}>0\}}}$$

este cociente no dependerá de θ cuando y sólo cuando sea $\sum_{i=1}^n x_i = \sum_{i=1}^n x'_i$ y $X_{(n)} = X'_{(n)}$, con lo que T será suficiente minimal (*PIE*, pag. 170 y siguientes).

Problema 2

a) La cota de Frechet-Cramer-Rao para los estimadores insesgados T de θ será (*PIE*, pag. 225) $V_\theta(T) \geq 1/I(\theta)$, en donde la cantidad de información de Fisher $I(\theta)$ es

$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X_1,...,X_n)\right)^2\right]$$

igual a (*PIE*, pag. 231)

$$I(\theta) = ni(\theta)$$

en donde $i(\theta)$ es la información de Fisher de una muestra de tamaño 1.

Como en nuestro caso es

$$i(\theta) = E\left[\left(-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right)\right] = e^\theta E_\theta[e^{-X}] = e^\theta\, e^{-\theta} = 1$$

será en definitiva la cota de Frechet-Cramer-Rao buscada igual a

$$V_\theta(T) \geq 1/n.$$

b) Como la distribución dada es una familia exponencial de la forma

$$f_\theta(x) = e^\theta e^{-x} \exp\left\{e^{-x}(-e^\theta)\right\}$$

el estadístico $T = \sum_{i=1}^n e^{-X_i}$ es un estadístico suficiente y completo (y tambiín minimal) (*PIE*, pags. 174 y 273.)

Una manera de determinar el ECUMV para θ es buscar un estimador insesgado para θ que sea función del suficiente y completo T. Parece natural probar con el mismo T. No obstante, es

$$E[T] = nE[e^{-X}] = n\, e^{-\theta}$$

con lo que parece lógico intentarlo con $\log T$, por lo que determinaremos su función característica:

$$\varphi_T(u) = E[e^{iuT}] = \left(E_X\left[e^{iue^{-X}}\right]\right)^n = \left(1 - \frac{iu}{e^\theta}\right)^{-n}$$

con lo que será $T \rightsquigarrow \gamma(n, e^\theta)$ y haciendo operaciones,

$$E[\log T] = \frac{\Gamma'(n)}{\Gamma(n)} - \theta.$$

(De hecho está calculada en PIE-ejercicio 1.9). Por tanto,

$$\frac{\Gamma'(n)}{\Gamma(n)} - \log T$$

será el ECUMV para θ.

Problema 3

Nos piden un test de la χ^2 de *homogeneidad de varias muestras* (CB-sección 12.3), en donde la hipótesis nula que se establece es que ambos grupos de edades pueden considerase homogéneos. Esta hipótesis nula se rechazará cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson. Si queremos resolverlo con R (EAR-sección 7.3), ejecutaríamos la siguiente secuencia de instrucciones. Con `(1)` incluimos los datos, que tienen que venir en forma de matriz. Recordemos que, por defecto, los incorpora por columnas. Las sentencias `(2)` y `(3)` son opcionales y sirven para poner nombre a las filas y a las columnas de la tabla. Con `(4)` comprobamos que hemos incorporado bien los datos a R. Ejecutando `(5)` es como le pedimos que haga el test χ^2.

```
> utero<-matrix(c(42,7,203,114),ncol=2)                              (1)
> colnames(utero)<-c("SÍ","NO")                                      (2)
> rownames(utero)<-c("menores 25","mayores 25")                      (3)
> utero                                                              (4)
           SÍ  NO
menores 25 42 203
mayores 25  7 114

> chisq.test(utero)                                                  (5)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  utero
X-squared = 8.0579, df = 1, p-value = 0.004531                       (6)
```

En `(6)` vemos el valor del estadístico de Pearson, $\lambda = 8'0579$ y el p-valor del test, $0'004531$, suficientemente pequeño como para concluir que puede rechazarse la hipótesis nula de homogeneidad de ambas poblaciones. Es decir, puede concluirse que sí parece existir una influencia en el cáncer por cuello de útero en las mujeres que has tenido el primer embarazo después de los 25 años de edad.

– –

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

PIE: **Principios de Inferencia Estadística**, 2012. R. Vélez y A. García Pérez. Editorial UNED, Colección Grado (código:6102310GR01A01).

CB: **Estadística Aplicada: Conceptos Básicos** (Segunda Edición), 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código:0184011EP01A02).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código:0137352PB01A01).

ADD: **Fórmulas y Tablas Estadísticas**, (1998). Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 41206AD01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

Curso 2014-2015

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2014-2015.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{2\theta}{3^{2\theta}}\, x^{2\theta-1} \qquad \text{si} \quad 0 < x < 3$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar el estimador de máxima verosimilitud para θ y analizar si es suficiente.

2 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para $1/\theta$.

3 puntos

c) Determinar un test uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. Aplicarlo al caso de que sea $n = 3$, $\alpha = 0'05$, $\theta_0 = 1/4$ y que se obtengan los valores $X_1 = 1$, $X_2 = 1'2$ y $X_3 = 2$.

3 puntos

Problema 2

Se clasificaron 218 tumbas de la Edad de Bronce en Ricas y Pobres según los objetos de ajuar encontrados en ellas para los 6 Grupos de Edad en los que se divide la población femenina de aquella época. Los resultados obtenidos fueron los siguientes:

Grupos de Edad	Ricas	Pobres
Infantil I	**5**	**24**
Infantil II	**8**	**20**
Juvenil	**12**	**25**
Adulta	**29**	**35**
Madura	**20**	**27**
Senil	**6**	**7**

¿Existen diferencias significativas entre los seis grupos de edad?

2 puntos

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Esto tiene especial sentido cuando hablamos de la resolución con el paquete R de un ejercicio, dado que en examen no se puede utilizar ordenador.

Problema 1

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{2\theta}{3^{2\theta}}\right)^n \prod_{i=1}^{n} x_i^{2\theta-1} \qquad \text{si} \quad x_1,...,x_n \in (0,3)$$

con lo cual será

$$\log f_\theta(x_1,...,x_n) = n\left[\log 2\theta - 2\theta \log 3\right] + (2\theta-1)\sum_{i=1}^{n} \log x_i$$

obteniéndose la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta} \log f_\theta(x_1,...,x_n) = n\,\frac{2}{2\theta} - 2n\log 3 + 2\sum_{i=1}^{n} \log x_i = 0$$

y, por tanto, el estimador de máxima verosimilitud para θ

$$T = \frac{1/2}{\log 3 - \frac{1}{n}\sum_{i=1}^{n} \log X_i}.$$

La función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1,...,x_n) = h(x_1,...,x_n) \cdot g_\theta(T(x_1,...,x_n))$$

con

$$h(x_1,...,x_n) = 1$$

y

$$g_\theta(T(x_1,...,x_n)) = \left(\frac{2\theta}{3^{2\theta}}\right)^n \exp\left\{(2\theta-1)n\left[\log 3 - \frac{1}{2T}\right]\right\}$$

es decir, mediante el producto de una función, h, que no depende del parámetro y una función, g, que depende del parámetro y de la muestra sólo a través del estadístico T. Por el teorema de factorización (*PIE*, pág. 167), T ser suficiente para la familia de densidades dada.

b) De entre las diferentes maneras de determinar el ECUMV de $1/\theta$, la más directa es la de utilizar el teorema de Lehmann-Scheffé (*PIE*, pág. 218) aunque para ello es necesario contar con un estimador suficiente y completo.

Como la distribución modelo es una familia exponencial (*PIE*, pág. 174) de la forma

$$f_\theta(x) = \frac{2\theta}{3^{2\theta}}\, e^{(2\theta-1)\log x}$$

será $S(X_1,...,X_n) = \sum_{i=1}^n \log X_i$ un estadístico minimal suficiente. Por tanto, (*PIE*, pág. 273), S ser además completo si la imagen de $q(\theta) = 2\theta - 1$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$, la imagen de la función q será $(-1, \infty)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo; basta ahora con encontrar un estimador insesgado para $1/\theta$ función de S y ése será el ECUMV para $1/\theta$. Si no lo encontramos fácilmente, se puede determinar un insesgado para $1/\theta$, digamos δ, y determinar $\delta_1 = E[\delta/S]$ que, por el teorema de Rao-Blackwell (*PIE*, pág. 211) es también insesgado para $1/\theta$ y además, por construcción, función de S.

En este punto es siempre muy útil determinar la distribución del estadístico S. Para ello, vamos a determinar en primer lugar la distribución de $Y = \log X$. Su función de densidad,

$$f_Y(y) = f_\theta(e^y)\, e^y = \frac{2\theta}{3^{2\theta}}\, e^{y(2\theta-1)}\, e^y = \frac{2\theta}{3^{2\theta}}\, e^{2\theta y} \qquad -\infty < y < \log 3$$

sugiere la transformación $-Y + \log 3$ para obtener una distribución gamma, ya que la variable $Z = -Y + \log 3$ tendrá por densidad

$$f(z) = f_Y(\log 3 - z) = 2\theta e^{-2\theta z} \qquad 0 < z < \infty$$

es decir, una $\gamma(1, 2\theta)$. Por tanto,

$$\sum_{i=1}^n Z_i = -\sum_{i=1}^n \log X_i + n\log 3 = -S(X_1,...,X_n) + n\log 3 \rightsquigarrow \gamma(n, 2\theta)$$

y, en consecuencia, $-E[S] + n\log 3 = n/(2\theta)$. Por tanto, el ECUMV para $1/\theta$ será

$$-\frac{2}{n}\sum_{i=1}^n \log X_i + 2\log 3.$$

c) Al ser f_θ una familia exponencial, tendrá razón de verosimilitud monótona en el estadístico S y, por tanto, por el teorema de Karlin-Rubin (*PIE*, pág. 349) existe un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual viene dado por

$$\phi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad S(x_1, ..., x_n) > c \\ \\ 0 & \text{si} \quad S(x_1, ..., x_n) \leq c \end{cases}$$

en donde c se determina por la condición del nivel,

$$E_{\theta_0}\left[\phi(X_1, ..., X_n)\right] = \alpha,$$

en nuestro caso, $\alpha = P_{\theta_0}\{S(X_1, ..., X_n) > c\}$. Es decir,

$$\alpha = P_{\theta_0}\{-S(X_1, ..., X_n) + n\log 3 < -c + n\log 3\} = P\{\chi^2_{2n} < -4\theta_0 c + 4\theta_0 n \log 3\}.$$

Por tanto, deberá ser $\chi^2_{2n;1-\alpha} = -4\theta_0 c + 4\theta_0 n \log 3$ y, en consecuencia,

$$c = n\log 3 - \frac{\chi^2_{2n;1-\alpha}}{4\theta_0}.$$

Con los datos del enunciado es $c = 1'66$ y $S = 0'875$. Por tanto, se acepta H_0.

Problema 2

Se trata de un *Contraste de homogeneidad de varias muestras* (CB-sección 12.3). La tabla de frecuencias observadas y esperadas (entre paréntesis) es

Grupos de Edad	Ricas	Pobres
Infantil I	5 (10′642)	24 (18′358)
Infantil II	8 (10′275)	20 (17′725)
Juvenil	12(13′578)	25 (23′422)
Adulta	29(23′486)	35 (40′514)
Madura	20(17′248)	27 (29′752)
Senil	6 (4′771)	7 (8′229)

que, como se ve, presenta una celdilla con frecuencia esperada menor que 5, pero dado que es por muy poco y los resultados que siguen son suficientemente claros, no es necesario agrupar las dos últimas filas.

El valor del estadístico de Pearson es igual a $\lambda = 9'05$ y el p-valor,

$$P\{\chi^2_5 > 9'05\}$$

aparece acotado entre $0'1$ y $0'3$ (bastante cercano a $0'1$), lo que conduce a aceptar la hipótesis nula de homogeneidad de los seis grupos de edad y concluir con que no existen diferencias significativas entre ellos.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2014-2015.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{4}{\theta}\, x^3\, e^{-x^4/\theta} \qquad \text{si} \quad x > 0$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar el estimador de máxima verosimilitud para θ y analizar si es suficiente.

2 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para θ. ¿Es este estimador eficiente para θ?

2 puntos

c) Determinar un intervalo de confianza para θ de colas iguales, de nivel de confianza $1 - \alpha$.

2 puntos

d) Determinar un test uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

2 puntos

Problema 2

Dos procesos industriales independientes tardan en la elaboración de un producto similar los siguientes tiempos

Proceso 1	**14′1**	**13′2**	**9′1**	**16′1**	**17′6**	**19′3**
Proceso 2	**13′7**	**12′5**	**13′4**	**26′3**	**18′4**	**15′4**

Suponiendo que los datos anteriores proceden de poblaciones normales independientes con la misma varianza, ¿concluiría con que existen diferencias significativas en los tiempos medios de ambos procesos de producción?

2 puntos

Problema 1

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n)=\prod_{i=1}^{n} f_\theta(x_i)=\left(\frac{4}{\theta}\right)^n \prod_{i=1}^{n} x_i^3 \ \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} x_i^4\right\} \qquad \text{si} \quad x_1,...,x_n>0$$

con lo cual será

$$\log f_\theta(x_1,...,x_n)=n\log 4-n\log\theta+\sum_{i=1}^{n}\log x_i^3-\frac{1}{\theta}\sum_{i=1}^{n} x_i^4$$

obteniéndose la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta}\log f_\theta(x_1,...,x_n)=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n} x_i^4=0$$

y, por tanto, el estimador de máxima verosimilitud para θ

$$T=\frac{1}{n}\sum_{i=1}^{n} X_i^4.$$

La función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1,...,x_n)=h(x_1,...,x_n)\cdot g_\theta(T(x_1,...,x_n))$$

con

$$h(x_1,...,x_n)=\prod_{i=1}^{n} x_i^3$$

y

$$g_\theta(T(x_1,...,x_n))=\left(\frac{4}{\theta}\right)^n \exp\left\{-\frac{1}{\theta}\,nT\right\}$$

es decir, mediante el producto de una función, h, que no depende del parámetro y una función, g, que depende del parámetro y de la muestra sólo a través del estadístico T. Por el teorema de factorización (*PIE*, pág. 167), T será suficiente para la familia de densidades dada.

b) De entre las diferentes maneras de determinar el ECUMV de θ, la más directa es la de utilizar el teorema de Lehmann-Scheffé (*PIE*, pág. 218) aunque para ello es necesario contar con un estimador suficiente y completo.

Como la distribución modelo es una familia exponencial (*PIE*, pág. 174) de la forma

$$f_\theta(x) = \frac{4}{\theta}\, x^3\, \exp\left\{-\frac{1}{\theta}\, x^4\right\}$$

será $S(X_1, ..., X_n) = \sum_{i=1}^n X_i^4$ un estadístico minimal suficiente. Por tanto, (*PIE*, pág. 273), S será además completo si la imagen de $q(\theta) = -1/\theta$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$, la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo; basta ahora con encontrar un estimador insesgado para θ función de S y ese será el ECUMV para θ. Si no lo encontramos fácilmente, se puede determinar un insesgado para θ, digamos δ, y determinar $\delta_1 = E[\delta/S]$ que, por el teorema de Rao-Blackwell (*PIE*, pág. 211) es también insesgado para θ y además, por construcción, función de S.

En este punto es siempre muy útil determinar la distribución del estadístico S. Para ello, vamos a determinar en primer lugar la distribución de $Y = X^4$. Su función de densidad es

$$f_Y(y) = f_\theta(y^{1/4})\, \frac{1}{4}\, y^{-3/4} = \frac{1}{\theta}\, e^{-y/\theta} \qquad\qquad y > 0$$

es decir, la de una $\gamma(1, 1/\theta)$. Por tanto, será

$$S = \sum_{i=1}^n X_i^4 \rightsquigarrow \gamma(n, 1/\theta)$$

y, en consecuencia, $E[S] = n\theta$. Por tanto, el ECUMV para θ será

$$\frac{S}{n} = \frac{1}{n} \sum_{i=1}^n X_i^4 = T.$$

Como es

$$\frac{\partial}{\partial\theta} \log f_\theta(x_1, ..., x_n) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i^4$$

la cantidad de información de Fisher de la muestra será (*PIE*, pág. 225)

$$\begin{aligned} I(\theta) &= E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(X_1, ..., X_n)\right)^2\right] = E_\theta\left[\left(-\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i^4\right)^2\right] \\ &= E\left[\frac{n^2}{\theta^2} + \frac{1}{\theta^4} (\sum_{i=1}^n X_i^4)^2 - \frac{2n}{\theta^3} \sum_{i=1}^n X_i^4\right] \\ &= \frac{n}{\theta^2}. \end{aligned}$$

Entonces, la acotación de Fréchet-Cramer-Rao para estimadores R centrados de θ es

$$V_\theta(R) \geq \frac{1}{I(\theta)} = \frac{\theta^2}{n}.$$

Como es

$$V(T) = \frac{\theta^2}{n}$$

T será eficiente para θ (*PIE*, pág. 228).

c) Para construir una cantidad pivotal (*PIE*, pág. 96) parece razonable utilizar el estadístico S. Como es $S \rightsquigarrow \gamma(n, 1/\theta)$, será $2S/\theta \rightsquigarrow \chi^2_{2n}$ y, por tanto, se pueden determinar dos valores χ^2 en las tablas, tales que sea

$$P\left\{\chi^2_{2n;1-\alpha/2} < \frac{2S}{\theta} < \chi^2_{2n;\alpha/2}\right\} = 1 - \alpha$$

con lo que, el intervalo de confianza pedido será

$$\left[\frac{2\sum_{i=1}^n x_i^4}{\chi^2_{2n;\alpha/2}}\ ,\ \frac{2\sum_{i=1}^n x_i^4}{\chi^2_{2n;1-\alpha/2}}\right].$$

d) Al ser f_θ una familia exponencial, tendrá razón de verosimilitud monótona en el estadístico S y, por tanto, por el teorema de Karlin-Rubin (*PIE*, pág. 349) existe un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual viene dado por

$$\phi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad S(x_1, ..., x_n) > c \\ \\ 0 & \text{si} \quad S(x_1, ..., x_n) \leq c \end{cases}$$

en donde c se determina por la condición del nivel,

$$E_{\theta_0}\left[\phi(X_1, ..., X_n)\right] = \alpha,$$

en nuestro caso, $\alpha = P_{\theta_0}\{S(X_1, ..., X_n) > c\}$. Es decir,

$$\alpha = P_{\theta_0}\left\{\sum_{i=1}^n X_i^4 > c\right\} = P\left\{\chi^2_{2n} > \frac{2c}{\theta_0}\right\}.$$

Por tanto, deberá ser $\chi^2_{2n;\alpha} = 2c/\theta_0$ y, en consecuencia,

$$c = \frac{\theta_0\ \chi^2_{2n;\alpha}}{2}.$$

Problema 2

Nos solicitan un test sobre las medias de dos poblaciones normales independientes (EBR-sección 7.6, CB-sección 7.6, ID-sección 5.5, PIE-sección 8.4) con varianzas desconocidas pero supuestamente iguales, $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$, aceptándose H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}$$

Como es

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = 0'6657$$

a partir de la Tabla 5 de la t de Student, vemos que

$$0'2 < P\{t_{10} > 0'6657\} < 0'3$$

con lo que el p-valor, $2 \cdot P\{t_{10} > 0'6657\}$, estará entre $0'4$ y $0'6$, en todo caso suficientemente grande como para aceptar la hipótesis nula de no existencia de diferencias significativas entre los tiempos medios de fabricación por ambos procesos.

Para resolver este ejercicio con R se ejecutarían las siguientes sentencias (EAR-sección 4.2),

```
> x1<-c(14.1,13.2,9.1,16.1,17.6,19.3)
> x2<-c(13.7,12.5,13.4,26.3,18.4,15.4)
> t.test(x1,x2,alternative="two.sided",var.equal=T)

        Two Sample t-test

data:  x1 and x2
t = -0.6657, df = 10, p-value = 0.5207
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.462333  4.028999
sample estimates:
mean of x mean of y
 14.90000  16.61667
```

Aquí se ve que, en concreto, el p-valor es igual a $0'5207$.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2014-2015.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_{(\theta_1,\theta_2)}(x) = \theta_2\,\theta_1^{\theta_2}\,x^{-(\theta_2+1)} \qquad \text{si} \quad x > \theta_1$$

siendo $\theta_1, \theta_2 > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar el estimador de máxima verosimilitud $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ del parámetro $\theta = (\theta_1, \theta_2)$.

1 punto

b) Supuesto θ_2 conocido, determinar el estimador centrado uniformemente de mínima varianza para θ_1.

3 puntos

c) Supuesto θ_1 conocido, determinar el estimador centrado uniformemente de mínima varianza para $1/\theta_2$.

2 puntos

d) En este supuesto de ser θ_1 conocido, determinar un intervalo de confianza para θ_2 de colas iguales, de coeficiente de confianza $1-\alpha$.

2 puntos

Problema 2

En un artículo de Fombonne (1989) se investiga la posibilidad de que el mes de nacimiento esté relacionado con la psicosis infantil. Para ello, el autor del trabajo realizó un estudio en París con 1248 niños y los clasificó por mes de nacimiento y según si eran enfermos (Casos) o no (Controles) de esta enfermedad psiquiátrica. Los resultados obtenidos fueron los siguientes:

	Casos	Controles
Enero	**13**	**83**
Febrero	**12**	**71**
Marzo	**16**	**88**
Abril	**18**	**114**
Mayo	**21**	**86**
Junio	**18**	**93**
Julio	**15**	**87**
Agosto	**14**	**70**
Septiembre	**13**	**83**
Octubre	**19**	**80**
Noviembre	**21**	**97**
Diciembre	**28**	**88**

A la vista de estos datos, ¿cree que están relacionados el mes de nacimiento y esta enfermedad?

2 puntos

Problema 1

a) Llamando θ al parámetro bidimensional (θ_1, θ_2), la función de verosimilitud de la muestra será

$$f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \theta_2^n \, \theta_1^{n\theta_2} \left(\prod_{i=1}^{n} x_i \right)^{-(\theta_2+1)} \qquad \text{si} \quad x_1, ..., x_n \geq \theta_1.$$

Como es $\theta_2 > 0$, la función de verosimilitud es tanto mayor cuanto mayor sea θ_1; no obstante, θ_1 no puede ser mayor que el mínimo de las observaciones puesto que, en ese caso, f_θ sería cero. En consecuencia, será $\hat{\theta}_1 = X_{(1)}$ y además, de la forma habitual, obtenemos que es

$$\hat{\theta}_2 = \frac{-n}{n \log X_{(1)} - \sum_{i=1}^{n} \log X_i}.$$

b) La forma más razonable de determinar el ECUMV de θ_1 es la de determinar un estadístico suficiente y completo S para la familia de distribuciones $f_{\theta_1}(x)$ y luego, utilizando el teorema de Lehmann-Scheffé (*PIE*, pág. 218), determinar un estadístico función de S que sea insesgado para θ_1.

Como $X_{(1)}$ es el estimador de máxima verosimilitud y, además, factorizando la función de densidad de la muestra de la forma

$$f_{\theta_1}(x_1, ..., x_n) = \theta_2^n \left(\prod_{i=1}^{n} x_i \right)^{-(\theta_2+1)} \theta_1^{n\theta_2} \, I_{x_{(1)} > \theta_1}$$

$X_{(1)}$ será también suficiente por el teorema de factorización (*PIE*, pág. 167). Parece razonable, por tanto, analizar su completitud.

La función de densidad del mínimo es (*PIE*, pág. 28)

$$\begin{aligned} f_{\theta_1}(y) &= n \left[1 - F_{\theta_1}(y)\right]^{n-1} f_{\theta_1}(y) \\ &= n \left[\frac{\theta_1^{\theta_2}}{y^{\theta_2}} \right]^{n-1} \theta_2 \, \theta_1^{\theta_2} \, y^{-\theta_2-1} \\ &= n \, \theta_2 \, \theta_1^{n\theta_2} \, y^{-(n\theta_2+1)} \end{aligned}$$

si $y > \theta_1$.

Por tanto, será

$$E_{\theta_1}\left[g(X_{(1)})\right] = \int_{\theta_1}^{\infty} g(y) \, \frac{n\theta_2 \theta_1^{n\theta_2}}{y^{n\theta_2+1}} \, dy$$

la cual es idénticamente nula si $\int_{\theta_1}^{\infty} g(y)\, y^{-n\theta_2-1}\, dy = 0 \quad \forall \theta_1 > 0$. Derivando respecto a θ_1, resulta $g(\theta_1)\theta_1^{-n\theta_2-1} = 0$; es decir, $g(\theta_1) = 0 \ \ \forall \theta_1 > 0$. El mínimo es, por tanto, completo, además de suficiente.

Basta determinar una función de $X_{(1)}$ que sea insesgada para θ_1. Al ser

$$E[X_{(1)}] = \frac{n\theta_2}{n\theta_2 - 1}\, \theta_1$$

el ECUMV para θ_1 será

$$\frac{n\theta_2 - 1}{n\theta_2}\, X_{(1)}.$$

(Una situación muy similar puede seguirse en el Ejemplo 6.5, *PIE*, pág. 217, y en el Ejercicio 6.4, *PIE*, pág. 265.)

c) Si θ_1 es conocido, la distribución modelo es una familia exponencial (*PIE*, pág. 174) de la forma

$$f_{\theta_2}(x) = \theta_1^{\theta_2}\, \theta_2\, e^{-(\theta_2+1)\log x}$$

con lo que $S(X_1, ..., X_n) = \sum_{i=1}^{n} \log X_i$ será un estadístico minimal suficiente y, además, (*PIE*, p g. 273), completo si la imagen de $q(\theta_2) = -(\theta_2+1)$ contiene un abierto de $\mathbb{R}$; como es $\theta_2 > 0$, la imagen de la función q será $(-\infty, -1)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo; basta ahora con encontrar un estimador insesgado para $1/\theta_2$ función de S y ese será el ECUMV para $1/\theta_2$.

En este punto es muy útil determinar la distribución del estadístico S. Para ello, vamos a determinar en primer lugar la distribución de $Y = \log X$. Su función de densidad,

$$f_Y(y) = f_{\theta_2}(e^y)\, e^y = \theta_2\, \theta_1^{\theta_2}\, e^{-\theta_2 y} \qquad\qquad y > \log \theta_1$$

sugiere la transformación $Y - \log \theta_1$ para obtener una distribución gamma, ya que la variable $Z = Y - \log \theta_1$ tendrá por densidad

$$f(z) = f_Y(z + \log \theta_1) = \theta_2 e^{-\theta_2 z} \qquad\qquad z > 0$$

es decir, una $\gamma(1, \theta_2)$. Por tanto,

$$\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} \log X_i - n \log \theta_1 = S(X_1, ..., X_n) - n \log \theta_1 \rightsquigarrow \gamma(n, \theta_2)$$

y, en consecuencia, $E[S] - n \log \theta_1 = n/\theta_2$. Por tanto, el ECUMV para $1/\theta_2$ será

$$\frac{1}{n}\sum_{i=1}^{n}\log X_i - \log\theta_1.$$

d) Para construir una cantidad pivotal (*PIE*, pág. 96) parece razonable utilizar el estadístico S. Como es $S - n\log\theta_1 \rightsquigarrow \gamma(n,\theta_2)$, será $2\theta_2(S - n\log\theta_1) \rightsquigarrow \chi^2_{2n}$ y, por tanto, se pueden determinar dos valores χ^2 en las tablas, tales que sea

$$P\left\{\chi^2_{2n;1-\alpha/2} < 2\theta_2(S - n\log\theta_1) < \chi^2_{2n;\alpha/2}\right\} = 1-\alpha$$

con lo que, el intervalo de confianza pedido ser

$$\left[\frac{\chi^2_{2n;1-\alpha/2}}{2(\sum_{i=1}^n \log x_i - n\log\theta_1)}\ ,\ \frac{\chi^2_{2n;\alpha/2}}{2(\sum_{i=1}^n \log x_i - n\log\theta_1)}\right] = \left[\frac{\chi^2_{2n;1-\alpha/2}}{2\log\prod_{i=1}^n (x_i/\theta_1)}\ ,\ \frac{\chi^2_{2n;\alpha/2}}{2\log\prod_{i=1}^n (x_i/\theta_1)}\right].$$

Problema 2

Nos piden un test de la χ^2 de *independencia de caracteres* (CB-sección 12.4). El valor del estadístico de Pearson es igual a $\lambda = 8'8358$ y el p-valor,

$$P\{\chi^2_{11} > 8'8358\}$$

aparece acotado entre $0'3$ y $0'7$ (bastante cercano a $0'7$), lo que conduce a aceptar la hipótesis nula de independencia entre el mes de nacimiento y la enfermedad y poder concluir que el mes no tiene influencia en esta enfermedad.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

MR: **Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo**, 2005. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0186080EP03A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

Curso 2015-2016

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2015-2016.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\log 3}{3^\theta - 1}\, 3^x \qquad \text{si} \quad 0 < x < \theta.$$

Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a)Determinar el estimador de máxima verosimilitud para θ.

1 punto

b)Determinar el estimador centrado uniformemente de mínima varianza para 3^θ.

4 puntos

c)Determinar, por el método de Neyman, un intervalo de confianza para θ, de nivel de confianza $1-\alpha$. De entre ellos, ¿cuál es el de longitud mínima?

3 puntos

Problema 2

Se piensa que el nivel medio de tromboglobulina que se elimina en la orina de los enfermos con diabetes es significativamente mayor que el nivel medio de los no diabéticos, establecido en 13′5.

Para confirmar esta idea se eligieron al azar 36 pacientes diabéticos que proporcionaron un nivel medio de tromboglobulina de 22 y una cuasidesviación típica muestral de 20.

¿Puede admitirse que efectivamente los pacientes diabéticos presentan un nivel medio de tromboglobulina mayor de 13′5?

2 puntos

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Esto tiene especial sentido cuando hablamos de la resolución con el paquete R de un ejercicio, dado que en examen no se puede utilizar ordenador.

Problema 1

a) La función de verosimilitud de la muestra será

$$f_\theta(x_1,...,x_n)=\prod_{i=1}^{n} f_\theta(x_i)=\left(\frac{\log 3}{3^\theta-1}\right)^n\ 3^{x_1+...+x_n} \qquad \text{si} \quad 0<x_1,...,x_n<\theta.$$

La función de verosimilitud es tanto mayor cuanto menor sea θ; no obstante, θ no puede ser menor que el máximo de las observaciones puesto que, en ese caso, f_θ sería cero. En consecuencia, será $\hat{\theta}=X_{(n)}$.

b) La forma más razonable de determinar el ECUMV para 3^θ es determinar un estadístico suficiente y completo para la familia de distribuciones $f_\theta(x)$ y luego, utilizando el teorema de Lehmann-Scheffé (*PIE*, pp. 218), determinar un estadístico función suya que sea insesgado para 3^θ.

Como $X_{(n)}$ es el estimador de máxima verosimilitud y, además, factorizando la función de densidad de la muestra de la forma

$$f_\theta(x_1,...,x_n)=\left(\frac{\log 3}{3^\theta-1}\right)^n\ 3^{x_1+...+x_n}\ I_{x_{(n)}<\theta}$$

$X_{(n)}$ será también suficiente por el teorema de factorización (*PIE*, pp. 167), parece razonable analizar su completitud para ver si es el estimador buscado.

La función de densidad del máximo es (*PIE*, pp. 28)

$$\begin{aligned} g_\theta(y) &= n\,[F_\theta(y)]^{n-1}\ f_\theta(y) \\ &= n\left[\frac{3^y-1}{3^\theta-1}\right]^{n-1}\ \frac{\log 3}{3^\theta-1}\,3^y \\ &= \frac{n\,\log 3}{(3^\theta-1)^n}\,(3^y-1)^{n-1}\,3^y \end{aligned}$$

si $0<y<\theta$.

Por tanto, será

$$E_\theta\left[g(X_{(n)})\right] = \int_0^\theta g(y)\,\frac{n\,\log 3}{(3^\theta-1)^n}\,(3^y-1)^{n-1}\,3^y\,dy.$$

Si esta esperanza es cero $\forall\ \theta > 0$, entonces será $\int_0^\theta g(y)\,(3^y-1)^{n-1}\,3^y\,dy = 0 \quad \forall\theta > 0$. Derivando respecto a θ, resulta $g(\theta)(3^\theta-1)^{n-1}3^\theta = 0$; es decir, $g(\theta) = 0 \ \ \forall\theta > 0$. El máximo es, por tanto, completo, además de suficiente.

Basta ahora determinar una función de $X_{(n)}$ que sea insesgada para 3^θ. Al ser $X_{(n)}$ un "buen estimador" de θ, parece razonable probar con $3^{X_{(n)}}$. Como es, integrando por partes,

$$E[3^{X_{(n)}}] = \frac{3^\theta\,n+1}{n+1}$$

el ECUMV para 3^θ será pues

$$\frac{(n+1)\,3^{X_{(n)}}-1}{n}.$$

c) La función de distribución del máximo es

$$G_\theta(y) = \begin{cases} 0 & \text{si} \quad y < 0 \\ \left(\dfrac{3^y-1}{3^\theta-1}\right)^n & \text{si} \quad 0 \le y < \theta \\ 1 & \text{si} \quad y \ge \theta. \end{cases}$$

De las ecuaciones

$$G_\theta\left(c_1(\theta)\right) = \alpha_1 \qquad\qquad G_\theta\left(c_2(\theta)\right) = 1-\alpha_2$$

se obtienen las funciones

$$c_1(\theta) = \frac{\log\left[(3^\theta-1)\alpha_1^{1/n}+1\right]}{\log 3}$$

y

$$c_2(\theta) = \frac{\log\left[(3^\theta-1)(1-\alpha_2)^{1/n}+1\right]}{\log 3}$$

con lo que será (*PIE*, pp. 102)

$$\mathcal{P}_\theta\left\{\frac{\log\left[(3^\theta-1)\alpha_1^{1/n}+1\right]}{\log 3}<X_{(n)}<\frac{\log\left[(3^\theta-1)(1-\alpha_2)^{1/n}+1\right]}{\log 3}\right\}=1-\alpha$$

es decir,

$$\mathcal{P}_\theta\left\{\frac{\log\frac{3^{X_{(n)}}-1+(1-\alpha_2)^{1/n}}{(1-\alpha_2)^{1/n}}}{\log 3}<\theta<\frac{\log\frac{3^{X_{(n)}}-1+\alpha_1^{1/n}}{\alpha_1^{1/n}}}{\log 3}\right\}=1-\alpha$$

con lo que, el intervalo de confianza pedido será

$$\left[\frac{\log\frac{3^{X_{(n)}}-1+(1-\alpha_2)^{1/n}}{(1-\alpha_2)^{1/n}}}{\log 3}\ ,\ \frac{\log\frac{3^{X_{(n)}}-1+\alpha_1^{1/n}}{\alpha_1^{1/n}}}{\log 3}\right].$$

Como la densidad del máximo es creciente, el intervalo de longitud mínima se obtendrá tomando $\alpha_1=\alpha$ y $\alpha_2=0$; es decir,

$$\left[X_{(n)}\ ,\ \frac{\log\frac{3^{X_{(n)}}-1+\alpha^{1/n}}{\alpha^{1/n}}}{\log 3}\right].$$

Problema 2

Se trata de un test en donde la hipótesis de interés se establece como hipótesis alternativa $H_1:\mu>13'5$ siendo μ la media de la variable de interés: *nivel de tromboglobulina eliminado en la orina de enfermos de diabetes.*

La hipótesis alternativa será por tanto la complementaria $H_0:\mu\leq 13'5$. A partir del enunciado se establecen las suposiciones del problema de ser un caso de un test de hipótesis para la media de una población no necesariamente normal y con tamaños muestrales grandes (CB-sección 7.3), que indica rechzar la hipótesis nula cuando y sólo cuando sea

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}>z_\alpha$$

siendo, como siempre, z_α el valor de la abscisa de una normal estándar que deja a su derecha un área de probabilidad α.

El valor del estadístico de contraste es

$$T_n=\frac{\overline{x}-\mu_0}{S/\sqrt{n}}=\frac{22-13'5}{20/\sqrt{36}}=2'55$$

Como no nos fijan un nivel de significación concreto, podemos elegir nosotros uno, por ejemplo $\alpha = 0'05$, con lo que sería $z_\alpha = 1'645 < T_n$ concluyendo con el rechazo de la hipótesis nula, o mejor, calculamos el p-valor (lo que siempre debemos hacer en un test de hipótesis), que en este caso es igual a

$$P\{Z > 2'55\} = 0'005386$$

siendo Z una variable normal $N(0,1)$ ya que con R obtenemos que

```
> 1-pnorm(2.55)
[1] 0.005386146
```

Con unas tablas de la normal estándar obtendríamos el valor,

$$P\{Z > 2'55\} = 0'0054.$$

Lo que queda claro en todo caso es que debemos rechazar la hipótesis nula y concluir que los pacientes en estudio presentan una nivel medio mayor de $13'5$.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2015-2016.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria discreta con valores en los enteros no negativos, cuya función de probabilidad es

$$f_\theta(x) = \binom{r+x-1}{x} \theta^r (1-\theta)^x \qquad\qquad x = 0, 1, 2, ...$$

siendo r un número natural conocido y $\theta \in (0,1)$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Estimar θ por el método de la máxima verosimilitud y por el método de los momentos.

2 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para $g_1(\theta) = (1-\theta)/\theta$.

2 puntos

c) Determinar el estimador centrado uniformemente de mínima varianza para $g_2(\theta) = \theta^r$.

2 puntos

d) Determinar un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

2 puntos

Problema 2

Se quiere determinar un intervalo de confianza para la longitud media de los antebrazos de los hombres adultos, admitiendo una distribución normal para dicha variable en estudio.

Para ello se obtuvo una muestra de 15 individuos en los que se observó una media de 50′1 cm. y una cuasivarianza muestral igual a 4.

Determinar el mencionado intervalo de confianza para un coeficiente de confianza de 0′95.

2 puntos

a) La función de verosimilitud de la muestra (véase *PIE*, pp. 284) será

$$f_\theta(x_1,...,x_n)=\prod_{i=1}^{n} f_\theta(x_i)=\theta^{nr}\left[\prod_{i=1}^{n}\binom{r+x_i-1}{x_i}\right](1-\theta)^{\sum_{i=1}^{n}x_i}.$$

Tomando logaritmos y derivando respecto al parámetro obtenemos la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta}\log f_\theta(x_1,...,x_n)=\frac{nr}{\theta}-\frac{\sum_{i=1}^{n}x_i}{1-\theta}=0$$

de donde obtenemos el estimador de máxima verosimilitud

$$\hat{\theta}=\frac{nr}{nr+\sum_{i=1}^{n}x_i}=\frac{1}{1+\frac{\sum_{i=1}^{n}x_i}{nr}}.$$

Para utilizar el método de los momentos (*PIE*, pp. 279) formamos la ecuación $E[X]=\overline{x}$ y despejamos el parámetro. Dicha ecuación queda

$$\frac{r(1-\theta)}{\theta}=\overline{x}$$

de donde se obtiene como estimador de θ

$$\hat{\theta}=\frac{r}{r+\overline{x}}$$

es decir, el mismo que el obtenido por el método de la máxima verosimilitud.

b) Para determinar el ECUMV de cualquier función del parámetro, lo primero que necesitamos será determinar un estadístico suficiente y completo S para la familia de distribuciones $f_\theta(x)$.

Como la distribución modelo es una familia exponencial (*PIE*, pp. 174) de la forma

$$f_\theta(x)=\theta^r\binom{r+x-1}{x}e^{x\log(1-\theta)}$$

será $S(X_1,...,X_n)=\sum_{i=1}^{n}X_i$ un estadístico minimal suficiente. Por tanto, (*PIE*, pp. 273), S será además completo si la imagen de $q(\theta)=\log(1-\theta)$ contiene un abierto de $\mathbb{R}$; como es $\theta\in(0,1)$ la imagen de la función q será $(-\infty,0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo. Si ahora somos capaces de intuir un estadístico función de S que sea insesgado para $g_1(\theta)$, utilizando el teorema de Lehmann-Scheffé (*PIE*, pp. 218), ése será el ECUMV de $g_1(\theta)$. Para ello es

razonable determinar la esperanza de S, bien directamente (n veces la media de X) o, en este caso, al ser la distribución modelo una binomial negativa $BN(r,\theta)$, será $S \rightsquigarrow BN(nr,\theta)$ de media conocida. En todo caso,

$$E[S] = n\,\frac{r(1-\theta)}{\theta}$$

con lo que el ECUMV para $g_1(\theta) = (1-\theta)/\theta$ será $\sum_{i=1}^{n} X_i/(nr)$.

c) Para determinar el ECUMV de $g_2(\theta) = \theta^r$ utilizaremos otro método, el cual da buenos resultados en el caso de que el parámetro a estimar sea la probabilidad con la que la variable en estudio toma algún valor concreto y, en este caso es así ya que $P_\theta\{X=0\} = f_\theta(0) = \theta^r$.

El método consiste (véase por ejemplo *PIE*, pp. 213) primero en elegir un estimador trivial e insesgado para el parámetro, por ejemplo

$$T_1 = \begin{cases} 1 & \text{si} \quad X_1 = 0 \\ 0 & \text{si} \quad X_1 \neq 0 \end{cases}$$

(el cual claramente es insesgado, y demuestra por qué es útil este método para estimar probabilidades de la variable en estudio) y luego calcular $T_2 = E[T_1/S] = \mathcal{P}_\theta\{X_1 = 0/S\}$ siendo S el suficiente y completo, ya que por construcción T_2 será insesgado y función del suficiente y completo.

El ECUMV para $g_2(\theta) = \theta^r$ será por tanto,

$$T_2 = \mathcal{P}_\theta\left\{X_1 = 0 \,\middle|\, \sum_{i=1}^{n} X_i = s\right\} = \frac{\mathcal{P}_\theta\{X_1 = 0, \sum_{i=2}^{n} X_i = s\}}{\mathcal{P}_\theta\{\sum_{i=1}^{n} X_i = s\}}$$

$$= \frac{\theta^r \binom{(n-1)r+s-1}{s} \theta^{(n-1)r}(1-\theta)^s}{\binom{nr+s-1}{s}\theta^{nr}(1-\theta)^s} = \frac{\Gamma(nr+s-r)\,\Gamma(nr)}{\Gamma(nr+s)\,\Gamma(nr-r)}.$$

d) Al ser f_θ una familia exponencial, tendrá razón de verosimilitud monótona en el estadístico S y, por tanto, por el teorema de Karlin-Rubin (*PIE*, pp. 349) existe un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual viene dado por (al ser $q(\theta) = \log(1-\theta)$ monótona decreciente en θ)

$$\phi(x_1,...,x_n) = \begin{cases} 1 & \text{si} \quad -S(x_1,...,x_n) > c \\ \gamma & \text{si} \quad -S(x_1,...,x_n) = c \\ 0 & \text{si} \quad -S(x_1,...,x_n) < c \end{cases}$$

o bien,

$$\phi(x_1,...,x_n)=\begin{cases}1 & \text{si} \quad \sum_{i=1}^n X_i < k \\ \gamma & \text{si} \quad \sum_{i=1}^n X_i = k \\ 0 & \text{si} \quad \sum_{i=1}^n X_i > k\end{cases}$$

en donde γ y k se determinan por la condición del nivel,

$$E_{\theta_0}\left[\phi(X_1,...,X_n)\right]=\alpha,$$

es decir,

$$\mathcal{P}_{\theta_0}\left\{\sum_{i=1}^n X_i < k\right\}+\gamma\,\mathcal{P}_{\theta_0}\left\{\sum_{i=1}^n X_i = k\right\}=\alpha.$$

Problema 2

Se trata de un caso de Intervalo de Confianza para la media de una población normal con varianza desconocida y muestras pequeñas (CB-sección 6.2)

$$\left[\overline{x}-t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\;,\;\overline{x}+t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\right].$$

De los datos observados obtenemos que es $\overline{x}=50'1$ y $S=2$; además, al ser el coeficiente de confianza $0'95$, a partir de la tabla 5 de la distribución t de Student obtenemos que es $t_{n-1;\alpha/2}=t_{14;0'025}=2'145$. Por tanto, el intervalo de confianza para la media, de coeficiente de confianza $0'95$, es

$$\begin{aligned}&\left[\overline{x}-t_{n-1;\alpha/2}\,\tfrac{S}{\sqrt{n}}\;,\;\overline{x}+t_{n-1;\alpha/2}\,\tfrac{S}{\sqrt{n}}\right]\\ =\;&\left[50'1-2'145\,\tfrac{2}{\sqrt{15}}\;,\;50'1+2'145\,\tfrac{2}{\sqrt{15}}\right]\\ =\;&[48'992\,,\,51'2076].\end{aligned}$$

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2015-2016.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \theta\, e^{2-x}\, \left(1 - e^{2-x}\right)^{\theta-1} \qquad\qquad x > 2$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a)Determinar el estimador de máxima verosimilitud para θ.

1.5 puntos

b)Determinar el estimador centrado uniformemente de mínima varianza para $1/\theta$ y el estimador centrado uniformemente de mínima varianza para θ .

3.5 puntos

c) Determinar un intervalo de confianza para $1/\theta$, de nivel de confianza $1 - \alpha$.

3 puntos

Problema 2

En un estudio sobre caries dental en niños, se tomaron muestras en cuatro zonas geográficas con distintos niveles de flúor en el agua con objeto de analizar si existían diferencias significativas entre las cuatro zonas.

Para ello se tomó una muestra de 120 niños en cada zona, obteniéndose los siguientes datos:

Zona	Niños sin caries
A	48
B	18
C	40
D	54

¿Se pueden aceptar como equivalentes las cuatro zonas geográficas respeto a la presencia de caries?

2 puntos

Problema 1
a) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n)=\prod_{i=1}^{n} f_\theta(x_i)=\theta^n \exp\left\{\sum_{i=1}^{n}(2-x_i)\right\}\left[\prod_{i=1}^{n}\left(1-e^{2-x_i}\right)\right]^{\theta-1} \quad \text{si} \quad x_1,...,x_n>2$$

con lo cual será

$$\log f_\theta(x_1,...,x_n)=n\log\theta+2\,n-\sum_{i=1}^{n}x_i+(\theta-1)\sum_{i=1}^{n}\log\left(1-e^{2-x_i}\right)$$

obteniéndose de la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta}\log f_\theta(x_1,...,x_n)=\frac{n}{\theta}+\sum_{i=1}^{n}\log(1-e^{2-x_i})=0$$

el estimador de máxima verosimilitud para θ

$$\hat{\theta}=\frac{n}{-\sum_{i=1}^{n}\log\left(1-e^{2-X_i}\right)}.$$

b) Para determinar el ECUMV de cualquier función del parámetro, en primer lugar calcularemos un estadístico suficiente y completo S para la familia de distribuciones $f_\theta(x)$.

Como la distribución modelo es una familia exponencial (PIE, pp. 174) de la forma

$$f_\theta(x)=\theta\,e^{2-x}\,\exp\left\{(\theta-1)\log(1-e^{2-x})\right\}$$

será

$$S(X_1,...,X_n)=\sum_{i=1}^{n}\log(1-e^{2-X_i})$$

un estadístico minimal suficiente. Por tanto, (PIE, pp. 273), S será además completo si la imagen de $q(\theta)=\theta-1$ contiene un abierto de $\mathbb{R}$; como es $\theta>0$ la imagen de la función q será $(-1,\infty)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo. Si ahora somos capaces de intuir un estadístico función de S que sea insesgado para $1/\theta$, utilizando el teorema de Lehmann-Scheffé (PIE, pp. 218), ese estadístico será el ECUMV de $1/\theta$.

Para ello es razonable determinar la esperanza de S, para lo que calcularemos primero la función de densidad de $Y = -\log(1 - e^{2-X})$. Es

$$f_Y(y) = f_X\left(2 - \log(1 - e^{-y})\right) \frac{e^{-y}}{1 - e^{-y}} = \theta\, e^{-\theta y} \qquad y > 0$$

es decir, $Y \rightsquigarrow \gamma(1, \theta)$ y como esta distribución es reproductiva respecto al primer parámetro, será

$$-\sum_{i=1}^{n} \log\left(1 - e^{2-X_i}\right) \rightsquigarrow \gamma\left(n, \theta\right).$$

En consecuencia,

$$E[S] = E\left[\sum_{i=1}^{n} \log(1 - X_i)\right] = \frac{-n}{\theta}$$

y, por tanto, el ECUMV para $1/\theta$ será $-S/n$.

Con objeto de determinar el ECUMV para θ y, en vista de lo anterior, parece razonable empezar calculando $E[-1/S]$. Será

$$E\left[\frac{1}{-S}\right] = \int_0^\infty \frac{1}{x} \frac{\theta^n\, e^{-\theta x}\, x^{n-1}}{\Gamma(n)}\, dx = \frac{\theta}{n-1} \int_0^\infty \frac{\theta^{n-1}\, e^{-\theta x}\, x^{n-2}}{\Gamma(n-1)}\, dx = \frac{\theta}{n-1}.$$

El ECUMV para θ será, por tanto, $-(n-1)/S$.

c) Como es

$$-\sum_{i=1}^{n} \log\left(1 - e^{2-X_i}\right) = -S(X_1, ..., X_n) \rightsquigarrow \gamma\left(n, \theta\right)$$

será

$$-2\,\theta\, S \rightsquigarrow \gamma\left(\frac{2n}{2}, \frac{1}{2}\right) = \chi^2_{2n}$$

estadístico que actúa de cantidad pivotal (PIE, pp. 96), mediante el cual se pueden determinar dos valores $\chi^2_{2n;1-\alpha/2}$ y $\chi^2_{2n;\alpha/2}$ tales que

$$P\left\{\chi^2_{2n;1-\alpha/2} < -2\,\theta\, S < \chi^2_{2n;\alpha/2}\right\} = 1 - \alpha$$

con lo que, despejando $1/\theta$, será

$$P\left\{\frac{-2S}{\chi^2_{2n;\alpha/2}} < \frac{1}{\theta} < \frac{-2S}{\chi^2_{2n;1-\alpha/2}}\right\} = 1-\alpha$$

y, por tanto,

$$\left[\frac{-2\sum_{i=1}^n \log\left(1-e^{2-X_i}\right)}{\chi^2_{2n;\alpha/2}}\ ,\ \frac{-2\sum_{i=1}^n \log\left(1-e^{2-X_i}\right)}{\chi^2_{2n;1-\alpha/2}}\right]$$

será un intervalo de confianza para $1/\theta$, de nivel de confianza $1-\alpha$.

Problema 2

Primero de todo decir que, en realidad tenemos una tabla con dos columnas ya que sabemos que se muestrearon 120 niños en cada zona. Por tanto, la tabla a estudiar es, en realidad,

Zona	Niños sin caries	Niños con caries	Totales
A	48	72	120
B	18	102	120
C	40	80	120
D	54	66	120

Como los datos aportados son recuentos de observaciones clasificados por clases, comparar las cuatro zonas debe hacerse mediante un test de la χ^2 de *homogeneidad de varias muestras* (CB-sección 12.3), en donde la hipótesis nula que se establece es que las cuatro zonas pueden considerase homogéneas. Esta hipótesis nula se rechazará cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson. En nuestro caso, toma el valor $\lambda = 27'9$.

De la Tabla 4 de la χ^2 de Pearson vemos que el p-valor es

$$P\{\chi^2_{(r-1)(s-1)} > 27'9\} = P\{\chi^2_3 > 27'9\} < 0'005$$

suficientemente pequeño como para rechazar la hipótesis nula de homogeneidad con bastante seguridad.

Si queremos resolverlo con R (EAR-sección 7.3), ejecutaríamos la siguiente secuencia de instrucciones. Con `(1)` incluimos los datos, que tienen que venir

en forma de matriz. Recordemos que, por defecto, los incorpora por columnas. Las sentencias (2) y (3) son opcionales y sirven para poner nombre a las filas y a las columnas de la tabla. Con (4) comprobamos que hemos incorporado bien los datos a R. Ejecutando (5) es como le pedimos que haga el test χ^2.

```
> caries<-matrix(c(48,18,40,54,72,102,80,66),ncol=2)                    (1)
> colnames(caries)<-c("Sin caries","Con caries")                        (2)
> rownames(caries)<-c("A","B","C","D")                                  (3)

> caries                                                                (4)
  Sin caries Con caries
A         48         72
B         18        102
C         40         80
D         54         66

> chisq.test(caries)                                                    (5)

        Pearson's Chi-squared test

data:  caries
X-squared = 27.9, df = 3, p-value = 3.812e-06                           (6)
```

En (6) vemos el valor del estadístico de Pearson, $\lambda = 27'9$ y el p-valor del test, $3'812e$-06, suficientemente pequeño como para concluir que puede rechazarse la hipótesis nula de homogeneidad de las cuatro zonas geográficas. Es decir, puede concluirse que existen diferencias significativas entre las zonas estudiadas.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

Curso 2016-2017

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2016-2017.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{2}{\theta^2\,\pi}\,\sqrt{\theta^2 - x^2} \qquad\qquad -\theta < x < \theta$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Estimar el parámetro θ por el método de los momentos.

2 puntos

b) Determinar un estimador insesgado para θ^2.

3 puntos

c) Supuesto que el tamaño muestral n es suficientemente grande, determinar un intervalo de confianza aproximado (es decir, basado en una distribución asintótica) de nivel de confianza $1-\alpha$, para θ^2. ¿Cuál sería el de θ?

3 puntos

(El cambio de variable $x = \theta\,\cos y$ suele dar buenos resultados, así como la fórmula $\cos^2 x = (1+\cos 2x)/2$.)

Problema 2

Se quiere analizar si pueden considerarse significativamente independientes o no el tener problemas respiratorios en la adolescencia y el tener un historial de bronquitis en la infancia. Para ello se seleccionaron al azar 1000 adolescentes y se les clasificó según tuvieran en la actualidad problemas respiratorios o no y según su historial clínico de bronquitis en la infancia. Los resultados obtenidos fueron los siguientes:

		Historial de bronquitis		
		Sí	*No*	**Total**
Problemas respiratorios	*Sí*	**25**	**40**	
	No	**200**	**735**	
	Total			**1000**

¿Puede rechazarse la hipótesis nula de independencia entre ambas variables?

2 puntos

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Problema 1

a) Según PIE, pp. 279 y siguientes, debemos igualar tantos momentos (respecto al origen) poblacionales a los correspondientes muestrales, hasta obtener un número suficiente de ecuaciones de donde poder obtener una solución para el parámetro. Para ello, calculamos primero la media de X,

$$E[X] = \frac{2}{\theta^2\,\pi}\int_{-\theta}^{\theta} x\,\sqrt{\theta^2 - x^2}\,dx = \frac{2}{\theta^2\,\pi}\,\frac{[\theta^2 - x^2]^{1/2+1}}{3/2}\bigg|_{-\theta}^{\theta} = 0$$

con lo que la primera ecuación, $0 = \overline{x}$, no nos sirve. Para utilizar la segunda ecuación debemos calcular

$$\begin{aligned}
E[X^2] &= \frac{2}{\theta^2\,\pi}\int_{-\theta}^{\theta} x^2\,\sqrt{\theta^2 - x^2}\,dx \\
&= \frac{2\,\theta}{\pi}\int_{-\theta}^{\theta}\left(\frac{x}{\theta}\right)^2\sqrt{1-\left(\frac{x}{\theta}\right)^2}\,dx \\
&= \frac{2\,\theta^2}{\pi}\int_0^{\pi}\cos^2 y\ \mathrm{sen}^2 y\ \,dy \\
&= \frac{\theta^2}{2\,\pi}\int_0^{\pi}(1-\cos 2y)(1+\cos 2y)\,dy \\
&= \frac{\theta^2}{2\,\pi}\int_0^{\pi}(1-\cos^2 2y)\,dy \\
&= \frac{\theta^2}{2\,\pi}\int_0^{\pi}\mathrm{sen}^2 2y\,dy \\
&= \frac{\theta^2}{2\,\pi}\int_0^{\pi}\frac{1-\cos 4y}{2}\,dy \\
&= \frac{\theta^2}{4}.
\end{aligned}$$

Por tanto, la segunda ecuación a utilizar en el método de los momentos será

$$\frac{\theta^2}{4} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

de donde obtenemos (despejando) como estimador de θ,

$$T = 2\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}.$$

b) Por el apartado anterior, parece razonable calcular la esperanza de

$$T^2 = 4\,\frac{1}{n}\sum_{i=1}^{n} X_i^2$$

para lo que necesitamos, previamente, determinar la distribución de la variable $Y = X^2$, la cual tendrá como función de densidad (véase LCP, pp. 132, ejemplo 4.14.7),

$$g(y) = \frac{1}{2\sqrt{y}}\left(f_\theta(-\sqrt{y}) + f_\theta(\sqrt{y})\right) = \frac{2}{\theta^2\,\pi}\frac{\sqrt{\theta^2 - y}}{\sqrt{y}}$$

si $0 < y < \theta^2$. Por tanto, la densidad de $Z = X^2/\theta^2 = Y/\theta^2$, será

$$g(z) = \frac{2}{\pi}\,(1-z)^{3/2-1}\,z^{1/2-1}$$

si $0 < z < 1$; es decir, $Z = X^2/\theta^2 \rightsquigarrow \beta(1/2, 3/2)$, con lo que será

$$E\left[\frac{X^2}{\theta^2}\right] = \frac{1/2}{2} = \frac{1}{4}$$

y, por tanto,

$$E[T^2] = 4E[X^2] = 4\,\frac{\theta^2}{4} = \theta^2$$

es decir, T^2 ser el estimador insesgado para θ^2 buscado.

Obsérvese que se puede calcular $E[T^2]$ directamente a partir de la definición:

$$E[T^2] = \frac{4}{n}\sum_{i=1}^{n} E[X_i^2] = \frac{4}{n}\,n\,\frac{\theta^2}{4} = \theta^2.$$

No obstante, en la siguiente sección necesitaremos determinar $V(T^2)$ que resulta más fácil de determinar que calculando primero $E[T^4]$ y luego

$$V(T^2) = E[T^4] - \left(E[T^2]\right)^2 .$$

c) Como acabamos de ver que T^2 es un buen estimador para θ^2, parece razonable utilizarlo en la búsqueda de la cantidad pivotal.

Como T^2 es la media aritmética de las variables $4X_i^2$, (las cuales tienen media y varianza finita, entonces (véase, por ejemplo, LCP, pp. 430, ó PIE, pp. 46) será

$$T^2 \longrightarrow N\left(E[T^2]\,,\,\sigma_{T^2}\right)$$

en distribución, cuando $n \to \infty$, siendo

$$E[T^2] = \theta^2$$

y

$$\sigma_{T^2}^2 = V(T^2) = \frac{\theta^4}{n}$$

por ser $X^2/\theta^2 \rightsquigarrow \beta(1/2, 3/2)$. Es decir,

$$T^2 \longrightarrow N\left(\theta^2\,,\,\frac{\theta^2}{\sqrt{n}}\right).$$

Por tanto, (ver PIE, pp. 120) fijado un nivel de confianza $1-\alpha$, se pueden determinar en la tabla de la $N(0,1)$, dos valores tales que

$$P\left\{-z_{\alpha/2} < \frac{T^2-\theta^2}{\theta^2/\sqrt{n}} < z_{\alpha/2}\right\} \approx 1-\alpha$$

es decir,

$$P\left\{\frac{\sqrt{n}\,T^2}{\sqrt{n}+z_{\alpha/2}} < \theta^2 < \frac{\sqrt{n}\,T^2}{\sqrt{n}-z_{\alpha/2}}\right\} \approx 1-\alpha$$

con lo que un intervalo para θ^2, de confianza de coeficiente de confianza $1-\alpha$, será

$$\left[\frac{\sqrt{n}\,T^2}{\sqrt{n}+z_{\alpha/2}}\,,\,\frac{\sqrt{n}\,T^2}{\sqrt{n}-z_{\alpha/2}}\right]$$

y para θ,

$$\left[\frac{T}{\sqrt{1+\frac{z_{\alpha/2}}{\sqrt{n}}}}\,,\,\frac{T}{\sqrt{1-\frac{z_{\alpha/2}}{\sqrt{n}}}}\right].$$

Problema 2

Se trata de un contraste de independencia de caracteres (CB-sección 12.4) en donde la hipótesis nula es la independencia de ambas variables. Para realizar dicho contraste utilizaremos el estadístico λ de Pearson el cual mide las discrepancias entre las frecuencias observadas n_{ij} y las esperadas $n_{i\cdot}\, n_{\cdot j}/n$ en cada casilla, siendo, respectivamente, $n_{i\cdot}\ \ i = 1, ..., a$ y $n_{\cdot j}\ \ j = 1, ..., b$ los totales por filas y columnas de la tabla de doble entrada que contiene los datos. Dicho estadístico tiene por expresión,

$$\lambda = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(n_{ij} - n_{i\cdot}\, n_{\cdot j}/n)^2}{n_{i\cdot}\, n_{\cdot j}/n}$$

el cual sigue, aproximadamente, una distribución χ^2 con $(a-1)(b-1)$ grados de libertad.

La tabla de frecuencias esperadas es

		Historial de bronquitis		
		Sí	*No*	Total
Problemas respiratorios	*Sí*	14′6	50′4	65
	No	210′4	724′6	935
	Total	225	735	1000

siendo cada uno de los sumandos del estadístico λ a utilizar en el contraste χ^2 igual a

		Historial de bronquitis		
		Sí	*No*	Total
Problemas respiratorios	*Sí*	7′4	2′1	
	No	0′5	0′1	
	Total			10′2

A la vista de estos resultados, el estadístico λ de Pearson de distribución χ^2_1 antes de tomar la muestra, toma el valor $\lambda = 10'2$. (Con más precisión $\lambda = 10'157$).

Como no se especifica ningún nivel de significación en el enunciado se calcula el p-valor y si éste es muy pequeño se rechaza la hipótesis nula y si es relativamente grande se acepta. La hipótesis nula de independencia de ambos caracteres es rechazada al ser $P\{\chi^2_1 > 10'2\} < 0'005$.

De hecho, este razonamiento, aunque habitual entre los usuarios de la Estadística es algo informal. Lo correcto hubiera sido fijar un nivel de significación α —habitualmente 0′1, 0′05 ó 0′01— y para ese nivel determinar el punto

crítico. Si ahora es λ mayor que ese punto crítico, rechazaremos la hipótesis nula; luego calcularíamos el p-valor para valorar la decisión tomada de la forma antes mencionada. Lo que ocurre es que con el p-valor determinado en este ejemplo —menor que $0'005$— si se hubiera tomado otra decisión que no fuera el rechazo de H_0, ésta sería muy poco fiable. Además, el cálculo del p-valor ya nos da para qué niveles de significación se rechaza H_0 —los mayores que dicho p-valor— y para cuáles se acepta —los menores. En este caso, deberíamos haber elegido un nivel de significación mucho menor que $0'005$ para haber aceptado la hipótesis nula (elección absurda).

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2016-2017.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria discreta con valores naturales y función de probabilidad

$$p_\theta(x) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{x-2} \qquad x = 2, 3, 4, ...$$

siendo $\theta > 0$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Estimar θ por el método de la máxima verosimilitud y por el método de los momentos.

2.5 puntos

b) Determinar el estadístico suficiente minimal y su distribución en el muestreo.

2.5 puntos

c) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = \theta$. ¿Es eficiente el estimador obtenido?

2.5 puntos

Problema 2

Se está estudiando el tiempo de vida entre los pacientes a una determinada enfermedad. A tal fin se eligieron al azar 100 fichas de pacientes fallecidos por la enfermedad en estudio, obteniéndose una media muestral de 740 días y una cuasidesviación típica muestral de 32 días.

¿Puede admitirse para los pacientes de la enfermedad en cuestión un tiempo medio de vida superior a 730 días?

2.5 puntos

Problema 1

a) La función de verosimilitud de la muestra (véase PIE, pp. 284) será

$$p_\theta(x_1,...,x_n)=\prod_{i=1}^{n}p_\theta(x_i)=\left(\frac{1}{1+\theta}\right)^n\left(\frac{\theta}{1+\theta}\right)^{\sum_{i=1}^n x_i-2n}.$$

Tomando logaritmos y derivando respecto al parámetro, obtenemos la siguiente ecuación de verosimilitud,

$$\frac{\partial}{\partial\theta}\log p_\theta(x_1,...,x_n)=-\frac{n}{1+\theta}+\left[(\sum_{i=1}^{n}x_i)-2n\right]\frac{1}{\theta(1+\theta)}=0$$

de donde obtenemos como estimador de máxima verosimilitud

$$\hat{\theta}=\overline{x}-2.$$

(Véase también PIE, pp. 317, ejercicio 7.5.)

Para utilizar el método de los momentos (PIE, pp. 279) formamos la ecuación $E[X]=\overline{x}$, y despejamos de ella el parámetro. Por tanto, lo primero que debemos hacer es calcular la media de la distribución de X.

Aunque más adelante veremos que la variable X es una transformación de una geométrica o, más en general, de una binomial negativa, lo que permite obtener su media a partir de la media de dicha distribución, a continuación vamos a calcularla directamente. Dicho cálculo es similar al de todas las distribuciones de tipo discreto y, en particular, al de la distribución geométrica (véase LCP, pp. 292).

La media de la distribución será

$$E[X]=\sum_{x=2}^{\infty}x\,p_\theta(x)=\frac{1}{1+\theta}\sum_{x=2}^{\infty}x\left(\frac{\theta}{1+\theta}\right)^{x-2}.$$

Llamando

$$S=\sum_{x=2}^{\infty}x\left(\frac{\theta}{1+\theta}\right)^{x-2}=2+3\left(\frac{\theta}{1+\theta}\right)+4\left(\frac{\theta}{1+\theta}\right)^2+\cdots$$

será

$$S\left(\frac{\theta}{1+\theta}\right)=2\left(\frac{\theta}{1+\theta}\right)+3\left(\frac{\theta}{1+\theta}\right)^2+4\left(\frac{\theta}{1+\theta}\right)^3+\cdots$$

con lo que

$$\begin{aligned} S - S\left(\frac{\theta}{1+\theta}\right) &= 2 + \left(\frac{\theta}{1+\theta}\right) + \left(\frac{\theta}{1+\theta}\right)^2 + \left(\frac{\theta}{1+\theta}\right)^3 + \cdots \\ &= 1 + 1 + \left(\frac{\theta}{1+\theta}\right) + \left(\frac{\theta}{1+\theta}\right)^2 + \left(\frac{\theta}{1+\theta}\right)^3 + \cdots \\ &= 1 + \frac{1}{1-\theta/(1+\theta)} \\ &= 2 + \theta \end{aligned}$$

por la conocida fórmula de la suma de los infinitos términos de una progresión geométrica ilimitada de razón, en este caso, $\theta/(1+\theta)$.

Por tanto,

$$S \cdot \left(1 - \frac{\theta}{1+\theta}\right) = 2 + \theta$$

o bien,

$$S = (2+\theta)(1+\theta)$$

con lo que

$$E[X] = \frac{1}{1+\theta} \cdot S = \frac{1}{1+\theta} \cdot (2+\theta)(1+\theta) = 2 + \theta$$

y, por tanto, la ecuación para obtener el estimador de los momentos será

$$2 + \theta = \overline{x}$$

de donde se obtiene como estimador para θ

$$\hat{\theta} = \overline{x} - 2$$

es decir, el mismo que se obtenía por el método de la máxima verosimilitud.

b) Como la distribución modelo es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$p_\theta(x) = \frac{1+\theta}{\theta^2} \exp\{x \log[\theta/(1+\theta)]\}$$

será

$$T(X_1, ..., X_n) = \sum_{i=1}^{n} X_i$$

un estadístico minimal suficiente, resultado que también se podría obtener directamente a partir del teorema de factorización (PIE, pp. 167) y del análisis directo de minimalidad (PIE, pp. 171, teorema 5.2).

Respecto a su distribución en el muestreo, un simple cambio de variable demuestra que la función de masa de la variable $Y = X - 2$ es

$$p_\theta(y) = P\{Y = y\} = P\{X = y+2\} = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^y \qquad y = 0, 1, 2, ...$$

es decir, una binomial negativa de par metros 1 y $1/(1+\theta)$, es decir, $Y \rightsquigarrow BN(1, 1/(1+\theta))$. (De aquí se podría obtener directamente la media de Y y, por tanto, la de X.)

Al ser esta distribución reproductiva respecto al primer parámetro, será $T = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i + 2n = W + 2n$, siendo $W \rightsquigarrow BN(n, 1/(1+\theta))$, con lo que la función de masa o cuantía de T será, $t = 2n, 2n+1, 2n+2, ...$

$$\begin{aligned} P\{T = t\} &= P\{W + 2n = t\} = P\{W = t - 2n\} \\ &= \binom{n+t-2n-1}{t-2n}\left(\frac{1}{1+\theta}\right)^n\left(\frac{\theta}{1+\theta}\right)^{t-2n} \end{aligned}$$

c) Para determinar el ECUMV de cualquier función del parámetro, lo primero que necesitamos es determinar un estadístico suficiente y completo para la familia de distribuciones $p_\theta(x)$.

Como vimos en el apartado anterior, al ser la distribución modelo una familia de tipo exponencial el estadístico $T = \sum_{i=1}^{n} X_i$ es suficiente minimal. Además (PIE, pp. 273), T será completo si la imagen de $q(\theta) = \log(\theta/(1+\theta))$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$ la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, T será suficiente y completo. Ahora sólo necesitamos determinar una función de T que sea insesgada para θ, ya que entonces, por el teorema de Lehmann-Scheffé (PIE, pp. 218), ese estadístico será el ECUMV buscado.

Bien por el cálculo directo efectuado en el primer apartado, o bien por la distribución en el muestreo de T determinada en el segundo, es

$$E[T] = E\left[\sum_{i=1}^{n} X_i\right] = nE[X] = n(2+\theta)$$

con lo que el estimador insesgado para θ, función del suficiente y completo, será $T/n - 2$; es decir, el estimador de máxima verosimilitud o el de los momentos.

Respecto a su varianza, ésta será

$$V(\overline{x} - 2) = V(\overline{x}) = \frac{V(X)}{n} = \frac{V(Y)}{n} = \frac{\theta/(1+\theta)}{n/(1+\theta)^2} = \frac{\theta(1+\theta)}{n}$$

mientras que la cota de Frechet-Cramer-Rao es (PIE, pp. 225)

$$\frac{[g'(\theta)]^2}{I(\theta)} = \frac{1}{n/\theta(1+\theta)} = \frac{\theta(1+\theta)}{n} = V(\overline{x} - 2)$$

siendo, por tanto, el ECUMV eficiente.

Problema 2

Si representamos por X la variable aleatoria *tiempo de vida de los pacientes con la enfermedad en estudio*, y por μ su media, estamos interesados en analizar si puede admitirse la hipótesis $\mu > 730$ la cual, como siempre, se plantea como hipótesis alternativa H_1, reservando la hipótesis nula al suceso complementario $H_0 : \mu \leq 730$.

En la situación que nos movemos de contrastes para la media, μ, de una población no necesariamente normal de varianza desconocida siendo el tamaño muestral suficientemente grande (CB-sección 7.3), se rechaza $H_0 : \mu \leq 730$ cuando y sólo cuando sea

$$\frac{\overline{x} - 730}{S/\sqrt{n}} > z_\alpha$$

siendo z_α el valor de la abscisa de una normal $N(0,1)$ que deja a la derecha un área de probabilidad α, siendo α el nivel de significación del test.

Si fijamos como nivel de significación $\alpha = 0'05$, la Tabla 3 de la normal $N(0,1)$ nos proporciona el punto crítico $z_\alpha = z_{0'05} = 1'645$, al obtenerse a partir de la mencionada tabla que es $P\{Z > 1'64\} = 0'0505$ y $P\{Z > 1'65\} = 0'0495$. Al ser la probabilidad cola requerida como nivel de significación la semisuma de las dos anteriores, el punto crítico también será la semisuma de las dos abscisas anteriores: $(1'64 + 1'65)/2 = 1'645$.

Como es

$$\frac{\overline{x} - 730}{S/\sqrt{n}} = \frac{740 - 730}{32/\sqrt{100}} = 3'125 > 1'645 = z_{0'05}$$

rechazaremos la hipótesis nula de ser $H_0 : \mu \leq 730$, aceptando la alternativa $H_1 : \mu > 730$, de ser el tiempo medio de supervivencia entre los pacientes con la enfermedad en estudio, significativamente mayor de 730 días.

El p-valor del test es

$$P\left\{\frac{\overline{x}-730}{S/\sqrt{n}} > 3'125\right\} = P\{Z > 3'125\} = 0'0009$$

obtenido, de nuevo a partir de la tabla 3, por interpolación de dos valores (en este caso iguales). Un p-valor tan pequeño confirma la conclusión adoptada.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2016-2017.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{1}{2\theta_2} \qquad \text{si} \qquad \theta_1 - \theta_2 < x < \theta_1 + \theta_2$$

siendo $\theta = (\theta_1, \theta_2)$ un parámetro desconocido en el que $\theta_2 > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Estimar el parámetro bidimensional $\theta = (\theta_1, \theta_2)$ por el método de los momentos.

2 puntos

b) Determinar un estadístico suficiente minimal.

2 puntos

c) Determinar el estimador de máxima verosimilitud para el parámetro bidimensional $\theta = (\theta_1, \theta_2)$, así como su distribución en el muestreo.

2 puntos

d) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = \theta_2$, suponiendo que el estadístico $(X_{(1)}, X_{(n)})$ es suficiente y completo.

2 puntos

Problema 2

Se quiere analizar si puede admitirse que los niveles medios de colesterol en una población determinada se encuentran por debajo de 200 mg/dl. Para ello se tomó una muestra de 50 personas de dicha población que proporcionó una media de 196 mg/dl. y una cuasivarianza muestral igual a 90. Calcule el p-valor del test y diga las conclusiones que obtendría.

2 puntos

Problema 1

a) El método de los momentos (PIE, pp. 279) consiste en formar tantas ecuaciones (igualando los primeros momentos muestrales a los correspondientes poblacionales) como sean necesarias, en donde figuren como incógnitas los parámetros a estimar. En este caso basta con dos ecuaciones, obtenidas igualando los dos primeros momentos:

$$\left\{ \begin{array}{lcl} \overline{x} & = & E[X] \\ \sum_{i=1}^{n} X_i^2/n & = & E[X^2]. \end{array} \right.$$

Operando sobre el modelo o, simplemente observando que éste es una uniforme $U(\theta_1 - \theta_2, \theta_1 + \theta_2)$ se obtienen los dos primeros momentos poblacionales y, por tanto, el siguiente sistema de ecuaciones

$$\left\{ \begin{array}{lcl} \overline{x} & = & \theta_1 \\ \sum_{i=1}^{n} X_i^2/n & = & \theta_2^2/3 + \theta_1^2 \end{array} \right.$$

el cual, proporciona el estimador $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, siendo

$$\left\{ \begin{array}{l} \hat{\theta}_1 = \overline{x} \\ \\ \hat{\theta}_2 = \sqrt{3}\, s = \sqrt{3} \sqrt{\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (X_i - \overline{x})^2}. \end{array} \right.$$

b) La función de verosimilitud será

$$f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{(2\theta_2)^n}\, I_{\{\theta_1 - \theta_2 < x_{(1)} \leq x_{(n)} < \theta_1 + \theta_2\}} ==$$

$$= \frac{1}{(2g_3(\theta))^n}\, I_{\{g_1(\theta) < x_{(1)} \leq x_{(n)} < g_2(\theta)\}}$$

en donde g_1, g_2 y g_3 son las funciones del parámetro —bidimensional— θ siguientes, $g_1(\theta) = g_1(\theta_1, \theta_2) = \theta_1 - \theta_2$, $g_2(\theta) = g_2(\theta_1, \theta_2) = \theta_1 + \theta_2$, $g_3(\theta) = g_3(\theta_1, \theta_2) = \theta_2$. Por tanto, hemos descompuesto la verosimilitud en el producto de dos funciones, una (la identidad) y otra (todo lo demás) que depende del parámetro θ y del estadístico $(X_{(1)}, X_{(n)})$; éste es, por el teorema de factorización (PIE, pp. 167), un estadístico suficiente (véase también el ejemplo 5.13 de PIE, pp. 169). Además, es trivialmente minimal suficiente (véase el ejercicio 5.4 de PIE, pp. 198).

c) La función de verosimilitud

$$f_\theta(x_1,...,x_n)=\prod_{i=1}^{n} f_\theta(x_i)=\frac{1}{(2\theta_2)^n}\, I_{\{\theta_1-\theta_2<x_{(1)}\leq x_{(n)}<\theta_1+\theta_2\}}$$

es igual a cero excepto en los valores (θ_1,θ_2) dentro del recinto tales que $\theta_1-\theta_2<x_{(1)}\leq x_{(n)}<\theta_1+\theta_2$; es decir, es cero excepto cuando sea

$$\theta_1-\theta_2<x_{(1)}$$

y

$$x_{(n)}<\theta_1+\theta_2$$

en cuyo caso la función de verosimilitud toma el valor $1/(2\theta_2)^n$. Habrá que buscar, por tanto, el valor de θ_2 lo más pequeño posible dentro del recinto antes mencionado, valor que corresponde a la solución de las ecuaciones

$$\begin{cases} \hat\theta_1=\dfrac{X_{(n)}+X_{(1)}}{2} \\ \\ \hat\theta_2=\dfrac{X_{(n)}-X_{(1)}}{2}. \end{cases}$$

El par $(\hat\theta_1,\hat\theta_2)$ anterior es el estimador de máxima verosimilitud de $\theta=(\theta_1,\theta_2)$. Para calcular su distribución en el muestreo partimos de la distribución conjunta de $(X_{(1)},X_{(n)})$ (PIE, pp. 19)

$$\begin{aligned} h(s,t) &= n(n-1)\left[F_\theta(t)-F_\theta(s)\right]^{n-2} f_\theta(t) f_\theta(s) \\ &= n(n-1)\frac{[t-s]^{n-2}}{(2\theta_2)^n} \end{aligned}$$

en el recinto $s\leq t\,,\ \theta_1-\theta_2<s<\theta_1+\theta_2\ \ \theta_1-\theta_2<t<\theta_1+\theta_2\,$, al ser la función de distribución del modelo,

$$F_\theta(x)=\frac{1}{2\theta_2}(x-\theta_1+\theta_2) \qquad \text{si} \qquad \theta_1-\theta_2<x<\theta_1+\theta_2.$$

La densidad de $(\hat\theta_1,\hat\theta_2)$ será ahora, (LCP, pp. 233)

$$\begin{aligned} l(u,v) &= h(u-v,u+v)\cdot|2| \\ &= \frac{n(n-1)}{2\theta_2^n}\, v^{n-2} \end{aligned}$$

en el recinto $0 < v < \theta_2\,,\ v + \theta_1 - \theta_2 < u < -v + \theta_1 + \theta_2\,.$

d) A partir de la distribución conjunta determinada en el apartado anterior, obtenemos la distribución marginal de $(X_{(n)} - X_{(1)})/2$

$$l(v) = \frac{n(n-1)}{\theta_2^n}\, v^{n-2}(\theta_2 - v) \qquad \text{si} \qquad 0 < v < \theta_2$$

de media

$$E\left[\frac{X_{(n)} - X_{(1)}}{2}\right] = \frac{(n-1)\theta_2}{n+1}$$

con lo que un estimador insesgado para θ_2 será

$$T = \frac{X_{(n)} - X_{(1)}}{2(n-1)}\,(n+1).$$

Como es función de $(X_{(1)}, X_{(n)})$, por hipótesis suficiente y completo, T será el ECUMV para θ_2 (teorema de Lehmann-Scheffé, PIE, pp. 218).

Otra forma de haber actuado sería la de determinar las distribuciones marginales y sus medias de $X_{(1)}$ y de $X_{(n)}$, ya que el enunciado nos da la pista de que el ECUMV buscado debe ser función de estos estadísticos. Se hubiera obtenido el mismo resultado.

Problema 2

Del enunciado se desprende que se quiere contrastar la hipótesis nula $H_0 : \mu \geq 200$ frente a la alternativa $H_1 : \mu < 200$. Estamos en un caso de contrastes para la media de una población no necesariamente normal y muestras grandes (CB-sección 7.3) rechazándose la hipótesis nula cuando y sólo cuando sea

$$\frac{\overline{x} - \mu_0}{S/\sqrt{n}} < z_{1-\alpha}.$$

Dado que no nos dan nivel de significación, vamos a calcular el p-valor del test. El estadístico del contraste toma el valor:

$$\frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{196 - 200}{\sqrt{90/50}} = -2'98$$

y como es un test unilateral con región crítica la cola de la izquierda, el p-valor será, a partir de una tablas de la distribución normal

$$\text{p-valor} = P\{Z < -2'98\} = P\{Z > 2'98\} = 0'0014$$

pudiendo rechazarse con bastante seguridad la hipótesis nula y concluir que el nivel medio de colesterol de la población en estudio sí puede establecerse en menos de 200.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

Curso 2017-2018

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2017-2018.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \theta\,(1+\theta)\,\left(e^{-\theta\,x} - e^{-(\theta+1)\,x}\right) \qquad x > 0$$

siendo $\theta > 0$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

a) Determinar un estadístico suficiente minimal y analizar su completitud.

2 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para

$$g(\theta) = \frac{1+2\,\theta}{\theta(1+\theta)}\,.$$

3 puntos

Problema 2

Sea X una variable aleatoria positiva con función de densidad $f_0(x) = e^{-x}$, $x > 0$ bajo la hipótesis nula H_0 y con función de densidad $f_1(x) = 2\,e^{-x}/[1+e^{-x}]^2$, $x > 0$, bajo la hipótesis alternativa H_1. Determinar un contraste de máxima potencia de nivel $\alpha = 0'055$ para contrastar H_0 frente a H_1, utilizando una muestra aleatoria simple de tamaño dos de X.

3 puntos

Problema 3

Se cree que la frecuencia con la que las mujeres realizan una auto-exploración del pecho depende mucho de su Edad. En un estudio (Senie, Rosen, Lesser y Kinne, 1981) se obtuvieron los siguientes datos:

	Frecuencia de la auto-exploración		
Edad	Mensualmente	Ocasionalmente	Nunca
Menos de 45	91	90	51
Entre 45 y 59	150	200	155
60 o más	109	198	172

A la vista de estos datos, ¿cree que efectivamente la auto-exploración depende de la Edad?

2 puntos

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Problema 1

a) Como la distribución modelo es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$f_\theta(x) = \theta\,(1+\theta)\,(1-e^{-x})\,e^{-\theta\,x}$$

será $S(X_1,...,X_n) = \sum_{i=1}^n X_i$ un estadístico minimal suficiente. Por tanto, (PIE, pp. 273), S será además completo si la imagen de $q(\theta) = -\theta$ contiene un abierto de $\mathbb{R}$. Como es $\theta > 0$, la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$. En definitiva, S será suficiente y completo.

b) Basta ahora con encontrar un estimador, función del suficiente y completo S, que sea insesgado para

$$g(\theta) = \frac{1+2\theta}{\theta(1+\theta)}$$

y ese será el ECUMV para $g(\theta)$. Comencemos calculando la esperanza de S. Como es $E[S(X_1,...,X_n)] = E[\sum_{i=1}^n X_i] = n\,E[X]$, calcularemos ésta última, utilizando la técnica de la integración por partes.

$$\begin{aligned}
E[X] &= \theta(1+\theta)\int_0^\infty x\,(1-e^{-x})\,e^{-\theta\,x}\,dx \\
&= \theta(1+\theta)\left(\int_0^\infty x\,e^{-\theta x}\,dx - \int_0^\infty x\,e^{-(\theta+1)x}\,dx\right) \\
&= \theta(1+\theta)\left(\frac{1}{\theta^2} - \frac{1}{(1+\theta)^2}\right) \\
&= \frac{1+2\,\theta}{\theta(1+\theta)}.
\end{aligned}$$

Por tanto, el ECUMV para $g(\theta)$ será

$$\frac{S}{n} = \overline{x}.$$

Problema 2

El lema de Neyman-Pearson nos dice (PIE, pp. 338) que el test de máxima potencia, para contrastar una hipótesis nula simple frente a una alternativa simple, como las aquí planteadas, tiene por región crítica la siguiente

$$f_1(x_1, ..., x_n) > k f_0(x_1, ..., x_n)$$

es decir, si $x_1, ..., x_n > 0$,

$$\frac{2^n \, e^{-\sum x_i}}{\prod_{i=1}^n [1 + e^{-x_i}]^2} > k \; e^{-\sum_{i=1}^n x_i}$$

o lo que es lo mismo,

$$\prod_{i=1}^n [1 + e^{-x_i}]^2 < d$$

o bien,

$$T_n = \sum_{i=1}^n \log[1 + e^{-x_i}] < c$$

siendo, por tanto, el test buscado el dado por

$$\varphi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^n \log[1 + e^{-x_i}] < c \\ 0 & \text{si} \quad \sum_{i=1}^n \log[1 + e^{-x_i}] \geq c \end{cases}$$

en donde la constante c es tal que

$$P_{H_0}\{T_n < c\} = \alpha.$$

Un sencillo cambio de variable permite calcular la función de densidad de la variable $Y = \log[1 + e^{-X}]$. Es

$$g(y) = e^y \qquad\qquad 0 < y < \log 2$$

distribución no muy habitual. No obstante, se puede determinar el valor de la constante c, en el caso de muestras de tamaño dos, directamente, ya que c es el valor tal que

$$\begin{aligned} \alpha = P\{Y_1 + Y_2 < c\} &= \int_{\mathcal{T}} g(y_1, y_2) \; dy_1 \, dy_2 \\ &= \int_{\mathcal{T}} g(y_1) g(y_2) \; dy_1 \, dy_2 \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{T}} e^{y_1}\, e^{y_2}\, dy_1\, dy_2 \\
&= \int_0^c e^{y_1} \left(\int_0^{c-y_1} e^{y_2}\, dy_2 \right) dy_1 \\
&= \int_0^c (e^c - e^{y_1})\, dy_1 \\
&= (c-1)\, e^c + 1.
\end{aligned}
$$

En el caso de ser $\alpha = 0'055$, de la ecuación

$$0'055 = (c-1)\, e^c + 1$$

se obtiene el valor $c = 0'3$, quedando el test de máxima potencia igual a

$$\varphi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^2 \log[1 + e^{-x_i}] < 0'3 \\ 0 & \text{si} \quad \sum_{i=1}^2 \log[1 + e^{-x_i}] \geq 0'3 \end{cases}$$

Problema 3

Se trata de un contraste de independencia de caracteres (CB-sección 12.4) en donde la hipótesis nula es que ambas variables: Edad y Frecuencia de la auto-exploración, son independientes y la hipótesis alternativa es que no son independientes.

El estadístico de Pearson toma el valor

$$\lambda = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(n_{ij} - n_i m_j / n)^2}{n_i m_j / n} = 25'086.$$

El p-valor del test será:

$$P\{\chi_4^2 > 25'086\}$$

acotado por

$$P\{\chi_4^2 > 25'086\} < P\{\chi_4^2 > 14'86\} = 0'005$$

es decir, el p-valor será menor que $0'005$, suficientemente pequeño como para rechazar la hipótesis nula de independencia y concluir con que sí existe relación significativa entre ambas variables.

Si quisiéramos calcular el valor del estadístico de Pearson con el software R utilizaríamos la siguiente secuencia:

```
> X<-matrix(c(91,150,109,90,200,198,51,155,172),ncol=3)                    (1)
> colnames(X)<-c("Mensualmente","Ocasionalmente","Nunca")
> rownames(X)<-c("Menos de 45","Entre 45 y  59","60 o más")

> X
              Mensualmente Ocasionalmente Nunca
Menos de 45             91             90    51
Entre 45 y  59         150            200   155
60 o más               109            198   172

> chisq.test(X,correct=FALSE)                                              (2)

        Pearson's Chi-squared test

data:  X
X-squared = 25.086, df = 4, p-value = 4.835e-05
                                      (3)

> chisq.test(X)                                                            (4)

        Pearson's Chi-squared test

data:  X
X-squared = 25.086, df = 4, p-value = 4.835e-05
```

Primero introducimos los datos según se indica en (1); después se ejecuta (2) y, finalmente obtenemos el valor del estadístico de contraste y el p-valor en (3).

Observamos que, como la tabla no es 2×2, si no decimos nada, como ocurre en (4), R no calcula el valor del estadístico con la corrección de Yates sino el correcto.

Por otro lado, esta corrección no es necesaria porque las frecuencias esperadas son todas mayores que 5,

```
> chisq.test(X)$observed
              Mensualmente Ocasionalmente Nunca
Menos de 45             91             90    51
Entre 45 y  59         150            200   155
60 o más               109            198   172
```

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2017-2018.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{1}{x \ \log\theta} \qquad\qquad 1 < x < \theta$$

siendo $\theta > 1$ un parámetro desconocido (log = logaritmo neperiano). Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

(a) Estimar θ por el método de la máxima verosimilitud y analizar si el estimador obtenido es suficiente minimal.

2 puntos

(b) Analizar si el estimador obtenido en el apartado anterior es o no completo.

2 puntos

(c) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = \log\theta$.

2 puntos

(d) Determinar un test uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

2 puntos

Problema 2

Se quiere estimar, mediante un intervalo de confianza construido por el método de la cantidad pivotal, la proporción de personas de una población que nunca ha usado un ordenador. Para ello se seleccionó una muestra aleatoria simple de 200 personas de la población en estudio, en la que 104 personas jamás había usado un ordenador. Determinar el intervalo de confianza buscado, con coeficiente de confianza del 95 %.

2 puntos

Problema 1

a) La función de verosimilitud de la muestra (véase PIE, pp. 283)

$$L(\theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{1}{\log\theta}\right)^n \frac{1}{\prod_{i=1}^n x_i} I_{\{x_{(n)}<\theta\}} I_{\{x_{(1)}>1\}}$$

se anula para $\theta < x_{(n)}$ y decrece a medida que avanza por el intervalo $[x_{(n)}\,,\,\infty)$, por lo que, como función de θ, alcanza su máximo en el extremo del intervalo, $\theta = x_{(n)}$. El estimador de máxima verosimilitud para θ será, por tanto,

$$\widehat{\theta} = X_{(n)}.$$

Como es

$$\frac{f_\theta(x_1, ..., x_n)}{f_\theta(x'_1, ..., x'_n)} = \left(\prod_{i=1}^{n} \frac{x'_i}{x_i}\right) \cdot \frac{I_{\{x_{(n)}<\theta\}}}{I_{\{x'_{(n)}<\theta\}}} \cdot \frac{I_{\{x_{(1)}>1\}}}{I_{\{x'_{(1)}>1\}}}$$

este cociente no dependerá de θ cuando y sólo cuando sea $X_{(n)} = X'_{(n)}$, con lo que $X_{(n)}$ será suficiente minimal (PIE, pp. 170 y siguientes).

b) La función de densidad del máximo es

$$f(y) = n\ (F_\theta(y))^{n-1}\ f_\theta(y) = \frac{n}{(\log\theta)^n}\,(\log y)^{n-1}\,\frac{1}{y} \qquad\qquad 1 < y < \theta$$

con lo que si h es una función no dependiente de θ para la que es

$$E_\theta[h(X_{(n)}] = 0 \qquad\qquad \forall\,\theta > 1$$

es decir,

$$\int_1^\theta h(y)\,\frac{n}{(\log\theta)^n}\,(\log y)^{n-1}\,\frac{1}{y}\,dy = 0 \qquad\qquad \forall\,\theta > 1$$

o bien

$$\int_1^\theta h(y)\,\frac{(\log y)^{n-1}}{y}\,dy = 0 \qquad\qquad \forall\,\theta > 1$$

derivando respecto a θ, deberá ser

$$h(\theta)\,\frac{(\log\theta)^{n-1}}{\theta} = 0 \qquad\qquad \forall\,\theta > 1$$

y como es $(\log\theta)^{n-1}/\theta > 0$ por ser $\theta > 1$, será $h(\theta) = 0 \quad \forall\, \theta > 1$, con lo que $X_{(n)}$ será completo para la familia dada. (PIE, pp. 216-218.)

c) Para determinar el ECUMV de $g(\theta) = \log\theta$ debemos encontrar una función del suficiente y completo, $X_{(n)}$, insesgada para $\log\theta$. Por la forma de la función de densidad del máximo vamos a probar con $\log X_{(n)}$.

$$E[\log X_{(n)}] = \frac{n}{(\log\theta)^n} \int_1^\theta \log y \, \frac{(\log y)^{n-1}}{y}\, dy = \frac{n}{n+1}\, \log\theta.$$

Por tanto, el ECUMV de $g(\theta) = \log\theta$ será

$$\frac{n+1}{n}\, \log X_{(n)}.$$

d) Para $\theta < \theta'$, la razón de verosimilitudes

$$\frac{f_{\theta'}(x_1, ..., x_n)}{f_\theta(x_1, ..., x_n)} = \left(\frac{\log\theta}{\log\theta'}\right)^n \frac{I_{\{x_{(n)} < \theta'\}}}{I_{\{x_{(n)} < \theta\}}}$$

sólo está definida para $x_{(n)} < \theta'$, pero es función creciente de $x_{(n)}$, ya que toma el valor constante $(\log\theta/\log\theta')^n$ si $x_{(n)} < \theta < \theta'$ y vale $+\infty$ si $\theta < x_{(n)} < \theta'$. Por tanto, el teorema de Karlin-Rubin, (PIE, pp. 349) nos garantiza la existencia de un contraste uniformemente de máxima potencia, ϕ, para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual tiene región crítica de la forma $X_{(n)} > c$, en donde c se elige para que ϕ tenga tamaño α; es decir, tal que

$$\alpha = P_{\theta_0}\left\{X_{(n)} > c\right\} = 1 - \left(\frac{\log c}{\log\theta_0}\right)^n$$

al ser la función de distribución del máximo $X_{(n)}$

$$F(y) = [F_\theta(y)]^n = \left(\frac{\log y}{\log\theta}\right)^n$$

obteniéndose el valor

$$c = \theta_0 \, \exp\{(1-\alpha)^{1/n}\}.$$

El contraste buscado será, por tanto,

$$\phi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad x_{(n)} > \theta_0 \, \exp\{(1-\alpha)^{1/n}\} \\ \\ 0 & \text{si} \quad x_{(n)} \leq \theta_0 \, \exp\{(1-\alpha)^{1/n}\} \end{cases}$$

Problema 2

El problema en estudio (véase el problema 13.2 de PREB, pp. 253) se puede modelizar mediante una variable dicotómica, es decir, mediante una variable que tome sólo dos valores, $X = 1$ si el individuo cumple la condición en estudio, *no ha usado nunca un ordenador*, y $X = 0$ si el individuo no cumple la condición en estudio, es decir, *ha usado alguna vez un ordenador*.

Modelizado así el problema, la distribución de tal variable aleatoria será binomial $B(1, p)$ siendo p la proporción de individuos de la población que nunca ha usado un ordenador, es decir, el parámetro cuyo intervalo queremos determinar por el método de la cantidad pivotal.

El libro del curso ya nos indican cuál es dicha cantidad pivotal (PIE, pp. 122) y, de hecho, determinan el intervalo de confianza solicitado para la media (en este modelo p) de una población no normal, siendo el tamaño muestral suficientemente grande (de al menos 100 individuos). Se trata de (ver también CB-sección 6.3 o ADD, pp. 15)

$$I = \left[\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[0'52 \mp 1'96\sqrt{\frac{0'52 \cdot 0'48}{200}}\right] = [0'45076, 0'58924]$$

al ser la proporción muestral $\hat{p} = 104/200 = 0'52$ y, a partir de la tabla 3 (ADD-pp. 33) de la distribución normal $N(0, 1)$, ser $z_{\alpha/2} = z_{0'05/2} = z_{0'025} = 1'96$.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2017-2018.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{2\,(1-x)}{(1-\theta)^2} \qquad\qquad x \in (\theta\,,\,1)$$

siendo $0 < \theta < 1$. Utilizando una muestra aleatoria simple de tamaño n de X, determinar:

(a) El estimador de máxima verosimilitud para θ.

2 puntos

(b) Un estadístico suficiente minimal para la familia de densidades dada y analizar su completitud.

2 puntos

(c) El estimador centrado uniformemente de mínima varianza, si existe, para $g(\theta) = \theta$.

2 puntos

(d) Un intervalo de confianza para θ, de coeficiente de confianza $1-\alpha$, utilizando el método de Neyman. ¿Cuál es el de longitud mínima?

2 puntos

Problema 2

Se ha realizado un estudio sobre los niveles de radiación de un determinado modelo de pantalla, midiéndose la radiación en 10 pantallas de ese modelo elegidas al azar, de donde se obtuvo una cuasivarianza muestral de $S^2 = 402$. Suponiendo que la radiación de las pantallas sigue una distribución normal, contrastar, mediante un test insesgado uniformemente de máxima potencia de nivel $\alpha = 0'05$, la hipótesis nula de que la varianza poblacional es mayor o igual que 1000?

2 puntos

Problema 1

a) La función de verosimilitud de la muestra (PIE, pp. 283)

$$L(\theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{2^n}{(1-\theta)^{2n}} \left[\prod_{i=1}^{n}(1-x_i)\right] I_{\{x_{(n)}<1\}} I_{\{x_{(1)}>\theta\}}$$

como función de θ que es, se anula si θ es menor que 0 ó mayor que $x_{(1)}$, y crece en el intervalo $(0, x_{(1)})$, por lo que alcanza su máximo en el extremo del intervalo, $\theta = x_{(1)}$. El estimador de máxima verosimilitud para θ será, por tanto,

$$\widehat{\theta} = X_{(1)}.$$

b) Como es

$$\frac{f_\theta(x_1, ..., x_n)}{f_\theta(x'_1, ..., x'_n)} = \left[\prod_{i=1}^{n} \frac{(1-x_i)}{(1-x'_i)}\right] \cdot \frac{I_{\{x_{(n)}<1\}}}{I_{\{x'_{(n)}<1\}}} \cdot \frac{I_{\{x_{(1)}>\theta\}}}{I_{\{x'_{(1)}>\theta\}}}$$

este cociente no dependerá de θ cuando y sólo cuando sea $x_{(1)} = x'_{(1)}$, con lo que $X_{(1)}$ será suficiente minimal (PIE, pp. 170 y siguientes).

Para analizar su completitud debemos determinar primero su distribución. La función de distribución y densidad del mínimo son, respectivamente,

$$G_\theta(y) = 1 - [1 - F_\theta(y)]^n$$

y

$$g_\theta(y) = n\ [1 - F_\theta(y)]^{n-1}\ f_\theta(y)$$

siendo f_θ y F_θ las funciones de densidad y distribución del modelo. Como es

$$F_\theta(x) = 1 - \frac{(1-x)^2}{(1-\theta)^2} \qquad x \in (\theta, 1)$$

serán,

$$G_\theta(y) = 1 - \left[\frac{1-y}{1-\theta}\right]^{2n} \qquad y \in (\theta, 1)$$

y

$$g_\theta(y) = \frac{2\,n}{1-\theta}\ \left[\frac{1-y}{1-\theta}\right]^{2n-1} \qquad y \in (\theta, 1)$$

La condición de completitud (PIE, pp. 216) es

$$E[h(X_{(1)})] = 0 \qquad \forall\theta \in (0,1)$$

es decir,

$$\int_\theta^1 h(y)\,(1-y)^{2n-1}\,dy = 0 \qquad \forall\theta \in (0,1)$$

de donde derivando respecto a θ, resulta

$$h(\theta)\,(1-\theta)^{2n-1} = 0 \qquad \forall\theta \in (0,1)$$

y como es $1-\theta > 0$, para que se cumpla condición anterior, deberá ser

$$h(\theta) = 0 \qquad \forall\theta \in (0,1)$$

probando la completitud del mínimo.

c) Si existe, el ECUMV de $g(\theta) = \theta$ deberá ser una función del suficiente y completo, $X_{(1)}$, insesgada de θ. Para empezar, vamos a calcular su esperanza, intregrando por partes

$$E[X_{(1)}] = \frac{2\,n}{(1-\theta)^{2n}} \int_\theta^1 y\,(1-y)^{2n-1}\,dy = \frac{2\,n\,\theta + 1}{2\,n+1}.$$

Por tanto, el ECUMV para $g(\theta) = \theta$ será

$$\frac{(2n+1)X_{(1)} - 1}{2n}$$

ya que es función del suficiente y completo, $X_{(1)}$, e insesgado para θ.

d) El método de Neyman para determinar intervalos de confianza (PIE, sección 4.4) consiste, primero en determinar dos funciones del parámetro $c_1(\theta) < c_2(\theta)$ tales que

$$P\{X_{(1)} < c_1(\theta)\} = \alpha_1$$

$$P\{X_{(1)} > c_2(\theta)\} = \alpha_2$$

siendo $\alpha_1 + \alpha_2 = \alpha$. Estas ecuaciones, dado que hemos determinado más arriba la función de distribución del mínimo, serán

$$1 - \left(\frac{1-c_1}{1-\theta}\right)^{2n} = \alpha_1$$

$$\left(\frac{1-c_2}{1-\theta}\right)^{2n} = \alpha_2$$

de donde se obtienen las funciones

$$c_1(\theta) = 1 - (1-\theta)\,(1-\alpha_1)^{1/(2n)}$$

$$c_2(\theta) = 1 - (1-\theta)\,\alpha_2^{1/(2n)}$$

Intersecando ahora estas rectas (en θ) con una ordenada $X_{(1)} = x_{(1)}$, obtenemos las ecuaciones (en θ)

$$x_{(1)} = 1 - (1-\theta)\,(1-\alpha_1)^{1/(2n)}$$

$$x_{(1)} = 1 - (1-\theta)\,\alpha_2^{1/(2n)}$$

de donde se obtienen los extremos del intervalo

$$\left[1 - \frac{1-x_{(1)}}{\alpha_2^{1/(2n)}}\ ,\ 1 - \frac{1-x_{(1)}}{(1-\alpha_1)^{1/(2n)}}\right].$$

Su longitud (creciente en α_2),

$$\left(1-x_{(1)}\right)\left(\alpha_2^{-1/(2n)} - (1-\alpha+\alpha_2)^{-1/(2n)}\right)$$

se hace mínima (véase, por ejemplo el análogo ejercicio 4.1 de PIE, pp. 126) cuando es $\alpha_2 = \alpha$ y $\alpha_1 = 0$, resultando que el intervalo

$$\left[1 - \frac{1-x_{(1)}}{\alpha^{1/(2n)}}\ ,\ x_{(1)}\right]$$

es el de longitud mínima.

Problema 2 (Véase el problema 5.2 de PREB, pp. 103)

Llamando X a la variable aleatoria *niveles de radiación del modelo de pantalla en estudio*, del enunciado se deduce que es $X \rightsquigarrow N(\mu, \sigma)$, con μ y σ desconocidas. Además, el contraste que se solicita es del tipo $H_0 : \sigma^2 \geq \sigma_0^2$ frente a $H_1 : \sigma^2 < \sigma_0^2$.

En estas condiciones, el test insesgado uniformemente de máxima potencia es el que tiene como región crítica (véase PIE, pp. 362)

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1;1-\alpha}$$

(puede verse también CB-sección 7.4) es decir, se rechazar H_0 cuando y sólo cuando sea

$$\frac{9 \cdot 402}{1000} = 3'618 < \chi^2_{9;0'95}.$$

Como, a partir de la tabla 4 de la χ^2 (ADD, pp. 34), es $\chi^2_{9;0'95} = 3'325$ no puede rechazarse H_0 al nivel de significación propuesto, aceptándose en consecuencia.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

Curso 2018-2019

Pruebas Presenciales

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Primera semana. Curso 2018-2019.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Adenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = 2\,\theta\,x\,e^{-\theta x^2} \qquad\qquad x > 0$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:

(a) Determinar un estadístico suficiente minimal y analizar su completitud.

2 puntos

(b) Determinar el estimador centrado uniformemente de mínima varianza para $g_1(\theta) = 1/\theta$ y para $g_2(\theta) = \theta$.

2 puntos

(c) Determinar un intervalo de confianza de colas iguales para θ, de coeficiente de confianza $1-\alpha$, por el método de la cantidad pivotal.

2 puntos

(d) Suponiendo ahora que el parámetro θ de la distribución f_θ es una variable aleatoria con distribución gamma $\gamma(p, a)$, con $p, a > 0$, determinar el estimador Bayes de θ, suponiendo una función de pérdida cuadrática.

2 puntos

Problema 2

Se investigan los puntos de fusión de dos aleaciones utilizadas en la fabricación de soldadura. Para ello se funden 20 muestras de cada material, obteniendo para la primera aleación los valores $\overline{x}_1 = 421$ y $S_1 = 4$ grados Fahrenheit para, respectivamente, la media y cuasidesviación típica muestrales y, $\overline{x}_2 = 426$ y $S_2 = 3$ grados

Fahrenheit para la media y cuasidesviación típica muestrales de la segunda aleación. ¿Apoyan estos datos la afirmación de que las dos aleaciones tiene el mismo punto de fusión? Utilice $\alpha = 0'05$ y suponga que ambas poblaciones tienen distribución normal y las mismas desviaciones estándar. Determine también una acotación del p-valor y valore la decisión tomada.

2 puntos

Las soluciones que se exponen a continuación son, en muchos casos, excesivamente detalladas para un examen pero, con ellas, tratamos de que sirvan no sólo de soluciones de un examen sino para la formación de futuros alumnos.

Problema 1

a) Como la distribución modelo es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$f_\theta(x) = 2\,\theta\,x\,e^{-\theta\,x^2}$$

será $S(X_1, ..., X_n) = \sum_{i=1}^n X_i^2$ un estadístico minimal suficiente. Por tanto, (PIE, pp. 273), S será además completo si la imagen de $q(\theta) = -\theta$ contiene un abierto de $\mathbb{R}$. Como es $\theta > 0$, la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$. En definitiva, S será suficiente y completo.

b) Basta con encontrar un estimador, función del suficiente y completo S, que sea insesgado para la correspondiente función del parámetro y ese será el ECUMV buscado.

Comencemos calculando la esperanza de S. (Calculada en PIE, pp. 323.)

$$E[S(X_1, ..., X_n)] = E\left[\sum_{i=1}^n X_i^2\right] = n\,E[X^2] = n\,\frac{1}{\theta}.$$

Por tanto, el ECUMV para $g_1(\theta) = 1/\theta$ será

$$\frac{S}{n} = \frac{\sum_{i=1}^n X_i^2}{n}.$$

El resultado anterior sugiere que

$$\hat{\theta} = \frac{n}{S} = \frac{n}{\sum_{i=1}^n X_i^2}$$

sea el ECUMV para $g_2(\theta) = \theta$. Sabemos que es función del suficiente y completo. Calculemos su esperanza.

Después de un sencillo cambio de variable obtenemos que es $X^2 \rightsquigarrow \gamma(1, \theta)$, por lo que, al ser esta distribución reproductiva respecto al primer parámetro (LCP, pp. 311), será

$$\sum_{i=1}^n X_i^2 = S \rightsquigarrow \gamma(n, \theta)$$

por lo que

$$E[\hat{\theta}] = n \int_0^\infty \frac{1}{y} \frac{y^{n-1}\,\theta^n\,e^{-\theta y}}{\Gamma(n)}\,dy = \frac{n\,\theta}{n-1}$$

que conduce a que el ECUMV para $g_2(\theta) = \theta$ sea

$$\frac{n-1}{\sum_{i=1}^n X_i^2}.$$

c) El método de la cantidad pivotal para la construcción de intervalos de confianza (PIE, pp. 96), parte de la determinación de una cantidad pivotal; es decir, de un estadístico dependiente del parámetro cuyo intervalo queremos calcular, cuya distribución en el muestro no dependa de parámetros desconocidos.

Como es $\sum_{i=1}^n X_i^2 = S \rightsquigarrow \gamma(n,\theta)$, un sencillo cambio de variable conduce a que sea $2\theta S \rightsquigarrow \chi^2_{2n}$, por lo que este estadístico será la cantidad pivotal a utilizar ya que podremos ahora determinar, en las tablas de una χ^2, dos valores tales que sea

$$P\left\{\chi^2_{2n;1-\alpha/2} < 2\theta S < \chi^2_{2n;\alpha/2}\right\} = 1-\alpha$$

de donde obtendremos que es

$$P\left\{\frac{\chi^2_{2n;1-\alpha/2}}{2S} < \theta < \frac{\chi^2_{2n;\alpha/2}}{2S}\right\} = 1-\alpha$$

y, por tanto,

$$\left[\frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^n X_i^2}\,,\ \frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^n X_i^2}\right]$$

el intervalo de confianza buscado.

d) El estimador Bayes bajo pérdida cuadrática es (PIE, pp. 181) la media de la distribución a posteriori por lo que vamos a determinar ésta en primer lugar.

La *distribución de la muestra condicionada por el parámetro* es la obtenida a partir del modelo,

$$f_\theta(x_1,...,x_n) = \prod_{i=1}^n f_\theta(x_i) = 2^n\,\theta^n \left(\prod_1^n x_i\right) e^{-\theta\sum_1^n x_i^2} \qquad x_1,...,x_n > 0.$$

Según el enunciado, la *distribución a priori* del parámetro es $\gamma(p,a)$, es decir,

$$\pi(\theta) = \frac{a^p\,\theta^{p-1}\,e^{-a\theta}}{\Gamma(p)} \qquad\qquad \theta > 0$$

por lo que la *distribución a posteriori* (del parámetro condicionada por la muestra) $\pi(\theta/x_1,...,x_n)$ será igual a una constante (que denominamos k y que se determinará de forma que $\pi(\theta/x_1,...,x_n)$ integre 1 para que sea una densidad), que multiplica al producto de la distribución de la muestra condicionada por el parámetro, por la distribución a priori. (De hecho k no es mas que la integral del producto de estas dos funciones, pero ya veremos que no va a ser necesario calcularla explícitamente en la mayoría de los casos, puesto que el producto de las dos densidades mencionadas dar lugar, en general, a una distribución conocida):

$$\pi(\theta/x_1,...,x_n) = k{\cdot}f_\theta(x_1,...,x_n){\cdot}\pi(\theta) = k{\cdot}2^n\,\theta^n\left(\prod_1^n x_i\right)e^{-\theta\sum_1^n x_i^2}{\cdot}\frac{a^p\,\theta^{p-1}\,e^{-a\theta}}{\Gamma(p)} =$$

$$= c\cdot\theta^{p+n-1}\cdot e^{-(a+S)\theta}$$

siendo c una constante que incluye todas las constantes de la expresión (es decir, todo lo que no dependa de la variable θ, como por ejemplo k) y donde es $S = \sum x_i^2$.

De ahí se deduce (si quiere el lector puede determinar explícitamente la constante c para comprobarlo, aunque resulta innecesario) que la distribución a posteriori es una gamma $\gamma(p+n, a+S)$ y su media, el estimador Bayes,

$$\widehat{\theta}_{Bayes} = \frac{p+n}{a+S} = \frac{p+n}{a+\sum_{i=1}^n X_i^2}.$$

Problema 2

Si denominamos μ_1 y μ_2 a las medias de las dos poblaciones en estudio (punto de fusión de la aleación I y II), se trata de contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$ en el caso que nos ocupa de dos poblaciones normales independientes de varianzas desconocidas pero supuestamente iguales y muestras pequeñas (véase CB-página 170), en el que se rechaza H_0 cuando y sólo cuando sea

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} > t_{n_1+n_2-2;\alpha/2}$$

Como el nivel de significación es $\alpha = 0'05$ se obtiene de la tabla de la distribución t de Student que $t_{n_1+n_2-2;\alpha/2} = t_{38;0'025}$ es un valor que estará comprendido entre $2'021$ y $2'042$. Como es

$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{|421-426|}{\sqrt{\dfrac{19\cdot 4^2 + 19\cdot 3^2}{20+20-2}}\sqrt{\dfrac{1}{20}+\dfrac{1}{20}}} = 4'47$$

se rechaza la hipótesis nula de igualdad de los puntos medios de fusión.

El p-valor es $2 \cdot P\{t_{38} > 4'47\}$, valor que, a partir de la tabla de la t de Student es claramente menor que $2 \cdot 0'0025 = 0'005$, suficientemente pequeño como para confirmar la decisión de rechazo antes tomada.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio. Segunda semana. Curso 2018-2019.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Se están realizando ensayos clínicos de un nuevo medicamento. En cada uno de los ensayos se eligen al azar m pacientes enfermos y se anota, al final del ensayo, el número X de pacientes sanos. Como en cada ensayo uno de los pacientes no está afectado por la enfermedad sino que es utilizado como control, el valor $X = 0$ nunca se obtiene en el ensayo, de manera que X toma los valores $1, 2, ..., m$. Si estos ensayos clínicos se repiten n veces de forma independiente, se pide:

(a) Modelizar la distribución de la variable X en observación utilizando una binomial $B(m, \theta)$ en la que no se puede presentar el valor $X = 0$ por razones experimentales, siendo m conocido y $\theta \in (0, 1)$ un parámetro desconocido. (Obsérvese que X no sigue una distribución binomial.)

$2'5$ puntos

(b) Para la muestra $X_1, ..., X_n$ de tamaño n de X, determinar el estimador centrado uniformemente de mínima varianza, si existe, de

$$g(\theta) = \frac{\theta}{1 - (1 - \theta)^m}$$

$2'5$ puntos

(c) Determinar la cota de Fréchet-Cramer-Rao para los estimadores insesgados de la función $g(\theta)$ anterior y la varianza del estimador calculado en el apartado anterior. ¿Es este estimador eficiente?

$2'5$ puntos

Problema 2

Se desea estimar la contaminación por mercurio en peces de lagos de Florida. La concentración de mercurio (en ppm) en el tejido muscular de 15 peces elegidos al azar fue la siguiente

$$1'22\,,\ 0'48\,,\ 0'49\,,\ 1'1\,,\ 0'59\,,\ 0'28\,,\ 0'19$$

0′11 , 0′27 , 0′64 , 0′75 , 0′41 , 0′19 , 0′81 , 0′56

Determinar un intervalo de confianza para la contaminación media por mercurio, para un coeficiente de confianza de 0'95, admitiendo la normalidad de los datos anteriores.

En base a estos datos, ¿podría concluirse que existe contaminación significativa por mercurio en dichos lagos?

2'5 puntos

Problema 1

a) La distribución binomial $B(m,\theta)$ tiene por función de masa (LCP, pp. 279)

$$p_\theta(x) = \binom{m}{x} \theta^x \, (1-\theta)^{m-x} \qquad\qquad x = 0, 1, ..., m.$$

Como, por razones experimentales, el valor $x = 0$ no se puede presentar, la distribución de X tendrá por función de masa o probabilidad

$$f_\theta(x) = P(X = x / X > 0) = \frac{P(X = x, X > 0)}{P(X > 0)} = \frac{P(X = x, X > 0)}{1 - P(X = 0)} = \frac{\binom{m}{x} \theta^x \, (1-\theta)^{m-x}}{1 - (1-\theta)^m}$$

$x = 1, 2, ..., m.$

b) Como esta distribución modelo es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$f_\theta(x) = \binom{m}{x} \frac{(1-\theta)^m}{1-(1-\theta)^m} \left(\frac{\theta}{1-\theta}\right)^x = \binom{m}{x} \frac{(1-\theta)^m}{1-(1-\theta)^m} \, e^{x \log(\theta/(1-\theta))}$$

será $S(X_1, ..., X_n) = \sum_{i=1}^n X_i$ un estadístico minimal suficiente. Además, (PIE, pp. 273), S será completo puesto que la imagen de $q(\theta) = \log(\theta/(1-\theta))$ contiene un abierto de $\mathbb{R}$, ya que al ser $0 < \theta < 1$, la imagen de la función q será el intervalo $(-\infty, \infty) = \mathbb{R}$. En definitiva, S será suficiente minimal y completo.

Basta ahora con encontrar un estimador, función del suficiente y completo S, que sea insesgado para

$$g(\theta) = \frac{\theta}{1-(1-\theta)^m}$$

y ese será el ECUMV para $g(\theta)$. Como la esperanza de S es $E[S(X_1, ..., X_n)] = E\left[\sum_{i=1}^n X_i\right] = n\, E[X]$, vamos a determinar la media de la variable en observación. Será

$$E[X] = \frac{(1-\theta)^m}{1-(1-\theta)^m} \sum_{x=1}^{m} x \binom{m}{x} \left(\frac{\theta}{1-\theta}\right)^x$$

Pero por otro lado, la media de una $B(l,p)$ es (véase, por ejemplo, LCP pp. 280)

$$l \cdot p = \sum_{x=0}^{l} x \binom{l}{x} p^x \, (1-p)^{l-x} = \sum_{x=1}^{l} x \binom{l}{x} p^x \, (1-p)^{l-x}$$

de donde se obtiene que es

$$\sum_{x=1}^{l} x \binom{l}{x} \left(\frac{p}{1-p}\right)^x = \frac{lp}{(1-p)^l}$$

con lo que será

$$E[X] = \frac{(1-\theta)^m}{1-(1-\theta)^m} \frac{m\,\theta}{(1-\theta)^m} = \frac{m\,\theta}{1-(1-\theta)^m}$$

Por tanto, el ECUMV para $g(\theta)$ será

$$T = S/(nm) = \frac{1}{n\,m} \sum_{i=1}^{n} X_i.$$

c) Observemos en primer lugar que aunque el estimador T determinado en el apartado anterior es el ECUMV, no necesariamente alcanza la cota de Fréchet-Cramer-Rao. Esta cota puede ser algo inferior a la varianza de dicho estimador. Lo recrípoco sí que es cierto; si un estimador alcanza la cota (es decir, es eficiente), entonces es el ECUMV.

Para los estimadores insesgados de $g(\theta)$ la mencionada cota es (LCP, pp. 225 y 231)

$$\frac{(g'(\theta))^2}{n\,i(\theta)}.$$

Por un lado es

$$g'(\theta) = \frac{1-(1-\theta)^{m-1}(1-\theta+m\,\theta)}{(1-(1-\theta)^m)^2}$$

y la cantidad de información del modelo

$$\begin{aligned} i(\theta) &= E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2\right] \\ &= E\left[\left(\frac{X}{\theta(1-\theta)} - \frac{m\theta}{\theta(1-\theta)[1-(1-\theta)^m]}\right)^2\right] \\ &= \frac{V(X)}{\theta^2(1-\theta)^2}. \end{aligned}$$

Para calcular la $V(X)$ sólo necesito determinar la media de los cuadrados

$$E\left[X^2\right] = \frac{(1-\theta)^m}{1-(1-\theta)^m} \sum_{x=1}^{m} x^2 \binom{m}{x} \left(\frac{\theta}{1-\theta}\right)^x$$

Utilizando, como hicimos antes, el momento de segundo orden de una $B(l,p)$ se obtiene que es (véase, por ejemplo, LCP pp. 281)

$$\sum_{x=1}^{l} x^2 \binom{l}{x} \left(\frac{p}{1-p}\right)^x = \frac{l\,p(1-p) + l^2p^2}{(1-p)^l}$$

con lo que será

$$E\left[X^2\right] = \frac{m\theta(1-\theta+\theta m)}{1-(1-\theta)^m}$$

y

$$V(X) = \frac{m\theta(1-\theta)\left[1-(1-\theta)^{m-1}(1-\theta+m\theta)\right]}{\left[1-(1-\theta)^m\right]^2}$$

obteniéndose una cota de Fréchet-Cramer-Rao igual a

$$\frac{(g'(\theta))^2}{n\,i(\theta)} = \frac{(g'(\theta))^2\;\theta^2(1-\theta)^2}{n\,V(X)} = \frac{\left[1-(1-\theta)^{m-1}(1-\theta+m\theta)\right]\theta(1-\theta)}{\left[1-(1-\theta)^m\right]^2 n\,m}.$$

Por otro lado, la varianza de T será igual a

$$V(T) = \frac{V(X)}{n\,m^2} = \frac{\left[1-(1-\theta)^{m-1}(1-\theta+m\theta)\right]\theta(1-\theta)}{\left[1-(1-\theta)^m\right]^2 n\,m}$$

con lo que se alcanzará la cota y T será eficiente.

Problema 2

El intervalo de confianza de coeficiente de confianza $0'95$ para la media de una población normal de varianza desconocida es el dado en, por ejemplo, EBR-sección 6.2 y se puede obtener fácilmente. El intervalo buscado es, $[0'358\,,\,0'72]$.

En la segunda parte del problema nos piden contrastar la hipótesis nula $H_0 : \mu = 0$ (o $H_0 : \mu \leq 0$) frente a la alternativa $H_1 : \mu > 0$. Como el valor del estadístico de contraste es $T_n = 6'383$, el p-valor de este test será a partir de las Tablas de la t de Student,

$$P\{T_{14} > 6'383\} < 0'0025$$

tan bajo que indica rechazar la hipótesis nula y concluir, por tanto, que existe contaminación media significativa por mercurio.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Septiembre. Curso 2018-2019.

NOTAS IMPORTANTES:

1) Duración del examen: Dos horas.

2) Material permitido: Solamente una calculadora no programable y el original (no se permiten fotocopias, ni anotaciones, ni hojas sueltas dentro) de la Addenda "Fórmulas y tablas estadísticas".

3) No es necesario entregar esta hoja de enunciados.

Problema 1

Sea X una variable aleatoria discreta con función de probabilidad

$$f_\theta(x) = \frac{(\log\theta)^{x-1}}{\theta\,\Gamma(x)} \qquad\qquad x = 1, 2, 3, ...$$

siendo $\theta > 1$ un parámetro desconocido y Γ la función gamma. Utilizando una muestra aleatoria simple de tamaño n de X, determinar:

(a) El estimador de máxima verosimilitud para θ y el obtenido por el método de los momentos.

2 puntos

(b) Un estadístico suficiente minimal para la familia de densidades dada y analizar su completitud.

2 puntos

(c) El estimador centrado uniformemente de mínima varianza, si existe, para $g(\theta) = \log\theta$.

2 puntos

(d) El test de razón de verosimilitudes de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. El test obtenido, ¿es uniformemente de máxima potencia?

2 puntos

Problema 2

Se quiere estimar la diferencia de uso que se hace de dos fotocopiadoras de una determinada empresa, mediante un intervalo de confianza para la diferencia de tiempos medios de utilización de ambas máquinas. Para ello, se controló el tiempo de utilización de cada fotocopiadora durante una serie de días elegidos al azar, obteniéndose los siguientes resultados:

Fotocopiadora 1	**2′4**	**3′1**	**2′5**	**2′7**	**3′1**	**2′7**	**3**	**2′3**	**3′2**	**3**
Fotocopiadora 2	**2′8**	**2**	**2′4**	**2**	**1′9**	**2′7**				

Suponiendo que el tiempo de utilización de cada fotocopiadora sigue una distribución normal, determinar un intervalo de confianza al 95 % para la diferencia de utilización media de las fotocopiadoras suponiendo,
(a) Que las varianzas de utilización de cada fotocopiadora son $\sigma_1^2 = 0'1$ y $\sigma_2^2 = 0'15$.
(b) Que las varianzas de utilización de ambas máquinas son desconocidas pero que pueden suponerse iguales.

2 puntos

Problema 1

a) La función de verosimilitud de la muestra (PIE, pp. 283)

$$L(\theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{(\log\theta)^{\sum_{i=1}^n x_i - n}}{\theta^n \prod_{i=1}^n \Gamma(x_i)}$$

tiene por logaritmo

$$\log L(\theta) = \left(\sum_{i=1}^{n} x_i - n\right) \log\log\theta - n\,\log\theta - \log\prod_{i=1}^{n}\Gamma(x_i)$$

cuya derivada igualada a cero es

$$\frac{\partial}{\partial\theta}\log L(\theta) = \left(\sum_{i=1}^{n} x_i - n\right)\frac{1}{\log\theta}\,\frac{1}{\theta} - \frac{n}{\theta} = 0$$

de donde, despejando, se obtiene como estimador máximo-verosímil para θ,

$$\widehat{\theta} = e^{\overline{x}-1}.$$

Para determinar el estimador de θ por el método de los momentos debemos calcular primero la media de la distribución modelo

$$\begin{aligned} E[X] &= \sum_{x=1}^{\infty} x\,\frac{(\log\theta)^{x-1}}{\theta\,\Gamma(x)} = \sum_{x=1}^{\infty} x\,\frac{(\log\theta)^{x-1}}{\theta\,(x-1)!} \\ &= \sum_{y=0}^{\infty}(y+1)\,\frac{(\log\theta)^{y}}{\theta\,y!} \\ &= \frac{1}{\theta}\sum_{y=0}^{\infty} y\,\frac{(\log\theta)^{y}}{y!} + \frac{1}{\theta}\sum_{y=0}^{\infty}\frac{(\log\theta)^{y}}{y!} \end{aligned}$$

Como la media de una distribución de Poisson $\mathcal{P}(\lambda)$ es λ, será

$$\sum_{y=0}^{\infty} y\,\frac{\lambda^y}{y!} = \lambda\,e^{\lambda}$$

Por otro lado, de su función de probabilidad obtenemos que es

$$\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{\lambda}$$

para todo parámetro $\lambda > 0$. Por tanto,

$$E[X] = \frac{1}{\theta}\,(\log\theta)\,\theta + \frac{1}{\theta}\,\theta = \log\theta + 1.$$

(Sería más rápido observando que $Y = X - 1$ sigue una distribución $\mathcal{P}(\log\theta)$).

La ecuación para obtener el estimador por el método de los momentos será

$$E[X] = \overline{x}$$

de donde se obtiene como estimador para θ,

$$\widehat{\theta} = e^{\overline{x}-1}$$

es decir, el mismo que el obtenido con el método de la máxima verosimilitud.

b) La distribución modelo es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$f_\theta(x) = \frac{1}{\theta}\,\frac{1}{\Gamma(x)}\,e^{(x-1)(\log\log\theta)}$$

con lo que el estadístico $S(X_1, ..., X_n) = \sum_{i=1}^n X_i - n$ será un estadístico minimal suficiente. Por tanto, (PIE, pp. 273) S será además completo si la imagen de $q(\theta) = \log\log\theta$ contiene un abierto de $\mathbb{R}$; como es $\theta > 1$, la imagen de la función q será $(-\infty, \infty) = \mathbb{R}$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$. En definitiva, S será suficiente minimal y completo.

c) Basta con encontrar un estimador, función del suficiente y completo S, que sea insesgado para la correspondiente función del parámetro y ése será el ECUMV buscado.

Comencemos calculando la esperanza de S.

$$E[S(X_1, ..., X_n)] = E\left[\sum_{i=1}^n X_i - n\right] = n\,E[X] - n = n\,\log\theta + n - n = n\,\log\theta.$$

Por tanto, el ECUMV para $g_1(\theta) = \log\theta$ será

$$\frac{S}{n} = \overline{x} - 1.$$

d) El test de razón de verosimilitudes está basado en el estadístico de contraste (PIE, pp. 388)

$$\Lambda(x_1, ..., x_n) = \frac{\sup_{\theta\in\Theta_0} f_\theta(x_1, ..., x_n)}{\sup_{\theta\in\Theta_0\cup\Theta_1} f_\theta(x_1, ..., x_n)}$$

Ya vimos en el primer apartado que el máximo en $\Theta_0 \cup \Theta_1$ (que es todo el espacio paramétrico $\theta > 1$), de la función de verosimilitud $L(\theta)$ se obtiene en el estimador de máxima verosimilitud $\widehat{\theta} = e^T$,

$$L(\widehat{\theta}) = \frac{T^{nT}}{e^{nT} \prod_i \Gamma(x_i)}$$

siendo

$$T = T(X_1, ..., X_n) = \overline{x} - 1.$$

El máximo bajo la hipótesis nula $\theta \leq \theta_0$, dependerá de la relación entre θ_0 y T. Así,

$$\sup_{\theta \leq \theta_0} L(\theta) = \begin{cases} \dfrac{(\log \theta_0)^{nT}}{\theta_0^n \prod_i \Gamma(x_i)} & \text{si} \quad e^T > \theta_0 \\ \dfrac{T^{nT}}{e^{nT} \prod_i \Gamma(x_i)} & \text{si} \quad e^T \leq \theta_0 \end{cases}$$

Por lo tanto, el estadístico del test de razón de verosimilitudes será

$$\Lambda(x_1, ..., x_n) = \frac{\sup_{\theta \leq \theta_0} f_\theta(x_1, ..., x_n)}{\sup_{\theta > 1} f_\theta(x_1, ..., x_n)} = \begin{cases} \dfrac{(\log \theta_0)^{nT}\, e^{nT}}{\theta_0^n\, T^{nT}} & \text{si} \quad T > \log \theta_0 \\ 1 & \text{si} \quad T \leq \log \theta_0 \end{cases}$$

Fácilmente se puede demostrar (tomando logaritmos y derivando) que esta función del estadístico T es decreciente en T (de hecho es constantemente igual a 1 hasta $T = \log \theta_0$ y luego decreciente en T), por lo que la región crítica del test de razón de verosimilitudes definida por $\Lambda < c$ se transforma en $T > k$, quedando en suma el test de razón de verosimilitudes de la forma

$$\varphi = \begin{cases} 1 & \text{si} \quad T > k \\ \gamma & \text{si} \quad T = k \\ 0 & \text{si} \quad T < k \end{cases}$$

siendo k y γ dos constantes que se determinan por la condición del tamaño α del test

$$\alpha = P_{\theta_0}\{T > k\} + \gamma\, P_{\theta_0}\{T = k\}$$

a partir de la distribución de T, que es una transformación de una distribución de Poisson.

Como la familia de distribuciones de la muestra tiene razón de verosimilitud monótona en T, el test de razón de verosimilitudes es, en este caso, uniformemente de máxima potencia.

Problema 2

Si llamamos X_i, $i = 1, 2$ a la variable aleatoria *tiempo de utilización de la fotocopiadora i-ésima*, el enunciado dice que se supone $X_i \rightsquigarrow N(\mu_i, \sigma_i)$ $i = 1, 2$, siendo el objetivo determinar un intervalo de confianza para la diferencia de medias de las dos poblaciones normales independientes (CB-sección 6.6).

a) En este caso, al suponerse conocidas las varianzas poblacionales, el intervalo a utilizar será

$$\begin{aligned} I &= \left[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \, , \, \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] \\ &= \left[2'8 - 2'3 - 1'96 \sqrt{\frac{0'1}{10} + \frac{0'15}{6}} \, , \, 2'8 - 2'3 + 1'96 \sqrt{\frac{0'1}{10} + \frac{0'15}{6}} \right] \\ &= [0'133 \, , \, 0'867] \end{aligned}$$

al obtenerse, a partir de los datos del enunciado, que es $\overline{x}_1 = 2'8$, $\overline{x}_2 = 2'3$ y $z_{\alpha/2} = z_{0'05/2} = z_{0'025} = 1'96$.

b) Ahora las varianzas poblacionales son desconocidas pero pueden suponerse iguales. El intervalo a utilizar será el de extremos

$$\overline{x}_1 - \overline{x}_2 \mp t_{n_1+n_2-2;\alpha/2} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

es decir, el de extremos

$$2'8 - 2'3 \mp 2'145 \sqrt{\frac{9 \cdot 0'104 + 5 \cdot 0'152}{14}} \sqrt{\frac{1}{10} + \frac{1}{6}}$$

al obtenerse de los datos que las cuasivarianzas muestrales son $S_1^2 = 0'104$ y $S_2^2 = 0'152$.

Haciendo operaciones, se obtiene finalmente el intervalo $[0'114 \, , \, 0'885]$ que como puede apreciarse es bastante similar el obtenido en el apartado anterior.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

EBR: **Estadística Básica con R** (2010). Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio (2019-2020).
Sábado 30 de Mayo. Sesión de Tarde. De 16 a 18.
Letras A - B

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{1}{2\,\theta}\ \exp\left\{\frac{x-2}{\theta} - \exp\left(\frac{x-2}{2\,\theta}\right)\right\} \qquad -\infty < x < \infty$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, determinar (utilizando la función Gamma y sus derivadas):
(a) El estimador de θ por el método de los momentos. ¿Es insesgado para θ?

2 puntos

(b) Determinar la cota de Fréchet-Cramer-Rao para los estimadores insesgados de θ y analizar si el estimador obtenido en el apartado anterior es o no eficiente para θ, sabiendo que es $\Gamma'(2) = 0'42$, $\Gamma'(3) = 1'85$, $\Gamma''(2) = 0'82$ y $\Gamma''(3) = 2'43$.

2 puntos

(c) Determinar un intervalo de confianza para θ, de coeficiente de confianza $1-\alpha$, basado en la distribución asintótica del estimador obtenido en el apartado (a).

2 puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con distribución normal de media cero y varianza $V(X) = 1/\theta$, siendo $\theta > 0$ un parámetro desconocido. Suponiendo que θ es una variable aleatoria con distribución gamma $\gamma(p, a)$, con $p, a > 0$, y utilizando una muestra aleatoria simple de tamaño n de X, determinar el estimador Bayes de θ bajo una función de pérdida cuadrática.

2 puntos

Problema 3

Los siguientes datos son los precios de un determinado producto en dos Zonas de un país.

Zona I	5'21	5'16	6'10	6'55	5'80	4'60	4'19	4'75
Zona II	4'59	6'75	4'73	6'90	4'25	5'50	6'29	4'85

¿Se puede concluir que existen diferencias significativas entre los precios medios de ambos Grupos, admitiendo que los precios siguen distribuciones normales independientes?

2 puntos

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio (2019-2020).
Sábado 30 de Mayo. Sesión de Tarde. De 16 a 18.
Letras C - D

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{1}{2\,\theta}\ \exp\left\{\frac{x-2}{\theta} - \exp\left(\frac{x-2}{2\,\theta}\right)\right\} \qquad -\infty < x < \infty$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, determinar (utilizando la función Gamma y sus derivadas):
(a) El estimador de θ por el método de los momentos. ¿Es insesgado para θ?

2 puntos

(b) Determinar la cota de Fréchet-Cramer-Rao para los estimadores insesgados de θ y analizar si el estimador obtenido en el apartado anterior es o no eficiente para θ, sabiendo que es $\Gamma'(2) = 0'42$, $\Gamma'(3) = 1'85$, $\Gamma''(2) = 0'82$ y $\Gamma''(3) = 2'43$.

2 puntos

(c) Determinar un intervalo de confianza para θ, de coeficiente de confianza $1-\alpha$, basado en la distribución asintótica del estimador obtenido en el apartado (a).

2 puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con distribución normal de media cero y varianza $V(X) = 1/\theta$, siendo $\theta > 0$ un parámetro desconocido. Suponiendo que θ es una variable aleatoria con distribución gamma $\gamma(p, a)$, con $p, a > 0$, y utilizando una muestra aleatoria simple de tamaño n de X, determinar el estimador Bayes de θ bajo una función de pérdida cuadrática.

2 puntos

Problema 3

Consideremos dos grupos de datos en los que se observó una variable que tomaba los datos siguientes:

GRUPO I: 0'627 , 0'680 , 0'620 , 0'622 , 0'652

GRUPO II: 0'575 , 0'611 , 0'590 , 0'570 , 0'617

Suponiendo que dichos valores siguen una distribución normal, ¿es significativamente mayor la media del GRUPO I que las del GRUPO II?

2 puntos

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio (2019-2020).
Sábado 30 de Mayo. Sesión de Tarde. De 16 a 18.
Letras E - F

Problema 1

Dada una muestra aleatoria simple de tamaño n de una distribución con función de densidad

$$f_\theta(x) = \frac{1}{2\sqrt{\theta}}\, x^{-3/2} \qquad\qquad x \geq \frac{1}{\theta}$$

siendo θ un parámetro positivo desconocido, se pide:
a) Determinar el estimador de máxima verosimilitud de θ y estudiar si es suficiente.

$2'5$ puntos

b) Determinar, por el método de la cantidad pivotal, el intervalo de confianza para θ de nivel de confianza $1-\alpha$ que tenga longitud mínima.

2 puntos

c) Determinar un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

$3'5$ puntos

Problema 2

Los datos de la siguiente tabla de contingencia son frecuencias absolutas de dos variables:

	Modalidad 1	Modalidad 2	Modalidad 3	Modalidad 4	Modalidad 5
Clase 1	**96**	**32**	**10**	**4**	**5**
Clase 2	**95**	**23**	**13**	**0**	**4**
Clase 3	**32**	**12**	**5**	**1**	**5**
Clase 4	**22**	**6**	**5**	**0**	**0**

Determinar si puede aceptarse la independencia entre estas dos variables que forman la tabla de contingencia.

2 puntos

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio (2019-2020).
Sábado 30 de Mayo. Sesión de Mañana. De 10 a 12.
Letras M - O

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \theta\,(\theta+1)\,e^{-2\,x}\left(1-e^{-x}\right)^{\theta-1} \qquad\qquad x>0$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
(a) Determinar un estadístico suficiente minimal y analizar su completitud.

2 puntos

(b) Determinar el estimador centrado uniformemente de mínima varianza de $g(\theta) = [1 + 2\,\theta]/[\theta\,(1+\theta)]$.

2 puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{3\,x}{2\,\theta^2} - \frac{3\,x^2}{4\,\theta^3} \qquad\qquad x \in (0\,,\,2\,\theta)$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, determinar:
(a) El estimador de θ mediante el método de los momentos y analizar su consistencia.

2 puntos

(b) Supuesto que el tamaño muestral es suficientemente grande como para poder utilizar las distribuciones asintóticas, determinar el intervalo de confianza de colas iguales para θ, de coeficiente de confianza $1-\alpha$, primero utilizando el estimador del apartado anterior y luego la mediana muestral M, la cual tiene por distribución asintótica una normal $N(\,x_{1/2}, 1/(2f_\theta(x_{1/2})\sqrt{n})\,)$, siendo $x_{1/2}$ la mediana poblacional. ¿Cuál tiene menor longitud esperada?

2 puntos

Problema 3

Los siguientes datos son una muestra aleatoria simple de una distribución normal $N(\mu, \sigma)$.

$$1'8\,,\ 0'9\,,\ 1'5\,,\ 1'9\,,\ 1'2\,,\ 0'6\,,\ 1'8\,,\ 0'7\,,\ 2'9\,,\ 2'1$$

(a) Calcular un intervalo de confianza para μ de coeficiente de confianza 95 %.
(b) ¿Se podría concluir con que es μ mayor que $1'35$?

2 puntos

INFERENCIA ESTADÍSTICA

Prueba Presencial de Mayo/Junio (2019-2020).
Sábado 30 de Mayo. Sesión de Mañana. De 10 a 12.
Letras S - T

Problema 1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta^3}{2}\,(x-2)^2\,e^{-\theta\,x}\,e^{2\,\theta} \qquad\qquad x > 2$$

siendo $\theta > 0$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
(a) Estimar θ por el método de la máxima verosimilitud y por el método de los momentos.

2′5 puntos

(b) Determinar el estimador centrado uniformemente de mínima varianza para $g_1(\theta) = 1/\theta$ y para $g_2(\theta) = \theta$.

2′5 puntos

Problema 2

Sea X una variable aleatoria absolutamente continua con función de densidad $f_0(x) = 1\,,\ 0 < x < 1$ bajo la hipótesis nula H_0 y con función de densidad $f_1(x) = 2\,x\,,\ 0 < x < 1\,,$ bajo la hipótesis alternativa H_1. Determinar un contraste de máxima potencia de nivel α para contrastar H_0 frente a H_1, utilizando una muestra aleatoria simple de tamaño n de X. Calcular la potencia del contraste obtenido si es $\alpha = 0'05$ y $n = 3$.

3 puntos

Problema 3

Los datos que siguen corresponden al número de icebergs observados en el mar desde dos localizaciones,

	Mes											
	E	**F**	**M**	**A**	**M**	**Jn**	**Jl**	**A**	**S**	**O**	**N**	**D**
Localización 1	**4**	**9**	**31**	**83**	**140**	**78**	**25**	**13**	**9**	**4**	**3**	**2**
Localización 2	**10**	**11**	**4**	**9**	**20**	**15**	**3**	**2**	**1**	**0**	**0**	**0**

En base a estos datos, ¿existen o no diferencias significativas entre los avistamientos de icebergs desde uno y otro lugar?

2 puntos

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2012-2013

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Los Tutores deberán de haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Sobre 10 puntos, el primer problema valdría 8 puntos y el segundo 2 puntos.

Ejercicios para la Evaluación Continua

Problema 1.1

Se dispone de una muestra aleatoria simple de tamaño n de una variable aleatoria con función de densidad

$$f_\theta(x) = \frac{1}{2}\,\theta^3\,x^2\,e^{-\theta x} \qquad\qquad x > 0$$

siendo θ un parámetro positivo desconocido. Se pide:

a) Determinar un estimador de máxima verosimilitud para $1/\theta$ y estudiar si es suficiente.

b) Sabiendo que $T = \sum_{i=1}^n X_i$ es un estadístico completo, determinar el estimador centrado uniformemente de mínima varianza de $1/\theta$.

c) Determinar un contraste de máxima potencia de nivel α para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, siendo $\theta_1 > \theta_0$.

d) Razonar por qué el contraste obtenido en el apartado anterior es uniformemente de máxima potencia para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta > \theta_0$. Determinar su función de potencia.

e) Determinar el test de razón de verosimilitudes para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. ¿Es uniformemente de máxima potencia?

---

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1, ..., x_n) = \frac{1}{2^n}\,\theta^{3n}\,\prod_{i=1}^n x_i^2\,\exp\left\{-\theta\sum_{i=1}^n x_i\right\} \qquad \text{si} \quad x_1, ..., x_n > 0$$

con lo cual será

$$\log f_\theta(x_1, ..., x_n) = -n\log 2 + 3n\log\theta + \log\prod_{i=1}^n x_i^2 - \theta\sum_{i=1}^n x_i$$

obteniéndose, de la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta}\log f_\theta(x_1, ..., x_n) = \frac{3n}{\theta} - \sum_{i=1}^n x_i = 0$$

el estimador de máxima verosimilitud para θ

$$\hat\theta = \frac{3n}{\sum_{i=1}^n X_i}.$$

Por tanto, (véase *PIE*, pag. 290) el estimador de máxima verosimilitud de $1/\theta$ será

$$\frac{\sum_{i=1}^{n} X_i}{3n} = \frac{T}{3n}.$$

Como la función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1, ..., x_n) = h(x_1, ..., x_n) \cdot g_\theta(S(x_1, ..., x_n))$$

con

$$h(x_1, ..., x_n) = \frac{1}{2^n} \prod_{i=1}^{n} x_i^2$$

y

$$g_\theta(S(x_1, ..., x_n)) = \theta^{3n} \exp\left\{-3n\theta \frac{\sum_{i=1}^{n} x_i}{3n}\right\}$$

es decir, mediante el producto de una función que no depende del parámetro, h, y una función que depende del parámetro y de la muestra sólo a través del estadístico, $\sum X_i/3n$, por el teorema de factorización (*PIE*, pag. 167), éste será suficiente.

b) Como $T = \sum_{i=1}^{n} X_i$ es un estadístico suficiente y por hipótesis completo, por el teorema de Lehmann-Scheffé (*PIE*, pag. 218), sólo tenemos que buscar un estimador insesgado para el parámetro, función del suficiente y completo T.

Como $X \rightsquigarrow \gamma(3, \theta)$, será $T = \sum_{i=1}^{n} X_i \rightsquigarrow \gamma(3n, \theta)$, por lo que será $E[T] = 3n/\theta$ y, por tanto

$$E\left[\frac{T}{3n}\right] = \frac{1}{\theta}.$$

El ECUMV para $1/\theta$ será, por tanto, $T/3n$; es decir, el estimador de máxima verosimilitud.

c) El lema de Neyman-Pearson nos dice (*PIE*, pag. 338) que el test de máxima potencia tiene como región crítica la dada por

$$f_{\theta_1}(x_1, ..., x_n) > k f_{\theta_0}(x_1, ..., x_n)$$

es decir,

$$\frac{1}{2^n}\,\theta_1^{3n} \prod_{i=1}^{n} x_i^2 \exp\left\{-\theta_1 \sum_{i=1}^{n} x_i\right\} > k \frac{1}{2^n}\,\theta_0^{3n} \prod_{i=1}^{n} x_i^2 \exp\left\{-\theta_0 \sum_{i=1}^{n} x_i\right\}$$

de donde, haciendo operaciones y teniendo en cuenta que es $\theta_1 > \theta_0$, la región crítica será la dada por

$$\sum_{i=1}^{n} x_i < c$$

siendo c la constante para la que el test tiene tamaño α; es decir, tal que

$$\mathcal{P}_{\theta=\theta_0}\left\{\sum_{i=1}^{n} X_i < c\right\} = \alpha.$$

Sabemos que $\sum_{i=1}^{n} X_i \rightsquigarrow \gamma(3n, \theta)$. No obstante, para determinar en un caso concreto el punto crítico del test deberemos calcular abscisas de dicha distribución, la cual no está tabulada; por esta razón, es conveniente modificar ligeramente la distribución anterior. Como es $T = \sum X_i \rightsquigarrow \gamma(3n, \theta)$, su función característica será

$$\varphi_T(t) = \left(1 - \frac{it}{\theta}\right)^{-3n}.$$

La de $2\theta T$ será, por tanto,

$$E\left[e^{it2\theta T}\right] = \varphi_T(2\theta t) = \left(1 - \frac{it}{1/2}\right)^{-3n}$$

es decir,

$$2\,\theta \sum_{i=1}^{n} X_i \rightsquigarrow \gamma(6n/2, 1/2) = \chi^2_{6n}.$$

En general, tenemos un resultado, también fácilmente demostrable a través de las funciones características, que dice:

Si $Y \rightsquigarrow \gamma(p, a)$, entonces $2aY \rightsquigarrow \chi^2_{2p}$

Por tanto, será

$$\alpha = \mathcal{P}_{\theta=\theta_0}\left\{\sum_{i=1}^{n} X_i < c\right\} = \mathcal{P}_{\theta=\theta_0}\left\{2\,\theta_0 \sum_{i=1}^{n} X_i < 2\,\theta_0\ c\right\}$$

con lo que deberá ser $2\,\theta_0\, c = \chi^2_{6n;1-\alpha}$, resultando el test de máxima potencia igual a

$$\phi'(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad \sum x_i < \chi^2_{6n;1-\alpha}/(2\theta_0) \\ \\ 0 & \text{si} \quad \sum x_i \geq \chi^2_{6n;1-\alpha}/(2\theta_0) \end{cases}$$

d) Como el test ϕ' obtenido en el apartado anterior no depende del valor de la alternativa θ_1, sino simplemente de que sea mayor que θ_0 (es necesario el signo de su diferencia con θ_0 en la determinación de la región crítica), será (*PIE*, pag. 348) uniformemente de máxima potencia para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta > \theta_0$.

La función de potencia del test será

$$\beta_{\phi'}(\theta) = \mathcal{P}_\theta \left\{ \sum_{i=1}^{n} X_i < \frac{\chi^2_{6n;1-\alpha}}{2\theta_0} \right\} = P \left\{ \chi^2_{6n} < \theta \, \frac{\chi^2_{6n;1-\alpha}}{\theta_0} \right\}.$$

e) El test de razón de verosimilitudes es por definición, (*PIE, pag. 388*), el que tiene por región crítica la dada por

$$\Lambda(x_1, ..., x_n) < b$$

siendo

$$\Lambda(x_1, ..., x_n) = \frac{\sup_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)}{\sup_{\theta \in \Theta} f_\theta(x_1, ..., x_n)}.$$

Como vimos en el primer apartado, el estimador de máxima verosimilitud de θ es

$$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} X_i} = \frac{3n}{T}.$$

Por tanto,

$$\sup_{\theta \in \Theta} f_\theta(x_1, ..., x_n) = f_{\hat{\theta}}(x_1, ..., x_n) = \frac{1}{2^n} \frac{(3n)^{3n}}{T^{3n}} \prod_{i=1}^{n} x_i^2 \, e^{-3n}.$$

Por otra parte, como $f_\theta(x_1, ..., x_n)$ (o su logaritmo) es una función creciente para $\theta \in (0, \hat{\theta})$ y decreciente para $\theta \in (\hat{\theta}, \infty)$, el máximo para $\theta \leq \theta_0$ se alcanzará en

$$\hat{\theta}_0 = \begin{cases} \theta_0 & \text{si} \quad \hat{\theta} > \theta_0 \\ \\ \hat{\theta} & \text{si} \quad \hat{\theta} \leq \theta_0 \end{cases}$$

y vale

$$\sup_{\theta \leq \theta_0} f_\theta(x_1, ..., x_n) = \begin{cases} \frac{1}{2^n} \theta_0^{3n} \prod_{i=1}^{n} x_i^2 \, e^{-\theta_0 T} & \text{si} \quad \hat{\theta} > \theta_0 \\ \\ \frac{1}{2^n} \frac{(3n)^{3n}}{T^{3n}} \prod_{i=1}^{n} x_i^2 \, e^{-3n} & \text{si} \quad \hat{\theta} \leq \theta_0 \end{cases}$$

Por tanto será,

$$\Lambda(x_1,...,x_n) = \begin{cases} \dfrac{\theta_0^{3n}T^{3n}}{(3n)^{3n}}\, e^{-\theta_0 T+3n} & \text{si} \quad T < 3n/\theta_0 \\ \\ 1 & \text{si} \quad T \geq 3n/\theta_0 \end{cases}$$

Tomando logaritmos se comprueba que Λ es una función creciente de T; por tanto, $\Lambda(x_1,...,x_n) < b$ equivale a $T < c$. Es decir, el test ϕ' obtenido en el apartado c).

En el apartado d) vimos que el test ϕ' maximiza la potencia uniformemente en la hipótesis alternativa $\Theta_1 = \{\theta : \theta > \theta_0\}$, entre todos los tests ϕ que cumplen la condición $E_{\theta_0}[\phi(x_1,...,x_n)] \leq \alpha$.

Ahora buscamos el óptimo en un subconjunto de los tests antes considerados, el de los tests ϕ que cumplen la condición

$$E_\theta[\phi(x_1,...,x_n)] \leq \alpha \qquad\qquad \forall\, \theta \leq \theta_0$$

condición más restrictiva que la anterior, impuesta por la hipótesis nula $H_0 : \theta \leq \theta_0$.

Como el test ϕ' antes obtenido tiene función de potencia no decreciente en θ, valiendo $\beta_\phi(\theta_0) = \alpha$, será

$$E_\theta[\phi'(x_1,...,x_n)] \leq \alpha \qquad\qquad \forall\, \theta \leq \theta_0$$

por lo que ϕ' cumple la nueva restricción, es decir, pertenece a la clase de tests en donde se busca ahora el óptimo. En consecuencia, ϕ' será también uniformemente de máxima potencia para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. (*PIE*, pag. 348).

Una forma alternativa de razonar es observar que la distribución de la muestra

$$f_\theta(x_1,...,x_n) = \frac{1}{2^n}\,\theta^{3n} \prod_{i=1}^{n} x_i^2 \, \exp\left\{-\theta \sum_{i=1}^{n} x_i\right\} \qquad\qquad x_1,...,x_n > 0$$

es de tipo exponencial

$$f_\theta(x_1,...,x_n) = c(\theta)h(x_1,...,x_n)e^{q(\theta)T(x_1,...,x_n)}$$

con $q(\theta) = \theta$ y $T(x_1,...,x_n) = -\sum x_i$, teniendo en consecuencia razón de verosimilitud monótona en $T(x_1,...,x_n)$, por lo que, según el teorema de Karlin-Rubin, (*PIE*, pag. 349), el test uniformemente de máxima potencia buscado es aquel que tiene por región crítica $T(x_1,...,x_n) > k$, es decir, $\sum x_i < c$, que es el test ϕ' antes determinado.

Problema 1.2

Una muestra aleatoria de 10 clientes de una farmacia determinada mostró los siguientes tiempos de espera hasta que son atendidos, en minutos:

2 , 10 , 4 , 5 , 1 , 0 , 5 , 9 , 3 , 9

Determinar un intervalo de confianza, con coeficiente de confianza 0′9, para el tiempo medio de espera, admitiendo que el tiempo de espera en esa farmacia sigue una distribución normal.

Se trata de calcular el intervalo de confianza para la media de una población normal de varianza desconocida y tamaños muestrales pequeños, cuya expresión es (CB-sección 6.2)

$$\left[\overline{x} - t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\ ,\ \overline{x} + t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\right].$$

De los datos del enunciado se desprende que es $\overline{x} = 4'8$ y $S = 3'52$. Por tanto, como de la Tabla 5 de la t de Student se obtiene que es $t_{n-1;\alpha/2} = t_{9;0'05} = 1'833$, el intervalo de confianza solicitado será

$$\left[\overline{x} - t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\ ,\ \overline{x} + t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}\right] = \left[\,4'8 - 1'833\,\frac{3'52}{\sqrt{10}}\ ,\ 4'8 + 1'833\,\frac{3'52}{\sqrt{10}}\,\right] =$$

$$= [\,2'76\,,\,6'84\,].$$

Referencias

(**PIE**). Vélez, R. y García Pérez, A. *Principios de Inferencia Estadística*. Editorial UNED.

(**CB**). García Pérez, A. (2008). *Estadística Aplicada: Conceptos Básicos*. Editorial UNED. (Código 0184011EP01A02).

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2013-2014

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Los Tutores deberán de haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Sobre 10 puntos, el primer problema de la Evaluación Continua vale 6 puntos y el segundo 4 puntos.

Esta Evaluación Continua es un buen ejemplo de lo que será la Prueba Presencial de este año aunque en dicho examen presencial no se trabajaría con un número elevado de datos.

Ejercicios para la Evaluación Continua

Problema 1.1

Se dispone de una muestra aleatoria simple de tamaño n de la densidad

$$f_\theta(x) = \theta\,(1-x)^{\theta-1} \qquad\qquad 0 < x < 1$$

siendo θ un parámetro positivo desconocido. Se pide:

a) Determinar el estimador de máxima verosimilitud de θ y un estadístico suficiente para la familia de densidades dada.

b) Sabiendo que $T = \sum_{i=1}^{n} \log(1 - X_i)$ es un estadístico completo, determinar el estimador centrado uniformemente de mínima varianza para $1/\theta$ y el estimador centrado uniformemente de mínima varianza para θ.

c) Determinar un contraste de máxima potencia de nivel α para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta = \theta_1$, siendo $\theta_1 > \theta_0$.

d) Razonar por qué el contraste obtenido en el apartado anterior es uniformemente de máxima potencia para contrastar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta > \theta_0$. Determinar su función de potencia.

e) Determinar el test de razón de verosimilitudes para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. ¿Es uniformemente de máxima potencia?

---

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1, ..., x_n) = \theta^n \prod_{i=1}^{n} (1 - x_i)^{\theta-1} \qquad \text{si} \quad x_1, ..., x_n \in (0, 1)$$

con lo cual será

$$\log f_\theta(x_1, ..., x_n) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log(1 - x_i)$$

obteniéndose de la ecuación de verosimilitud

$$\frac{\partial}{\partial \theta} \log f_\theta(x_1, ..., x_n) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(1 - x_i) = 0$$

el estimador de máxima verosimilitud para θ

$$\hat{\theta} = \frac{n}{-\sum_{i=1}^{n} \log(1 - X_i)} = \frac{n}{-T}.$$

Como la función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1, ..., x_n) = h(x_1, ..., x_n) \cdot g_\theta(S(x_1, ..., x_n))$$

con

$$h(x_1, ..., x_n) = 1$$

y

$$g_\theta(S(x_1, ..., x_n)) = \theta^n \exp\left\{(\theta - 1)\sum_{i=1}^{n} \log(1 - x_i)\right\}$$

es decir, mediante el producto de una función que no depende del parámetro, h, y una función que depende del parámetro y de la muestra sólo a través del estadístico $T = \sum \log(1 - X_i)$, por el teorema de factorización (véase *PIE*, pag. 167), T será suficiente para la familia de densidades dada.

b) Como $T = \sum_{i=1}^{n} \log(1 - X_i)$ es un estadístico suficiente y por hipótesis completo, por el teorema de Lehmann-Scheffé (*PIE*, pag. 218), sólo tenemos que buscar un estimador insesgado para el parámetro, el cual sea función del suficiente y completo T.

Suele comenzarse calculando la media de T. Para ello, como la función de densidad de $Y = -\log(1 - X)$ es

$$f_Y(y) = f_X(1 - e^{-y})\, e^{-y} = \theta\, e^{-\theta y} \qquad\qquad y > 0$$

ser $Y \rightsquigarrow \gamma(1, \theta)$ y como esta distribución es reproductiva respecto al primer parámetro, será

$$-\sum_{i=1}^{n} \log(1 - X_i) \rightsquigarrow \gamma\,(n, \theta)\,.$$

En consecuencia,

$$E[-T] = E\left[-\sum_{i=1}^{n} \log(1 - X_i)\right] = \frac{n}{\theta}$$

y, por tanto, el ECUMV para $1/\theta$ será $-T/n$.

Con objeto de determinar el ECUMV para θ y en vista de lo anterior, parece razonable empezar calculando $E[-1/T]$. Será

$$E\left[\frac{1}{-T}\right] = \int_0^\infty \frac{1}{x}\, \frac{\theta^n\, e^{-\theta x}\, x^{n-1}}{\Gamma(n)}\, dx = \frac{\theta}{n-1} \int_0^\infty \frac{\theta^{n-1}\, e^{-\theta x}\, x^{n-2}}{\Gamma(n-1)}\, dx = \frac{\theta}{n-1}.$$

El ECUMV para θ será, por tanto, $-(n-1)/T$.

c) El lema de Neyman-Pearson nos dice (*PIE*, pag. 338) que el test de máxima potencia tiene como región crítica la dada por

$$f_{\theta_1}(x_1, ..., x_n) > k f_{\theta_0}(x_1, ..., x_n)$$

es decir,

$$\theta_1^n \prod_{i=1}^{n} (1 - x_i)^{\theta_1 - 1} > k\, \theta_0^n \prod_{i=1}^{n} (1 - x_i)^{\theta_0 - 1}$$

de donde, haciendo operaciones y teniendo en cuenta que es $\theta_1 > \theta_0$, la región crítica será la dada por

$$\sum_{i=1}^{n} \log(1 - x_i) > c'$$

o mejor,

$$-\sum_{i=1}^{n} \log(1 - x_i) < c$$

siendo c la constante para la que el test tiene tamaño α; es decir, tal que

$$\mathcal{P}_{\theta=\theta_0} \left\{ -\sum_{i=1}^{n} \log(1 - X_i) < c \right\} = \alpha.$$

Sabemos que

$$-\sum_{i=1}^{n} \log(1 - X_i) \rightsquigarrow \gamma\,(n, \theta)$$

no obstante, para determinar en un caso concreto el punto crítico del test deberemos calcular abscisas de dicha distribución, la cual no está tabulada; por esta razón, es conveniente modificar ligeramente la distribución anterior utilizando el resultado que dice que

Si $Z = -\sum \log(1 - X_i) \rightsquigarrow \gamma(n, \theta)$, entonces su función característica es

$$\varphi_Z(t) = \left(1 - \frac{it}{\theta}\right)^{-n}.$$

La de $2\theta Z$ será, por tanto,

$$E\left[e^{it2\theta Z}\right] = \varphi_Z(2\theta t) = \left(1 - \frac{it}{1/2}\right)^{-n}$$

Por tanto, será

$$-2\,\theta\,\sum_{i=1}^{n}\log(1-X_i)\rightsquigarrow\gamma(2n/2,1/2)=\chi^2_{2n}.$$

Y en consecuencia, será

$$\alpha=\mathcal{P}_{\theta_0}\left\{-\sum_{i=1}^{n}\log(1-X_i)<c\right\}=\mathcal{P}_{\theta_0}\left\{-2\theta_0\sum_{i=1}^{n}\log(1-X_i)<2\theta_0 c\right\}$$

con lo que deberá ser $2\,\theta_0\,c=\chi^2_{2n;1-\alpha}$, resultando el test de máxima potencia igual a

$$\phi'(x_1,...,x_n)=\begin{cases}1 & \text{si} \quad -\sum\log(1-x_i)<\chi^2_{2n;1-\alpha}/(2\theta_0)\\ \\ 0 & \text{si} \quad -\sum\log(1-x_i)\geq\chi^2_{2n;1-\alpha}/(2\theta_0)\end{cases}$$

d) Como el test ϕ' obtenido en el apartado anterior no depende del valor de la alternativa θ_1, sino simplemente de que sea mayor que θ_0 (es necesario el signo de su diferencia con θ_0 en la determinación de la región crítica), será (*PIE*, pag. 348) uniformemente de máxima potencia para contrastar $H_0:\theta=\theta_0$ frente a $H_1:\theta>\theta_0$.

La función de potencia del test será

$$\beta_{\phi'}(\theta)=\mathcal{P}_\theta\left\{-\sum_{i=1}^{n}\log(1-X_i)<\frac{\chi^2_{2n;1-\alpha}}{2\theta_0}\right\}=P\left\{\chi^2_{2n}<\frac{\theta\,\chi^2_{2n;1-\alpha}}{\theta_0}\right\}$$

e) El test de razón de verosimilitudes es por definición, (*PIE, pag. 388*), aquel que tiene por región crítica la dada por

$$\Lambda(x_1,...,x_n)<b$$

siendo

$$\Lambda(x_1,...,x_n)=\frac{\sup_{\theta\in\Theta_0}f_\theta(x_1,...,x_n)}{\sup_{\theta\in\Theta}f_\theta(x_1,...,x_n)}.$$

Como vimos en el primer apartado, el estimador de máxima verosimilitud de θ es

$$\hat\theta=\frac{n}{-\sum_{i=1}^{n}\log(1-X_i)}=\frac{n}{-T}.$$

Por tanto,

$$\sup_{\theta\in\Theta} f_\theta(x_1,...,x_n) = f_{\hat\theta}(x_1,...,x_n) = \frac{n^n}{(-T)^n}\prod_{i=1}^n (1-x_i)^{-n/T-1} = \frac{n^n}{(-T)^n}\, e^{-n-T}.$$

Por otra parte, como $f_\theta(x_1,...,x_n)$ (o su logaritmo) es una función creciente para $\theta \in (0,\hat\theta)$ y decreciente para $\theta \in (\hat\theta,\infty)$, el máximo para $\theta \le \theta_0$ se alcanzará en

$$\hat\theta_0 = \begin{cases} \theta_0 & \text{si} \quad \hat\theta > \theta_0 \\ \\ \hat\theta & \text{si} \quad \hat\theta \le \theta_0 \end{cases}$$

y vale

$$\sup_{\theta\le\theta_0} f_\theta(x_1,...,x_n) = \begin{cases} \theta_0^n \prod_{i=1}^n (1-x_i)^{\theta_0-1} = \theta_0^n\, e^{(\theta_0-1)T} & \text{si} \quad \hat\theta > \theta_0 \\ \\ \dfrac{n^n}{(-T)^n}\, e^{-n-T} & \text{si} \quad \hat\theta \le \theta_0 \end{cases}$$

Por tanto será,

$$\Lambda(x_1,...,x_n) = \begin{cases} \dfrac{\theta_0^n}{n^n}\,(-T)^n e^{\theta_0 T+n} & \text{si} \quad T > -n/\theta_0 \\ \\ 1 & \text{si} \quad T \le -n/\theta_0 \end{cases}$$

Tomando logaritmos se comprueba que Λ es una función decreciente de T; por tanto, $\Lambda(x_1,...,x_n) < b$ equivale a $T > c'$. Es decir, el test ϕ' obtenido en el apartado c).

En el apartado d) vimos que el test ϕ' maximiza la potencia uniformemente en toda la hipótesis alternativa $\Theta_1 = \{\theta : \theta > \theta_0\}$, entre todos los tests ϕ que cumplen la condición impuesta por la hipótesis nula, $E_{\theta_0}[\phi(x_1,...,x_n)] \le \alpha$.

Ahora buscamos el óptimo en un subconjunto de los tests antes considerados, el de los tests ϕ que cumplen la condición

$$E_\theta[\phi(x_1,...,x_n)] \le \alpha \qquad\qquad \forall\, \theta \le \theta_0$$

condición más restrictiva que la anterior, impuesta por la hipótesis nula $H_0 : \theta \le \theta_0$.

Como el test ϕ' antes obtenido tiene función de potencia no decreciente en θ, valiendo $\beta_\phi(\theta_0) = \alpha$, será

$$E_\theta[\phi'(x_1,...,x_n)] \le \alpha \qquad\qquad \forall\, \theta \le \theta_0$$

por lo que ϕ' cumple la nueva restricción, es decir, pertenece a la clase de tests en donde se busca ahora el óptimo. En consecuencia, ϕ' será también uniformemente de máxima potencia para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$. (*PIE*, pag. 348.)

Una forma alternativa de razonar es observar que la distribución de la muestra

$$f_\theta(x_1, ..., x_n) = \left(\frac{\theta}{1-\theta}\right)^n \exp\left\{\frac{2\theta - 1}{1-\theta} \sum_{i=1}^{n} \log(1 - X_i)\right\} \qquad x_1, ..., x_n \in (0, 1)$$

es de tipo exponencial

$$f_\theta(x_1, ..., x_n) = c(\theta)h(x_1, ..., x_n)e^{q(\theta)T(x_1,...,x_n)}$$

con $q(\theta) = (2\theta - 1)/(1 - \theta)$ y $T(x_1, ..., x_n) = \sum \log(1 - x_i)$, teniendo en consecuencia razón de verosimilitud monótona en $T(x_1, ..., x_n)$, por lo que, según el terorema de Karlin-Rubin, (*PIE*, pag. 349), el test uniformemente de máxima potencia buscado es aquel que tiene por región crítica $T(x_1, ..., x_n) > k$, es decir, $\sum \log(1 - x_i) > c$, que es el test ϕ' antes determinado.

Problema 1.2

Los datos `captopril.txt` que aparecen a continuación y que tenéis en el curso virtual

	Paciente	Sistoantes	Sistodespues	Diastoantes	Diastodespues
1	1	210	201	130	125
2	2	169	165	122	121
3	3	187	166	124	121
4	4	160	157	104	106
5	5	167	147	112	101
6	6	176	145	101	85
7	7	185	168	121	98
8	8	206	180	124	105
9	9	173	147	115	103
10	10	146	136	102	98
11	11	174	151	98	90
12	12	201	168	119	98
13	13	198	179	106	110
14	14	148	129	107	103
15	15	154	131	100	82

corresponden (MacGregor et al., 1979) a la presión sanguínea sistólica y diastólica, en milímetros de mercurio (mmHg), de 15 pacientes a los que se les midió la tensión arterial inmediatamente antes, Sistoantes y Diastoantes, y dos horas después Sistodespues y Diastodespues, de tomar el medicamento Captopril. ¿Puede

decirse que el Captopril es eficaz en una reducción significativa de la presión sanguínea sistólica? ¿Y en la diastólica?

Resolveremos este problema con R por simplificar pero el alumno podrá resolverlo simplemente con la ayuda de una simple calculadora, especialmente en el examen.

Se trata de hacer un par de tests para datos apareados (CB-sección 5.10). Los cálculos los haremos con R. Primero incorporamos los datos ejecutando

```
> captopril<-read.table("d:\\datos\\captopril.txt",header=T)
```

Los datos apareados de Antes-Después de la presión sanguínea sistólica se obtienen ejecutando

```
> sisto<-captopril[,2]-captopril[,3]
> sisto
 [1]  9  4 21  3 20 31 17 26 26 10 23 33 19 19 23
```

Tenemos ahora 15 datos de una variable de la que queremos analizar si puede admitirse que su media μ_{sisto} es mayor que cero (el Captopril es eficaz), por lo que contrastaremos la hipótesis nula $H_0 : \mu_{sisto} \leq 0$ frente a la alternativa $H_0 : \mu_{sisto} > 0$. Dado que no tenemos más de 30 datos en cada población, deberemos analizar antes la posible normalidad de los datos contrastando ésta mediante un test de Kolmogorov-Smirnov (CB-sección 13.3 y EAR-sección 8.3)

```
> ks.test(sisto,"pnorm",mean(sisto),sd(sisto))

        One-sample Kolmogorov-Smirnov test

data:  sisto
D = 0.1696, p-value = 0.7812
alternative hypothesis: two-sided
```

Con un p-valor igual a $0'7812$ podemos concluir con la aceptación de la hipótesis nula, es decir, la normalidad de los datos. Ahora ya podemos aplicar un test de la t de Student para las hipótesis antes especificadas, en el caso de población normal, tamaños muestrales pequeños y varianza desconocida (CB-sección 7.2), rechazándose la hipótesis nula cuando y sólo cuando sea

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}} > t_{n-1;\alpha}$$

el cual podemos ejecutar como sigue (EAR-sección 4.2.1)

```
> t.test(sisto,alternative="greater")

        One Sample t-test

data:  sisto
t = 8.1228, df = 14, p-value = 5.732e-07
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 14.82793      Inf
sample estimates:
mean of x
 18.93333
```

Un p-valor tan pequeño permite concluir que el Captopril es eficaz en la disminución de la presión sanguínea sistólica.

Repetimos el proceso ahora con la presión sanguínea diastólica:

```
> diasto<-captopril[,4]-captopril[,5]

> ks.test(diasto,"pnorm",mean(diasto),sd(diasto))

        One-sample Kolmogorov-Smirnov test

data:  diasto
D = 0.1565, p-value = 0.8562
alternative hypothesis: two-sided
```

Con un p-valor igual a $0'8562$ podemos concluir con la aceptación de la hipótesis nula, es decir, la normalidad de los datos. Ahora ya podemos aplicar un test de la t de Student para las hipótesis nula $H_0 : \mu_{diasto} \leq 0$ frente a la alternativa $H_0 : \mu_{diasto} > 0$ como sigue:

```
> t.test(diasto,alternative="greater")

        One Sample t-test

data:  diasto
t = 4.1662, df = 14, p-value = 0.0004755
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
```

```
 5.349067      Inf
sample estimates:
mean of x
 9.266667
```

Un p-valor tan pequeño permite concluir que el Captopril también es eficaz en la disminución de la presión sanguínea diastólica.

Referencias

(**CB**). García Pérez, A. (2008). *Estadística Aplicada: Conceptos Básicos*. Editorial UNED. (Código 0184011EP01A02).

(**EAR**). García Pérez, A. (2008). *Estadística Aplicada con R*. Editorial UNED. (Código 0137352PB01A01).

MacGregor, G.A., Markandau, N.D., Roulston, J.E. y Jones, J.C. (1979). inhibitor of angiotensin-converting enzyme. *British Medical Journal*, **2**, 1106-1109.

(**PIE**). Vélez, R. y García Pérez, A. *Principios de Inferencia Estadística*. Editorial UNED.

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2014-2015

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Los Tutores deberán de haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Sobre 10 puntos, los dos primeros problemas de la Evaluación Continua valen 4 puntos cada uno y el tercero, 2 puntos.

Esta Evaluación Continua es un buen ejemplo de lo que será la Prueba Presencial de este año aunque en dicho examen presencial no se trabajaría con un número elevado de datos.

Ejercicios para la Evaluación Continua

Problema 1.1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_{\lambda\mu}(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\ \exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\} \qquad \text{si} \quad x > 0$$

siendo $\lambda > 0$ y $\mu > 0$. Determinar, utilizando una muestra aleatoria simple de tamaño n de X, los estimadores de máxima verosimilitud T_1 de μ y T_2 de $1/\lambda$. Analizar su suficiencia minimal.
Supuesto que T_2 es completo y que $\lambda n T_2$ sigue una distribución χ^2 con $n-1$ grados de libertad, determinar el estimador centrado uniformemente de mínima varianza de $1/\lambda$, así como un intervalo de confianza de nivel de confianza $1-\alpha = 0'95$ para $1/\lambda$, supuesto que es $n = 20$.

La función de verosimilitud de la muestra es

$$f_{\lambda,\mu}(x_1,...,x_n) = \frac{\lambda^{n/2}}{(2\pi)^{n/2}}\ \frac{1}{\prod_{i=1}^n x_i^{3/2}} \exp\left\{-\frac{\lambda}{2\mu^2}\sum_{i=1}^n x_i + \frac{n\lambda}{\mu} - \frac{\lambda}{2}\sum_{i=1}^n \frac{1}{x_i}\right\}$$

con lo que, derivando parcialmente respecto a λ y μ, se obtiene el sistema

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\log f_{\lambda,\mu}(x_1,...,x_n) &= \frac{n}{2\lambda} - \frac{1}{2\mu^2}\sum_{i=1}^n x_i + \frac{n}{\mu} - \frac{1}{2}\sum_{i=1}^n \frac{1}{x_i} = 0 \\
\frac{\partial}{\partial\mu}\log f_{\lambda,\mu}(x_1,...,x_n) &= \frac{\lambda}{\mu^3}\sum_{i=1}^n x_i - \frac{n\lambda}{\mu^2} = 0
\end{aligned}$$

a partir del cual se obtienen los estimadores máximo verosímiles

$$\begin{aligned}
T_1 = \hat{\mu} &= \overline{x} \\
T_2 = \frac{1}{\hat{\lambda}} &= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{X_i} - \frac{1}{\overline{x}}\right).
\end{aligned}$$

Bien utilizando el teorema de factorización o, por ser el modelo una familia exponencial, el estimador bidimensional $(\sum X_i\,,\ \sum(1/X_i))$ es suficiente minimal. Como T_2 es función del suficiente y completo, si su esperanza es $1/\lambda$, será el ECUMV. Al ser $\lambda n T_2 \rightsquigarrow \chi^2_{n-1}$ será $E[T_2] = (n-1)/(\lambda n)$, por lo que, en definitiva, el ECUMV para $1/\lambda$ será

$$T_3 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{X_i} - \frac{1}{\overline{x}}\right).$$

Por último, al ser $\lambda n T_2 \rightsquigarrow \chi^2_{n-1}$ se podrán determinar dos valores $\chi^2_{n-1;1-\alpha/2}$ y $\chi^2_{n-1;\alpha/2}$ tales que

$$P\left\{\chi^2_{n-1;1-\alpha/2} < \lambda n T_2 < \chi^2_{n-1;\alpha/2}\right\} = 1-\alpha$$

con lo que despejando,

$$\left[\frac{nT_2}{\chi^2_{n-1;\alpha/2}} < \frac{1}{\lambda} < \frac{nT_2}{\chi^2_{n-1;1-\alpha/2}}\right]$$

será el intervalo de confianza buscado. Para $n = 20$ y $\alpha = 0'05$ queda igual a

$$\left[\frac{20T_2}{32'85} < \frac{1}{\lambda} < \frac{20T_2}{8'906}\right].$$

Problema 1.2

Determinar un contraste de máxima potencia de nivel α $(0 < \alpha < 0'5)$ para contrastar $H_0 : f_0$ frente a $H_1 : f_1$, en donde

$$f_0(x) = \begin{cases} 4x & \text{si } 0 < x < 1/2 \\ 4 - 4x & \text{si } 1/2 \le x < 1 \end{cases}$$

y

$$f_1(x) = 1 \quad \text{si} \quad 0 < x < 1$$

basado en una muestra aleatoria simple de tamaño uno de X. Determinar su función de potencia.

El lema de Neyman-Pearson nos dice que el test de máxima potencia es el que tiene como región crítica los x tales que

$$f_1(x) > k\, f_0(x)$$

o, equivalentemente, tales que

$$f_0(x) < k'$$

es decir, (Figura 1)

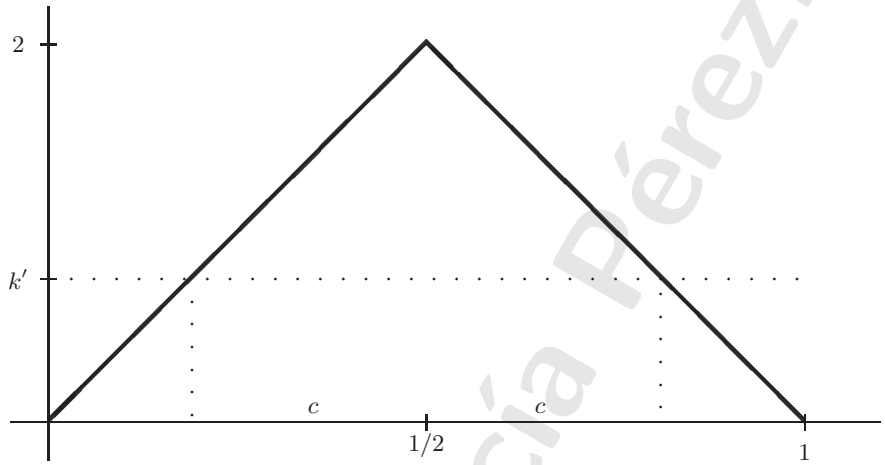

Figura 1

el contraste de máxima potencia será

$$\varphi(x) = \begin{cases} 1 & \text{si } |x - \frac{1}{2}| > c \\ 0 & \text{si } |x - \frac{1}{2}| \leq c \end{cases}$$

siendo c tal que $\alpha = E_0[\varphi(X)]$; es decir,

$$\begin{aligned} \alpha &= P_0\{|X - \frac{1}{2}| > c\} \\ &= P_0\{X < \frac{1}{2} - c\} + P_0\{X > \frac{1}{2} + c\} \\ &= \int_0^{1/2-c} 4x\,dx + \int_{1/2+c}^{1} (4 - 4x)\,dx \end{aligned}$$

De donde se obtiene que debe ser $c = (1 - \sqrt{\alpha})/2$.

La potencia será

$$P_1\left\{\left|X - \frac{1}{2}\right| > \frac{1 - \sqrt{\alpha}}{2}\right\} = \sqrt{\alpha}.$$

Problema 1.3

Se piensa que el tipo de vida que llevaban los Neandertales no se diferenciaba mucho de la que llevan hoy en día los cowboys americanos, en cuanto a su "relación" con la naturaleza. Para analizar esta idea, se compararon fracturas en fósiles de este homínido y en actuales cowboys americanos, obteniéndose los siguientes resultados,

	Pie	Pierna	Pelvis	Cabeza	Brazo	Tronco
Neandertales	**12**	**10**	**9**	**27**	**15**	**7**
Cowboys	**5**	**3**	**4**	**18**	**15**	**5**

¿Qué conclusiones obtendría?

Como lo que queremos hacer es *comparar dos poblaciones*, la de Neandertales y la de Vaqueros, en las que tenemos datos que son recuentos de observaciones, el test adecuado es el de la χ^2 de *homogeneidad de varias muestras* (CB-sección 12.3), en donde la hipótesis nula que se establece es que ambas poblaciones pueden considerase homogéneas. Ésta se rechaza cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson. En nuestro caso, toma el valor $\lambda = 3'9978$.

De la tabla de la χ^2 de Pearson vemos que el p-valor es

$$P\{\chi^2_{(r-1)(s-1)} > 3'9978\} = P\{\chi^2_5 > 3'9978\} > 0'3$$

suficientemente grande como para aceptar la hipótesis nula de homogeneidad con bastante seguridad.

Si queremos resolverlo con R (EAR-sección 7.3), ejecutaríamos la siguiente secuencia de instrucciones. Con `(1)` incluimos los datos, que tienen que venir en forma de matriz. Recordemos que, por defecto, los incorpora por columnas. Si quisiéramos que los incorporara por filas, deberíamos ponerlos por filas en la función `c`, e incluir el argumento `byrow = T`. Las sentencias `(2)` y `(3)` son opcionales y sirven para poner nombre a las filas y a las columnas de la tabla. Con `(4)` comprobamos que hemos incorporado bien los datos a R. Ejecutando `(5)` es como le pedimos que haga el test χ^2.

```
> vidaNean<-matrix(c(12,5,10,3,9,4,27,18,15,15,7,5),ncol=6)                    (1)
> colnames(vidaNean)<-c("Pie","Pierna","Pelvis","Cabeza","Brazo","Tronco")     (2)
> rownames(vidaNean)<-c("Neandertales","Vaqueros")                             (3)
> vidaNean                                                                     (4)
             Pie Pierna Pelvis Cabeza Brazo Tronco
Neandertales  12     10      9     27    15      7
Vaqueros       5      3      4     18    15      5
> chisq.test(vidaNean)                                                        (5)

        Pearson's Chi-squared test

data:  vidaNean
X-squared = 3.9978, df = 5, p-value = 0.5497                                  (6)

Warning message:
In chisq.test(vidaNean) : Chi-squared approximation may be incorrect
```

En (6) vemos el valor del estadístico de Pearson, $\lambda = 3'9978$ y el p-valor del test, $0'5497$, suficientemente grande como para concluir que puede aceptarse la hipótesis nula de homogeneidad de ambas poblaciones. Es decir, puede concluirse que llevaban un tipo de vida semejante en cuanto a las fracturas corporales.

El aviso final que aparece, indica que alguna de las frecuencias esperadas es menor que 5, como podemos comprobar ejecutando (7) y viendo que el último elemento de la matriz lo es.

```
> chisq.test(vidaNean)$expected                                               (7)
                  Pie Pierna Pelvis   Cabeza    Brazo   Tronco
Neandertales 10.461538      8      8 27.69231 18.46154 7.384615
Vaqueros      6.538462      5      5 17.30769 11.53846 4.615385
Warning message:
In chisq.test(vidaNean) : Chi-squared approximation may be incorrect
```

Dado que es menor que 5 por muy poco y que las conclusiones del análisis son muy claras, no merece la pena utilizar la corrección de Yates, que deberíamos calcular mediante su fórmula, puesto que R no lo suministra al no ser la tabla 2×2.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

MR: **Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo**, 2005. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0186080EP03A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2015-2016

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Se aconseja a los alumnos no esperar al último momento ya que el sistema puede fallar y sólo se admitirán trabajos a través del Curso Virtual. Los Tutores deberán de haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Esta Evaluación Continua es un buen ejemplo de lo que será la Prueba Presencial de este año.

Ejercicios para la Evaluación Continua

Problema 1.1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta^2}{2}\, e^{x-\theta\,(e^x/2)} \qquad\qquad -\infty < x < \infty$$

siendo $\theta > 0$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
a)Determinar el estimador de máxima verosimilitud para θ.

1 punto

b) Determinar el estimador centrado uniformemente de mínima varianza para $1/\theta$.

3 puntos

c)Determinar un intervalo de confianza para θ, de nivel de confianza $1-\alpha$.

2 puntos

d)Determinar un test uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$.

2 puntos

---

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{\theta^{2n}}{2^n}\right) \exp\left\{\sum_1^n x_i - \theta \sum_1^n e^{x_i/2}\right\}$$

con lo cual será

$$\log f_\theta(x_1, ..., x_n) = 2n \log\theta - \log 2^n + \sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} e^{x_i/2}$$

obteniéndose la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta} \log f_\theta(x_1, ..., x_n) = \frac{2n}{\theta} - \sum_{i=1}^{n} e^{x_i/2} = 0$$

y, por tanto, el estimador de máxima verosimilitud para θ

$$T = \frac{2n}{\sum_{i=1}^n e^{X_i/2}}.$$

b) De entre las diferentes maneras de determinar el ECUMV de $1/\theta$, la más directa es la de utilizar el teorema de Lehmann-Scheffé (*PIE*, pp. 218) aunque para ello es necesario contar con un estimador suficiente y completo.

Como la distribución modelo es una familia de tipo exponencial (*PIE*, pp. 174) de la forma

$$f_\theta(x) = \frac{\theta^2}{2}\, e^x\, e^{-\theta\, e^{x/2}}$$

será $S(X_1, ..., X_n) = \sum_{i=1}^n e^{X_i/2}$ un estadístico minimal suficiente. Por tanto, (*PIE*, pp. 273), S será además completo si la imagen de $q(\theta) = -\theta$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$, la imagen de la función q será $(-\infty, 0)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo; basta ahora con encontrar un estimador insesgado para $1/\theta$ función de S y ése será el ECUMV para $1/\theta$.

En este punto es siempre muy útil determinar la distribución del estadístico S. Para ello, vamos a determinar en primer lugar la distribución de $Y = e^{X/2}$. Su función de densidad,

$$g_Y(y) = f_\theta(2\log y)\, 2\, \frac{1}{y} = \theta^2\, y\, e^{-\theta\, y} \qquad\qquad y > 0$$

es la de una gamma $\gamma(2, \theta)$. Por tanto, $S = \sum_{i=1}^n e^{X_i/2}$ seguirá una distribución $\gamma(2n, \theta)$ y, en consecuencia, $E[S] = 2n/\theta$. Por tanto, el ECUMV para $1/\theta$ será

$$\frac{S}{2n} = \frac{\sum_{i=1}^n e^{X_i/2}}{2n}.$$

c) Al ser $S \rightsquigarrow \gamma(2n, \theta)$, será $2\theta\, S \rightsquigarrow \gamma(4n/2, 1/2) = \chi^2_{4n}$, con lo que se podrán determinar dos valores $\chi^2_{4n;1-\alpha/2}$ y $\chi^2_{4n;\alpha/2}$ tales que

$$P\left\{\chi^2_{4n;1-\alpha/2} < 2\theta\, S < \chi^2_{4n;\alpha/2}\right\} = 1 - \alpha$$

con lo que, despejando,

$$\left[\frac{\chi^2_{4n;1-\alpha/2}}{2S}\ ,\ \frac{\chi^2_{4n;\alpha/2}}{2S}\right]$$

será el intervalo de confianza buscado.

d) Al ser f_θ una familia de tipo exponencial, tendrá razón de verosimilitud monótona en el estadístico S; en concreto, si $\theta_1 < \theta_2$ será

$$\frac{f_{\theta_2}(x_1,...,x_n)}{f_{\theta_1}(x_1,...,x_n)} = \left(\frac{\theta_2}{\theta_1}\right)^{2n} e^{-S(\theta_2-\theta_1)}$$

creciente en $-S$, por lo que por el teorema de Karlin-Rubin (*PIE*, pp. 349) existe un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \geq \theta_0$ frente a $H_1 : \theta < \theta_0$, el cual viene dado por

$$\phi(x_1,...,x_n) = \begin{cases} 1 & \text{si} \quad -S(x_1,...,x_n) < c \\ \\ 0 & \text{si} \quad -S(x_1,...,x_n) \geq c \end{cases}$$

es decir,

$$\phi(x_1,...,x_n) = \begin{cases} 1 & \text{si} \quad S(x_1,...,x_n) > k \\ \\ 0 & \text{si} \quad S(x_1,...,x_n) \leq k \end{cases}$$

en donde k se determina por la condición del nivel,

$$E_{\theta_0}\left[\phi(X_1,...,X_n)\right] = \alpha,$$

en nuestro caso, $\alpha = P_{\theta_0}\{S(X_1,...,X_n) > k\}$. Es decir,

$$\alpha = P\{\chi^2_{4n} > 2\theta_0 k\}.$$

Por tanto, deberá ser $\chi^2_{4n;\alpha} = 2\theta_0 k$ y, en consecuencia,

$$k = \frac{\chi^2_{4n;\alpha}}{2\theta_0}.$$

Problema 1.2

En un estudio sobre la calidad en la atención proporcionada por cuatro residencias de ancianos, se tomó una muestra de 100 ancianos residentes válidos en cada una de las cuatro y se les preguntó si ellos calificaban la atención recibida como Buena o no. Se obtuvieron los siguientes datos:

Residencia	Atención considerada como Buena
A	**60**
B	**62**
C	**40**
D	**54**

¿Se pueden aceptar como equivalentes las cuatro residencias en opinión de los ancianos?

Como lo que queremos hacer es *comparar cuatro poblaciones*, las cuatro residencias, en las que tenemos datos que son recuentos de observaciones, el test adecuado es el de la χ^2 de *homogeneidad de varias muestras* (CB-sección 12.3), en donde la hipótesis nula que se establece es que las cuatro residencias pueden considerase homogéneas. Ésta se rechaza cuando y sólo cuando sea

$$\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$$

siendo

$$\lambda = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_i\, m_j/n)^2}{n_i\, m_j/n}$$

el estadístico de Pearson. En nuestro caso, toma el valor $\lambda = 11'9163$.

De la tabla de la χ^2 de Pearson vemos que el p-valor es

$$P\{\chi^2_{(r-1)(s-1)} > 11'9163\} = P\{\chi^2_3 > 11'9163\} < 0'01$$

suficientemente pequeño como para rechazar con claridad la hipótesis nula de homogeneidad.

Si queremos resolverlo con R (EAR-sección 7.3), ejecutaríamos la siguiente secuencia de instrucciones. Con `(1)` incluimos los datos, que tienen que venir en forma de matriz. Recordemos que, por defecto, los incorpora por columnas. Si quisiéramos que los incorporara por filas, deberíamos ponerlos por filas en la función `c`, e incluir el argumento `byrow = T`. Las sentencias `(2)` y `(3)` son opcionales y sirven para poner nombre a las filas y a las columnas de la tabla. Con `(4)` comprobamos que hemos incorporado bien los datos a R. Ejecutando `(5)` es como le pedimos que haga el test χ^2.

```
> calidad<-matrix(c(60,62,40,54,40,38,60,46),ncol=2)                    (1)
> colnames(calidad)<-c("Buena","Mala")                                  (2)
> rownames(calidad)<-c("A","B","C","D")                                 (3)

> calidad                                                               (4)
  Buena Mala
A    60   40
B    62   38
C    40   60
D    54   46

> chisq.test(calidad)                                                   (5)


        Pearson's Chi-squared test

data:  calidad
```

```
X-squared = 11.9163, df = 3, p-value = 0.007676                    (6)

Warning message:
In chisq.test(vidaNean) : Chi-squared approximation may be incorrect
```

En (6) vemos el valor del estadístico de Pearson, $\lambda = 11'9163$ y el p-valor del test, $0'007676$, suficientemente pequeño como para concluir que debe rechazarse la hipótesis nula de homogeneidad de las cuatro poblaciones. Es decir, puede concluirse que, en opinión de los ancianos residentes, las cuatro residencias no pueden considerarse homogéneas.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

MR: **Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo**, 2005. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0186080EP03A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2016-2017

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Se aconseja a los alumnos no esperar al último momento ya que el sistema puede fallar y sólo se admitirán trabajos a través del Curso Virtual. Los Tutores Intercampus deberán haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Ejercicios para la Evaluación Continua

Problema 1.1

Sea X una variable aleatoria discreta, de media $1/(1-\theta)$ y varianza $\theta/(1-\theta)^3$, cuya función de masa o probabilidad es

$$f_\theta(x) = \frac{1}{(x-1)!}\, x^{x-2}\, e^{-\theta x}\, \theta^{x-1} \qquad \text{para} \quad x = 1, 2, 3, ...$$

siendo θ un parámetro desconocido tal que $0 < \theta < 1$. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
a) Determinar un estadístico suficiente minimal para la familia anterior.

1.5 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = 1/(1-\theta)$.

2 puntos

c) Determinar la cota de Fréchet-Cramer-Rao para los estimadores insesgados de $g(\theta) = 1/(1-\theta)$. El estimador determinado en el apartado anterior, ¿es eficiente para $g(\theta) = 1/(1-\theta)$?

3 puntos

---

a) La determinación del estadístico suficiente minimal se puede hacer directamente utilizando el teorema de factorización (PIE, pp. 167) y el teorema 5.2 (PIE, pp. 171) aunque resulta más fácil si utilizamos el hecho de que la variable aleatoria dada es una familia de tipo exponencial (PIE, pp. 174) de la forma

$$f_\theta(x) = \frac{1}{(x-1)!}\, x^{x-2}\, \frac{1}{\theta}\, \exp\left\{x(\log\theta - \theta)\right\}.$$

Por tanto, (ver Ejemplo 5.17, PIE, pp. 174) el estadístico

$$T(X_1, ..., X_n) = \sum_{i=1}^{n} X_i$$

es minimal suficiente.

b) Para determinar el ECUMV de cualquier función del parámetro, lo primero que necesitamos es determinar un estadístico suficiente y completo para la familia de distribuciones $f_\theta(x)$.

Ya hemos visto, en el apartado anterior, que el estadístico $T(X_1, ..., X_n) = \sum_{i=1}^{n} X_i$ es un estadístico minimal suficiente. Por tanto, (PIE, pp. 273), T

será además completo si la imagen de $q(\theta) = \log(\theta) - \theta$ contiene un abierto de $\mathbb{R}$; como es $\theta \in (0,1)$, la imagen de la función q será $(-\infty, -1)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, T será suficiente y completo. Si ahora somos capaces de intuir un estadístico función de T que sea insesgado para $g(\theta) = 1/(1-\theta)$, utilizando el teorema de Lehmann-Scheffé (PIE, pp. 218), ése será el ECUMV de $g(\theta)$. Para ello es razonable determinar la esperanza de T,

$$E[T] = E\left[\sum_{i=1}^{n} X_i\right] = n\,E[X] = \frac{n}{1-\theta}$$

con lo que el ECUMV para $g(\theta) = 1/(1-\theta)$ será la media muestral $\overline{x} = \sum_{i=1}^{n} X_i/n$.

c) La cota de Fréchet-Cramer-Rao para los estimadores insesgados S de $g(\theta) = 1/(1-\theta)$ será (PIE, pp. 225)

$$V_\theta(T) \geq \frac{[g'(\theta)]^2}{I(\theta)}$$

en donde la cantidad de información de Fisher $I(\theta)$ es

$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X_1, ..., X_n)\right)^2\right].$$

En nuestro caso es

$$f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{x_i^{x_i-2}}{(x_i-1)!}\; e^{-\theta\sum_1^n x_i}\;\theta^{\sum_1^n x_i - n}$$

con lo que será

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f_\theta(x_1, ..., x_n) &= -\sum_{i=1}^{n} x_i + \frac{\sum_{i=1}^{n} x_i - n}{\theta} \\
&= \frac{1}{\theta}\left(\sum_{i=1}^{n}[x_i(1-\theta) - 1]\right) \\
&= \frac{1-\theta}{\theta}\sum_{i=1}^{n}\left[x_i - \frac{1}{1-\theta}\right]
\end{aligned}$$

y, por tanto,

$$\begin{aligned} I(\theta) &= E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X_1,...,X_n)\right)^2\right] \\ &= \frac{(1-\theta)^2}{\theta^2}\, n\, E[(X-E[X])^2] \\ &= \frac{(1-\theta)^2\, n}{\theta^2}\, V(X) \\ &= \frac{n}{\theta(1-\theta)} \\ &= E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_1,...,X_n)\right] \end{aligned}$$

con lo que la cota de Fréchet-Cramer-Rao buscada será igual a

$$V_\theta(T) \geq \frac{1/(1-\theta)^4}{n/(\theta(1-\theta))} = \frac{\theta}{n(1-\theta)^3}.$$

El estimador determinado en el apartado anterior, la media muestral, será eficiente, por definición (PIE, pp. 228), si su varianza coincide con la cota de Fréchet-Cramer-Rao. La varianza de $\overline{x}$ es

$$V(\overline{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{V(X)}{n} = \frac{\theta}{n(1-\theta)^3}$$

siendo, por tanto, $\overline{x}$ eficiente para $g(\theta) = 1/(1-\theta)$.
(La distribución modelo $f_\theta(x)$ es una distribución de Borel-Tanner. LCP, pp. 296.)

Problema 1.2

Sea X una variable aleatoria simple con distribución $f_0 = N(0,1)$ bajo la hipótesis nula H_0 y con distribución de Cauchy de función de densidad, $f_1(x) = 1/(\pi(1+x^2))$, bajo la hipótesis alternativa H_1. Determinar un contraste de máxima potencia de nivel α para contrastar H_0 frente a H_1 utilizando una muestra de tamaño uno de X.

2 puntos

El lema de Neyman-Pearson nos dice (PIE, pp. 338) que el test de máxima potencia tiene como región crítica la dada por

$$f_1(x_1, ..., x_n) > k f_0(x_1, ..., x_n)$$

es decir, si el tamaño muestral es $n = 1$, la constituida por aquellos x tales que

$$\frac{e^{x^2/2}}{1+x^2} > k'$$

es decir, aquellos x para los que la función

$$g(x) = \frac{e^{x^2/2}}{1+x^2}$$

sea mayor que la constante k'.

Esta función tiene como representación gráfica la Figura 1.1 obtenida al ejecutar con R

```
> x<-seq(-2,2,len=100)
> plot(x,(exp(x^2/2))/(1+x^2),type="l",lwd=2,ylab="g(x)",main="Función g")
> points(0,1,pch=16,col=4)
> text(0,1.02,"(0,1)")
```

estando k' determinada por la condición del nivel, es decir, tal que

$$P_{H_0}\left\{(\exp(X^2/2))/(1+X^2) > k'\right\} = \alpha$$

con $X \rightsquigarrow N(0,1)$.

Si *alpha* es pequeño, los x que verifican la condición anterior irán completando de forma simétrica las dos colas de la normal $N(0,1)$ desde los extremos $(-\infty, -c)$ y (c, ∞) (proyección sobre el eje de abscisas de línea roja de la 1.1) formando la región crítica $|x| > c$ hasta que el el punto crítico llegue a $k' = 1$ a partir del cual tendrán que ir llenando otra parte alrededor de $x = 0$ (proyección sobre el eje de abscisas de línea azul de la 1.1).

En definitiva, si α es pequeño será $k' < 1$ (punto azul de la figura) y la región crítica de la forma

$$\frac{e^{x^2/2}}{1+x^2} > k' \Leftrightarrow |x| > c$$

en donde $c = z_{\alpha/2}$.

Si fuera α grande, la región crítica sería de la forma $\{|x| > c\} \cup \{|x| < c'\}$ siendo c y c' tales que

$$P_0\{0 < X < c'\} + P_0\{X > c\} = \alpha/2$$

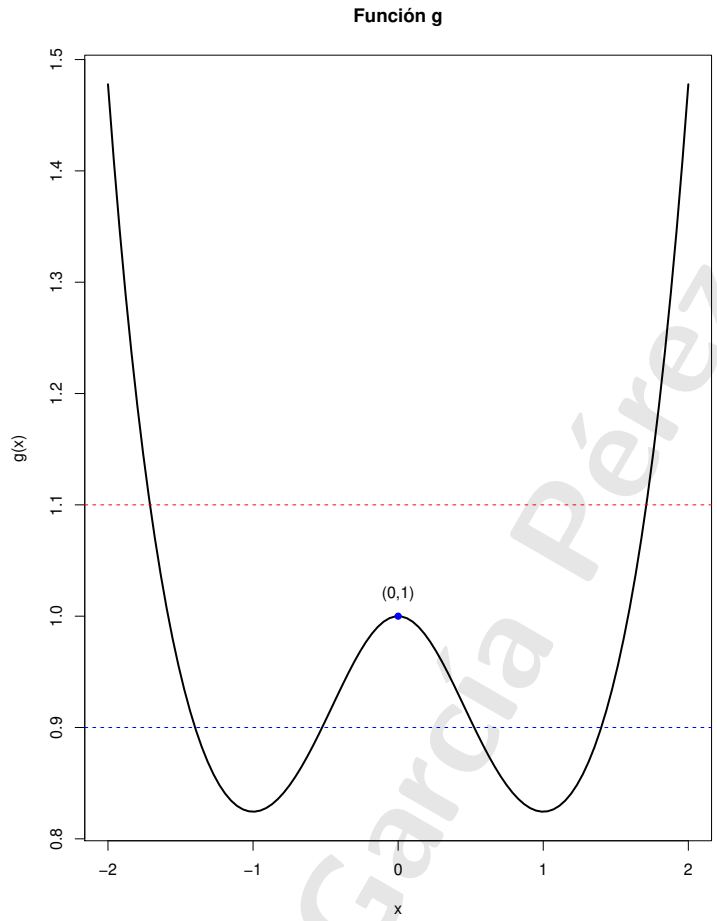

Figura 1.1 Función g

y además verificando la relación

$$\frac{e^{c'^2/2}}{1+c'^2} = \frac{e^{c^2/2}}{1+c^2}$$

El α_0 que diferencia ambas regiones críticas será aquel en el que el correspondiente c_0 (con $P_0\{X > c_0\} = \alpha_0/2$) cumpla

$$\frac{e^{c_0^2/2}}{1+c_0^2} = 1.$$

Por completar el ejercicio, digamos que es $c_0 = 1'585$ y $\alpha_0 = 0'113$.

Problema 1.3

Se cree que los niveles de beta-endorfina en sangre aumentan cuando aumenta el estrés emocional. Para analizar esta hipótesis se midió dicho nivel en 19 pacientes que iban a ser sometidos a una intervención quirúrgica efectuando las mediciones, (a) 12-14 horas antes de la operación y (b) 10 minutos antes de la intervención. Los datos obtenidos (Hoaglin et al., 1985) de los niveles de beta-endorfina en fmol/ml fueron los siguientes, en donde el lugar de cada dato se corresponde en ambas situaciones con un mismo individuo: es decir, al primer individuo le observaron unos niveles de 10 en la situación (a) y de 6′5 en la situación (b), ..., al individuo 19 un nivel de 2 en la situación (a) y 2 en la situación (b).

(a)	10	6′5	8	12	5	11′5	5	3′5	7′5	5′8
	4′7	8	7	17	8′8	17	15	4′4	2	
(b)	6′5	14	13′5	18	14′5	9	18	42	7′5	6
	25	12	52	20	16	15	11′5	2′5	2	

Analizar la hipótesis del incremento en los niveles de beta-endorfina, verificando las condiciones necesarias para que el test que aplique sea válido.

1.5 puntos

El enunciado nos pide un test para datos apareados ya que será la variable diferencia Di=(b)-(a) la que observemos, estableciendo las hipótesis a contrastar de $H_0 : \mu_{Di} \leq 0$ frente a la alternativa $H_1 : \mu_{Di} > 0$, siendo μ_{Di} el nivel medio de la variable diferencia Di.

Como es sabido, el caso de datos apareados (CB-sección 5.10) se trata como un caso de datos de la variable unidimensional diferencia, Di. En este caso estamos ante el caso un test para la media poblacional de unos datos de tamaño muestral pequeño (menor que 30) por lo que necesitamos analizar primero si puede admitirse que éstos proceden de una distribución normal.

Los cálculos los haremos con R. Primero incorporamos los datos ejecutando

```
> a<-c(10,6.5,8,12,5,11.5,5,3.5,7.5,5.8,4.7,8,7,17,8.8,17,15,4.4,2)
> b<-c(6.5,14,13.5,18,14.5,9,18,42,7.5,6,25,12,52,20,16,15,11.5,2.5,2)
```

Los datos apareados de Di=(b)-(a) se obtienen ejecutando

```
> Di<-b-a
> Di
 [1] -3.5  7.5  5.5  6.0  9.5 -2.5 13.0 38.5  0.0  0.2 20.3  4.0 45.0  3.0  7.2
[16] -2.0 -3.5 -1.9  0.0
```

Tenemos ahora 19 datos de una variable de la que queremos analizar si puede admitirse que su media μ_D es mayor que cero (el estrés aumenta). Dado que no tenemos más de 30 datos en cada población, deberemos analizar antes la posible normalidad de los datos contrastando ésta mediante, por ejemplo, un test de Kolmogorov-Smirnov (CB-sección 13.3 y EAR-sección 8.3)

```
> ks.test(Di,"pnorm",mean(Di),sd(Di))

        One-sample Kolmogorov-Smirnov test

data:  Di
D = 0.2427, p-value = 0.2128
alternative hypothesis: two-sided
```

Con un p-valor igual a $0'2128$ podemos concluir con la aceptación de la hipótesis nula, es decir, la normalidad de los datos. Ahora ya podemos aplicar un test de la t de Student para las hipótesis antes especificadas, en el caso de población normal, tamaños muestrales pequeños y varianza desconocida (CB-sección 7.2), rechazándose la hipótesis nula cuando y sólo cuando sea

$$\frac{\overline{x}-\mu_0}{S/\sqrt{n}} > t_{n-1;\alpha}$$

el cual podemos ejecutar como sigue (EAR-sección 4.2.1)

```
> t.test(Di,alternative="greater")

        One Sample t-test

data:  Di
t = 2.4827, df = 18, p-value = 0.01156
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 2.321786      Inf
sample estimates:
mean of x
      7.7
```

El p-valor, $0'01156$, no es concluyente aunque sí lo suficientemente pequeño como para rechazar la hipótesis nula y decir que el estrés sí incrementa los niveles de beta-endorfina significativamente.

Referencias

Hoaglin, D.C., Mosteller, F. y Tukey, J,W. (1985). *Exploring data tables, trends and shapes.* Ed. Wiley.

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Addendas (código: 0141206AD01A01).

MR: **Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo**, 2005. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0186080EP03A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.

INFERENCIA ESTADÍSTICA

Soluciones a los Ejercicios para la Evaluación Continua

Curso 2017-2018

Alfonso García Pérez

Universidad Nacional de Educación a Distancia

Edita: Universidad Nacional de Educación a Distancia

Capítulo 1

Soluciones a los Ejercicios para la Evaluación Continua

Si se quiere optar por la modalidad de Evaluación Continua, estos ejercicios deberán ser entregados antes del 6 de Mayo obligatoriamente en el Curso Virtual. Se ruega entregar en un formato fácil de acceder como por ejemplo pdf. Se aconseja a los alumnos no esperar al último momento ya que el sistema puede fallar y sólo se admitirán trabajos a través del Curso Virtual. Los Tutores Intercampus deberán haber calificado estas pruebas con una nota de 0 a 10 (que ponderadas en la nota final por $0'2$ sólo sumarán la calificación de la Prueba Presencial entre 0 y 2) antes del final de la primera semana de las pruebas presenciales. Y, como mucho, al comienzo de la segunda semana de pruebas presenciales los alumnos pueden haber reclamado al Tutor por la nota con la que les calificó, de manera que estas calificaciones serán definitivas al final de la segunda semana de exámenes.

Las calificaciones así obtenidas se sumarán a la de la Prueba Presencial, si en ésta se obtuvo una puntuación de 4 o más puntos, truncando a 10 aquellas notas que superen este valor. Así, el alumno podrá obtener hasta una calificación de 10 puntos. No obstante, para obtener una calificación de Matrícula de Honor deberá haber obtenido un 10 en la Prueba Presencial. Por ejemplo, si un alumno obtiene un 1 en la Evaluación Continua y un 4 en la Prueba Presencial, su calificación final será de 5; si obtiene un 1 en la Evaluación Continua y un 10 en la Prueba Presencial, su calificación final será de 10 (MH); si obtiene un 2 en la Evaluación Continua y un $3'5$ en la Prueba Presencial, su calificación final será de $3'5$; si obtiene un 1 en la Evaluación Continua y un 9 en la Prueba Presencial, su calificación final será de 10.

Ejercicios para la Evaluación Continua

Problema 1.1

Sea X una variable aleatoria absolutamente continua con función de densidad

$$f_\theta(x) = \frac{\theta+1}{2^{\theta+1}}\, x^\theta \qquad\qquad 0 < x < 2$$

siendo $\theta > 0$ un parámetro desconocido. Utilizando una muestra aleatoria simple de tamaño n de X, se pide:
a) Determinar el estimador de máxima verosimilitud para θ y analizar si es suficiente.

2 puntos

b) Determinar el estimador centrado uniformemente de mínima varianza para $g(\theta) = 1/(1+\theta)$.

3 puntos

c) Determinar un test uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$.

3 puntos

---

a) La función de verosimilitud de la muestra es

$$f_\theta(x_1,...,x_n) = \prod_{i=1}^n f_\theta(x_i) = \left(\frac{\theta+1}{2^{\theta+1}}\right)^n \prod_{i=1}^n x_i^\theta \qquad \text{si} \quad x_1,...,x_n \in (0,2)$$

con lo cual será

$$\log f_\theta(x_1,...,x_n) = n\left[\log(1+\theta) - (1+\theta)\log 2\right] + \theta \sum_{i=1}^n \log x_i$$

obteniéndose la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta} \log f_\theta(x_1,...,x_n) = \frac{n}{1+\theta} - n\log 2 + \sum_{i=1}^n \log x_i = 0$$

y, por tanto, el estimador de máxima verosimilitud para θ

$$T = \frac{1}{\log 2 - \frac{1}{n}\sum_{i=1}^n \log X_i} - 1.$$

La función de densidad de la muestra se puede factorizar de la forma

$$f_\theta(x_1, ..., x_n) = h(x_1, ..., x_n) \cdot g_\theta(T(x_1, ..., x_n))$$

con

$$h(x_1, ..., x_n) = 1$$

y

$$g_\theta(T(x_1, ..., x_n)) = \left(\frac{\theta+1}{2^{\theta+1}}\right)^n \exp\left\{\theta\, n\, \left[\log 2 - \frac{1}{1+T}\right]\right\}$$

es decir, mediante el producto de una función, h, que no depende del parámetro y una función, g, que depende del parámetro y de la muestra sólo a través del estadístico T. Por el teorema de factorización (*PIE*, pp. 167), T será suficiente para la familia de densidades dada.

b) De entre las diferentes maneras de determinar el ECUMV de $1/\theta$, la más directa es la de utilizar el teorema de Lehmann-Scheffé (*PIE*, pp. 218) aunque para ello es necesario contar con un estimador suficiente y completo.

Como la distribución modelo es una familia de tipo exponencial (*PIE*, pp. 174) de la forma

$$f_\theta(x) = \frac{\theta+1}{2^{\theta+1}}\, e^{\theta \log x}$$

será $S(X_1, ..., X_n) = \sum_{i=1}^n \log X_i$ un estadístico minimal suficiente. Por tanto, (*PIE*, pp. 273), S será además completo si la imagen de $q(\theta) = \theta$ contiene un abierto de $\mathbb{R}$; como es $\theta > 0$, la imagen de la función q será $(0, \infty)$ la cual contiene un abierto (i.e., un intervalo abierto) de $\mathbb{R}$.

En definitiva, S será suficiente y completo; basta ahora con encontrar un estimador insesgado para $1/(1+\theta)$ función de S y ese será el ECUMV para $1/(1+\theta)$.

En este punto es siempre muy útil determinar la distribución del estadístico S. Para ello, vamos a determinar en primer lugar la distribución de $Y = \log X$. Su función de densidad,

$$f_Y(y) = f_\theta(e^y)\, e^y = \frac{\theta+1}{2^{\theta+1}}\, e^{(\theta+1)y} \qquad -\infty < y < \log 2$$

sugiere la transformación $-Y + \log 2$ para obtener una distribución gamma, ya que la variable $Z = -Y + \log 2$ tendrá por densidad

$$f(z) = f_Y(\log 2 - z) = (\theta+1)\, e^{-(\theta+1)z} \qquad 0 < z < \infty$$

es decir, una $\gamma(1, \theta+1)$. Por tanto,

$$\sum_{i=1}^{n} Z_i = -\sum_{i=1}^{n} \log X_i + n \log 2 = -S(X_1, ..., X_n) + n \log 2 \rightsquigarrow \gamma(n, \theta+1)$$

y, en consecuencia, $-E[S] + n \log 2 = n/(\theta+1)$. Por tanto, el ECUMV para $1/(1+\theta)$ será

$$-\frac{1}{n} \sum_{i=1}^{n} \log X_i + \log 2.$$

c) Al ser f_θ una familia exponencial, tendrá razón de verosimilitud monótona en el estadístico S y, por tanto, por el teorema de Karlin-Rubin (*PIE*, pp. 349) existe un contraste uniformemente de máxima potencia de nivel α para contrastar $H_0 : \theta \leq \theta_0$ frente a $H_1 : \theta > \theta_0$, el cual viene dado por

$$\phi(x_1, ..., x_n) = \begin{cases} 1 & \text{si} \quad S(x_1, ..., x_n) > c \\ \\ 0 & \text{si} \quad S(x_1, ..., x_n) \leq c \end{cases}$$

en donde c se determina por la condición del nivel,

$$E_{\theta_0} [\phi(X_1, ..., X_n)] = \alpha,$$

en nuestro caso, $\alpha = P_{\theta_0}\{S(X_1, ..., X_n) > c\}$. Es decir,

$$\alpha = P_{\theta_0}\{-S(X_1, ..., X_n) + n \log 2 < -c + n \log 2\} = P\{\chi^2_{2n} < 2(1+\theta_0)[-c + n \log 2]\}.$$

Por tanto, deberá ser $\chi^2_{2n;1-\alpha} = 2(1+\theta_0)[-c + n \log 2]$ y, en consecuencia,

$$c = n \log 2 - \frac{\chi^2_{2n;1-\alpha}}{2(1+\theta_0)}.$$

Problema 1.2

Las guardias en un determinado hospital parecen ser muy monótonas (es decir, siempre se presentan casi el mismo número de pacientes) o muy variables (es decir, unos días hay muchos pacientes y otros casi ninguno) dependiendo de las horas y de los días. Para comprobar esta idea, un grupo de residentes calculó la cuasivarianza muestral en $n_1 = 13$ horas-días del primer Grupo, obteniendo el valor $S_1^2 = 50'8$ y la cuasivarianza muestral en $n_2 = 16$ horas-días del segundo Grupo, obteniendo el valor $S_2^2 = 31'3$. Admitiendo que el número de pacientes analizados en ambos grupos son independientes y con distribuciones normales, calcular el intervalo de confianza para el cociente de varianzas poblacionales de coeficiente de confianza, del 95 %. A partir del intervalo calculado, ¿se puede concluir que no hay diferencias significativas entre ambas varianzas?

2 puntos

El Intervalo de Confianza para el cocientes de varianzas de dos poblaciones normales independientes de medias desconocidas, CB-sección 6.5 ó ID-sección 3.5, es

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}} \ , \ \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;1-\alpha/2}} \right].$$

Como el coeficiente de confianza es $1-\alpha = 0'95$, será, a partir de la Tabla 6 de la Distribución F de Snedecor, es

$$F_{n_1-1,n_2-1;\alpha/2} = F_{12,15;0'025} = 2'9633$$

y

$$F_{n_1-1,n_2-1;1-\alpha/2} = F_{12,15;0'975} = \frac{1}{F_{15,12;0'025}} = \frac{1}{3'1772} = 0'31474$$

por la propiedad de la F de Snedecor, CB-sección 5.3.3.

Estos cuantiles también se pueden determinar con R fácilmente ejecutando (ID-sección 2.6 o EAR-sección 3.5.3)

```
> qf(1-0.025,12,15)
[1] 2.963282

> qf(1-0.975,12,15)
[1] 0.3147424
```

El Intervalo de Confianza buscado será por tanto,

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}} \, , \, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;1-\alpha/2}} \right] = \left[\frac{50'8/31'3}{2'9633} \, , \, \frac{50'8/31'3}{0'31474} \right] = [0'5477 \, , \, 5'1566].$$

Como el 1 pertenece al anterior intervalo de confianza cabe inferir que no existen diferencias significativas entre las varianzas poblacionales, es decir, que no existen diferencias significativas en la dispersión o monotonía de pacientes en ambos días-horas.

Referencias

CB: **Estadística Aplicada: Conceptos Básicos**, segunda edición, 2008. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0184011EP01A02).

ID: **La Interpretación de los Datos. Una Introducción a la Estadística Aplicada**, 2014. Alfonso García Pérez. Editorial UNED (código: 0105008CT01A01).

PREB: **Problemas Resueltos de Estadística Básica**, 1998. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 84011EP31A01).

EEA: **Ejercicios de Estadística Aplicada**, 2008. Alfonso García Pérez. Editorial UNED, Colección Cuadernos de la UNED (código: 0135284CU01A01).

EAR: **Estadística Aplicada con R**, 2008. Alfonso García Pérez. Editorial UNED, Colección Varia (código: 0137352PB01A01).

EBR: **Estadística Básica con R**, 2010. Alfonso García Pérez. Editorial UNED, Colección Grado (código: 6102104GR01A01).

ADD: **Fórmulas y tablas estadísticas**, 1998. Alfonso García Pérez. Editorial UNED, Colección Adendas (código: 0141206AD01A01).

MR: **Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo**, 2005. Alfonso García Pérez. Editorial UNED, Colección Educación Permanente (código: 0186080EP03A01).

PIE: **Principios de Inferencia Estadística**, Vélez, R. y García Pérez, A. . Editorial UNED.

LCP: **Lecciones de Cálculo de Probabilidades**, 1988. V. Quesada y A. García Pérez. Editorial Díaz de Santos.